
MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

# An Application of Predictive Modeling to New Business Claims in Automobile Insurance

Magdalena Brach

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# AN APPLICATION OF PREDICTIVE MODELING TO NEW BUSINESS CLAIMS IN AUTOMOBILE INSURANCE

Magdalena Brach

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

**Advisor:** Fernando José Ferreira Lucas Bação

February 2020

# ACKNOWLEDGMENTS

I would like to thank Ageas Seguros for giving me an opportunity for an internship. I am grateful to all my colleagues from the Pricing and Business Analytics Team. I have enjoyed working with them and contributing to their work. I would like to thank my supervisors for giving me guidance and support.

I am grateful to all the amazing people I met in Lisbon, who changed my life and point of view. Also, to my family that was always supporting my unconventional ideas and letting me live my life the way I wanted.

# ABSTRACT

One of the biggest struggles of insurance companies is to acquire new clients and retain the most profitable ones. Consequently, companies often accept new clients with less restricted requirements and offer a much better price than at policy renewal. Thus, many companies make big losses in new business. Actuary and predictive models might help to cut down this problem by discovering the characteristics of unprofitable clients and limit the losses. This report presents a modeling approach suggested by the company and developed during a 9-month internship.

The main objective of this project was to build the models on the probability of having a claim within 90 days from the start date of a policy in new business auto insurance. Also, the profiles of the riskiest clients were created, and further actions defined. A comparison of the models for claims that happened within 90 days and those that happened in the rest of the first annuity revealed some differences in characteristics of those clients/policies. The results were obtained by training four classifiers, namely logistic regression, decision tree, random forest, and neural network on the original and balanced datasets. This paper discusses also different performance metrics, and an imbalanced data problem. Finally, the paper suggests important aspects that must be taken into consideration in future work.

# KEYWORDS

Insurance; Claims; Fraud; Automobile; New Business; Risk; Predictive Models; Unbalanced Dataset; Logistic Regression; Decision Trees; Neural Networks; Evaluation Metric

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **ABT** | Analytical Base Table |
| **AIC** | Akaike Information Criterion |
| **AUC** | Area under the ROC Curve |
| **BM** | Bonus/Malus – a system of rewards and penalties applied to insurance price |
| **CCC** | Own Damage (Choque, Colisão ou Comportamento) |
| **DT** | Decision Tree |
| **GLM** | General Linear Models |
| **FN** | False Negatives |
| **FP** | False Positives |
| **FRB** | Theft (Furto ou Roubo) |
| **LoB** | Line of Business |
| **LR** | Linear Regression |
| **NB** | New Business |
| **NN** | Artificial Neural Networks |
| **QIV** | Windscreen Breakage (Quebra de Vidros) |
| **RCV** | Third Party Liability – TPL (Responsabilidade Civil) |
| **RF** | Random Forest |
| **ROC** | Receiver Operating Characteristic |
| **SAS EG** | SAS Enterprise Guide |
| **SAS EM** | SAS Enterprise Miner |
| **SMOTE** | Synthetic Minority Oversampling Technique |
| **TN** | True Negatives |
| **TP** | True Positives |

# 1. INTRODUCTION

Insurance might be described by its basic characteristics that typically include pooling of losses, payment of unintentional losses, risk transfer, and indemnification. It means that insured clients are sharing the losses so that in this process average loss is substituted for actual loss. Additionally, the loss must be accidental, and a pure risk is transferred from the insured to the insurer, who commonly is in a stronger financial position to pay the loss than the client. This, in consequence, should allow the insured to be restored to his approximate financial position prior to the occurrence of the accident (Rejda & McNamara, 2017, pp. 21-22).

Private auto insurance is one of the products the insurance companies offer to their clients. The insurer protects the insured against legal liability occurring in case of auto accidents that might cause property damage or bodily injury to others. Sometimes the insurance also protects the passengers travelling with the insured driver, as well as, baggage, and often covers breakdown assistance. The most basic insurance, third party liability for land-based motor vehicles and their trailers, is a legal requirement in Portugal. Consequently, it is held by all car owners. Failure to insure is punishable by law and can result in confiscation of the vehicle, payment of a fine or compensation of the parties affected in an accident. However, car owners usually want to be wider protected and choose a more comprehensive coverage plan. Thus, the insurance provider must adjust their offer according to the clients' needs.

One of the biggest struggles of insurance companies is to acquire new clients and retain the most profitable ones. Pricing differentiation can help to achieve this goal by rewarding valuable clients and penalizing loss-making ones. It is not an easy task to determine them, especially if the clients are new and do not have any track record in the insurance company. In order to achieve balance, companies often accept new business clients with less restricted requirements and then tighten the risk requirements at policy renewal. The consequence of this is that many companies make big losses in NB. However, actuary and predictive models might help to cut down this problem by discovering the characteristics of unprofitable clients. This, in turn, might help to limit the losses.

Fraud is another problem that insurance companies are facing. Both confirmed and undiscovered cases are highly affecting the losses. Statistics of Ageas Seguros in auto insurance presented in Figure 1.1 show that 38% of overall detected frauds within 2015-2018 were in NB (48% in 2018 only). Within those years, on average, 17% of frauds occurred within 90 days but if we take NB frauds only, it was 44% of them (49% in 2018 only) within 90 days. Those numbers reflect the frauds that were confirmed, so potentially there are many more that were not detected or possible to be confirmed.

Moreover, the study made by the Portuguese Association of Insurers in 2016 shows that 29% of the respondents consider fraud as normal and deserved. 30% of the respondents claim that fraud in insurance is unacceptable and unjustifiable, while the rest, that is 41%, consider it unacceptable but justifiable in some cases. This proves the Portuguese nation, in general, does not coherently perceive fraud and many people do not consider fraud as a vice. The overall attitude towards fraud is rather negative but still many people consider it a customary behavior (Associação Portugesa de Seguradores, 2016, pp. 6-7).

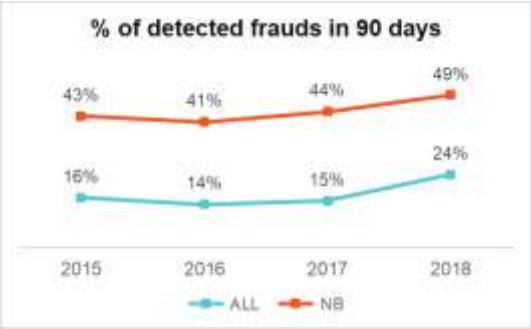


Figure 1.1 - Fraud Statistics in Ageas

The main objective of this project was to build the models on the probability of having a claim within 90 days from the start date of a policy in new business auto insurance. This would allow to create the profiles of the riskiest clients and define further actions. Also, a comparison of the models for claims that happened within 90 days and those that happened after this time (but still within one year) can reveal differences in characteristics of those clients/policies. Therefore, pricing strategies might be differentiated and become more competitive. The results obtained might be analyzed against fraud data, in order to discover some potential relationships and suggest future research. Those objectives were to be achieved during a 10-months academic internship at Ageas Seguros Portugal. The work was done between September 2018 and June 2019.

The policyholder can choose different covers included in a policy, so there exist claims that are corresponding to them. Every policy has an obligatory minimum cover that is a third-part liability (RCV). This option covers the expenses of the damages made by a policyholder to the other party involved in the accident. Also, basic assistance is included in the simplest option. The client can choose one of the upgrades that consist of extended third-part liability, legal defense, passengers protection, windscreen (QIV), own damage (CCC), fire, lightning or explosion, theft (FRB), natural disasters and others. Having the abovementioned covers, we can distinguish corresponding claims. As the project required to separate the analysis by the type of cover and having a limited time, 3 most common claims were analyzed and presented in this project.

Following the main objective, other secondary objectives have emerged:

1. Construct the ABT table,
2. Build separate models for 3 most common types of claims,
3. Compare the results between modeling of the first 90 days and the rest of the first year of the policy,
4. Build profiles of NB customers,
5. Relate this to fraud data.

This report is organized as follows. In the second section literature review that is relevant to the project is presented. The third section gives a brief overview of the insurance offer and the main KPI's used to monitor the performance. The methodology used in the project is described in detail in the fourth section. It includes project planning. The fifth section presents and discusses the results of the modeling process. Then, the conclusions are drawn in the next, sixth section. Finally, the last section outlines some limitations of the project, as well as, recommendations for future work, both on the same topic and further steps.

# 2. LITERATURE REVIEW

Both, before and during the project development there was a literature review conducted as the theoretical and practical support was needed. Below there is a brief description of the paperwork that provided explanations, examples, and inspirations for completing the assignment. The review is not strictly scientific as the project presented was an internship assignment.

## 2.1. INSURANCE

In 2017, Chang and Nelson provided a summary of the challenging insurance industry regarding pricing, competition and clients' expectations. They talk about crucial usage of data in insurance to assess the risk, set the prices and retain the customers. Then some data opportunities in the underwriting area are presented like usage-based insurance, leveraging external and real-time data.

Werner and Modlin (2016) outlined basic insurance ratemaking concepts and techniques. They highlighted the measures that are useful to monitor portfolio performance. It is important not only in day-to-day routines but also to measure the impact of new strategies.

There are a few critical moments in the relationship between an insurance company and a client. One of them is a moment of claim when the client asks for compensation for the damages. Bond and Stone (2004) stated that the automotive claims experience is crucial to retain the clients and meet their expectations.

## 2.2. CLAIMS MODELING

To identify the most common techniques and workflows of claims modeling, a review of some related papers was led. However, many authors used actuarial and econometric models. The examples are given below.

Najim (2018) provided an overview of claim analytics and the general modeling process. The described processes are applied in the insurance domain. He presented also a case study in SAS Enterprise Guide® and SAS Enterprise Miner® with a remark that the process could be replicated in other statistical software.

The insurance industry has a close association with actuarial science. De Jong and Heller (2008, pp. 99-110) present an example of a General Linear Model (GML) application on the vehicle insurance claims. However, the example does not use many exploratory variables.

David and Jemna (2015) presented a pricing approach to claim modeling. The target for their model is a frequency of claims, also known in theory as count data. It allows for the identification of the main risk factors and prediction of the frequency given the risk factors. The methods used in the research are common in actuary namely, the Poisson regression that is a special case of GLMs and the negative binomial distribution.

Hurdle model and GLM-Gamma distribution have been analyzed on a portfolio of one of the largest non-life insurance company in Russia (Gilenko & Mironova, 2017). The predictions were provided for two targets claim frequency and severity. The results showed that additional factors should be considered in the tariff calculation of the company that provided the data.

In 2019, Burri, Burri, Bojja, and Buruga presented claim modeling using machine learning algorithms. Naïve Bayes Updatable, Naïve Bayes, Multi-Layer Perceptron, J48, Random Tree, Logistic Model Tree, and Random Forest methods were compared. The performance was measured using precision, recall, F-measure, Matthews Correlation Coefficient, AUROC, and PRC area. The dataset had only 8 features that were not explored in the paper.

A broader literature review on machine learning algorithms use in insurance might be found in a recently published survey by the Society of Actuaries (2019). It concludes that there is a growing literature on this topic but only limited evidence of ML algorithms usage in insurance problems. Morover, the study concluded that the evidence about the performance improvement by using the various machine-learning methods in claims prediction is still limited.

### 2.3. IMBALANCED DATA PROBLEM

Having an imbalanced data set, the literature on this subject was reviewed. Mainly the papers that include implementation guidelines were used during the work, as SAS software does not have many features that might deal with the problem.

Chawla, Bowyer, Hall, and Kegelmeyer (2002) presented a methodology and study on the potential of SMOTE, a method proposed by them to deal with unbalanced data. Experiments were performed using C4.5, ripper and a naïve Bayes classifier. The study on 9 different datasets proves that algorithms trained on balanced data improve the performance in ROC space. The study proposes also two extensions of the method that allow to use it on a dataset with nominal variables. However, their implementation is not explained sufficiently clear.

Authors like Guzman (2015), Wang, Lee and Wei (2015) or Boardman, Biron, and Rimbey (2018) applied methods by using SAS Code, as methods like SMOTE or TOMEK link undersampling are not standard methods in SAS Software. The first author used a logistic regression model on each balanced dataset and compared the results to prove that SMOTE gives better results in a churn classification problem in the study. A similar approach was presented in the second paper but using decision tree and neural network methods on a marketing dataset. Also, a cost-sensitive, random oversampling and undersampling approaches were applied. The authors of the third paper compared SMOTE to Tomek method that helps to clear majority class examples from class boundaries. Also, the combination of those both methods was tested and gave the best performance together with the pure SMOTE method. The results were obtained by training four classifiers for fraud target on transactions database.

An extensive literature review on class-imbalanced data can be found in Haixiang, et al. (2017). One of the topics in this review is model evaluation in the presence of rare cases. The paper enumerates how many articles talked about the most frequently used metrics such as AUC/ROC, Accuracy/Error, G-mean, F-Measure, Sensitivity, Specificity, Precision, Balanced Accuracy, and Matthews Correlation Coefficient.

# 3. BASIC INSURANCE TERMS

As part of business understanding, it is fundamental to know the basic insurance terms and ratios. This chapter explains how to calculate the most important KPIs. They are very crucial for every insurance company to measure the dynamics and performance of the portfolio of the clients. Especially, within different groups to be compared and targeted (Werner & Modlin, 2016, pp. 1-11).

## 3.1. INSURANCE RATIOS

1. **Exposure** – the basic unit of risk that underlies the insurance premium. It is measured in a fraction of an annuity of insurance. Thus, if the contract is valid for half a year, then the exposure equals 0.5. There are different ways that insurer may measure exposure but, in this case, it is a number of insured units that are exposed to loss at a given point in time. It is calculated as:

$$\text{Exposure} = \frac{\#\ years\ of\ contract}{\#\ years\ in\ force}$$

2. **Frequency** – measures claims ratio over risk exposure. Helps to measure the utilization of the insurance coverages or the effectiveness of specific underwriting actions and is calculated as:

$$\text{Frequency} = \frac{\#\ claims}{\sum exposure}$$

3. **Average severity** – expresses the average value of the cost per claim. Analyzing the dynamics of severity allows monitoring the loss trends and highlighting the impact of changes in claims handling procedures. It is calculated as:

$$\text{Severity} = \frac{\sum cost\ of\ claims}{\#\ claims}$$

4. **Loss Ratio normal and capped** – measures the portion of total losses that are paid for each currency unit of premium. It allows for assessing the profitability of the business. The capped version excludes extreme and less frequent values of claims. The basic version is calculated as:

$$\text{Loss Ratio} = \frac{\sum cost\ of\ claims}{\sum earned\ premiums}$$

   where the value of earned premiums is the amount that the insurer has already earned for a given exposure.

5. **Average premium** – measure how much premium is paid on average by a policyholder. Premium itself is the amount the insured pays for insurance coverage or the aggregate amount a group of insureds pays over a period of time. It is calculated as:

$$\text{Avg Premium} = \frac{\sum earned\ premiums}{\sum exposure}$$

6. **Pure premium or loss cost** – measures the average loss per exposure, so the amount that is necessary to cover the losses and expenses. It is calculated as:

$$\text{Pure premium} = \frac{\sum cost\ of\ claims}{\sum exposure} = Frequency * Severity$$

7. **Commercial premium** – measures the expenses that are added into pure premium and are related to costs of operation and maintenance.
8. **Conversion Rate or close ratio** – measures the rate at which prospective clients accept a new business simulation and it is calculated as:

$$\text{Conversion rate} = \frac{\#\ accepted\ quotes}{\#\ quotes}$$

All the above measurements should be used for monitoring the identified high-risk groups of having a claim. Nevertheless, the calculation might require other data than used in the project, as many of the measures are done over time. So, the first view of each policy is not enough to operationalize performance monitoring.

### 3.2. MOTOR COVERAGES

The other important point in business understanding is to get familiar with the offer. Ageas Seguros offers the following 12 covers in 5 different packs:

1. Third-Party Liability Obligatory

It is an obligatory minimum cover of every policy. This option covers the expenses of the damages (legal liability) made by a policyholder to the third party involved in the accident including property damage, bodily injury or death. Also, basic assistance is included in the simplest option.

2. Third-Party Liability Additional

Superior capital of 50M EUR to the third party liability cover.

3. Own Damage

It covers accidental loss or damage. In Ageas, this cover is called “Shock, collision or rollover”, where the definitions of each are as follows. Shock is when a vehicle is damaged by a crash against a fixed or immobilized object. Collision means vehicle crash with any moving object. Rollover is a loss of normal position by a vehicle not due to shock or collision.

4. Theft

Covers loss of the vehicle due to both psychological or physical violence, or without violence. Includes also damages resulting from a theft attempt.

5. Windscreen

Insurance that covers replacement or refund in case of glass or mirror breakage.

6. Assistance

This cover provides the first response in the event of a breakdown or accident. Should the vehicle be deemed not roadworthy, the driver and passengers will be transported to the destination of their journey and the vehicle will be towed to the nearest garage or repair shop. In case of VIP modality insurance covers also problems, such as flat tire, run out of fuel, loss of keys or need for replacement vehicle in the event of a breakdown.

7. Acts of God Protection/Natural Disasters

It might be called Catastrophic events insurance and it provides cover in the event of loss involving two or more vehicles caused by an insured peril, such as flood, hurricane, landslide or earthquake.

8. Fire, Lightning or Explosion

Fire is defined as accidental combustion, with the development of flames, foreign to a normal source of fire, even though it may have its origin, and which can spread by its means. Lightning means atmospheric discharge that occurs between the cloud and the ground, consisting of one or more current impulses that give the phenomenon a characteristic luminosity (lightning), and that causes permanent mechanical deformations in the insured vehicle. An explosion is a sudden and violent action of pressure or depression of gas or vapor due to a car defect.

9. Passengers Protection

Includes death or permanent disability cover, medical treatment expenses, expenses of the funeral, temporary full disability during the hospital treatment meaning lost wages. It can protect only the driver or all the passengers.

10. Legal Defense

This insurance covers costs incurred in taking a civil action against a third party or defending a civil action brought by a third party.

11. Luggage

It guarantees to the insured the reimbursement of material damages caused to luggage and personal objects, that were in the insured vehicle, of the vehicle occupants, which result directly from an accident covered by the policy, under the coverage of the Own Damage, Fire, Lightning or Explosion, Theft, or Natural Disasters.

12. Tires

Covers the damages of the tires. Ensures assistance in the event of a puncture or burst of the tire and the cost of repair or replacement.

# 4. METHODOLOGY

This chapter presents the approach of data understanding and preparation process, then the examples of the variables used are given. Next, the problem of data imbalance is discussed, and the possible solutions are presented. Moreover, three modeling approaches for a binary target are described, including their way of implementation. In the end, the most appropriate performance metrics that help to choose the best model are characterized.

The whole work was based on the following modeling process (Najim, 2018, pp. 4-5):

- Business Goals (and Model Design)
- Data Scope and Acquisition
- Data Preparation
- Variable Creation (a.k.a.: Feature Engineering)
- Variable Selection (a.k.a.: Feature Selection)
- Model Building (a.k.a.: Modeling Fitting)
- Model Validation
- Model Testing
- Model Implementation and Monitoring

Consequently, the timeline presented below in Figure 4.1 was suggested by the team. In the first step, business understanding included not only the definition of the business goals, data scope, and acquisition but also getting to know the team, stakeholders, timeline planning and deep-diving into auto insurance topic itself. Due to a complicated structure and a nonstandard data view, the data preparation phase was the longest step. It took almost two months longer than expected and included the abovementioned variable creation and selection. Those and further steps are described in detail in the following chapters. Unforeseen events forced some initially planned tasks to be excluded. The deployment step was removed during the project due to the delay and stakeholders' decision.

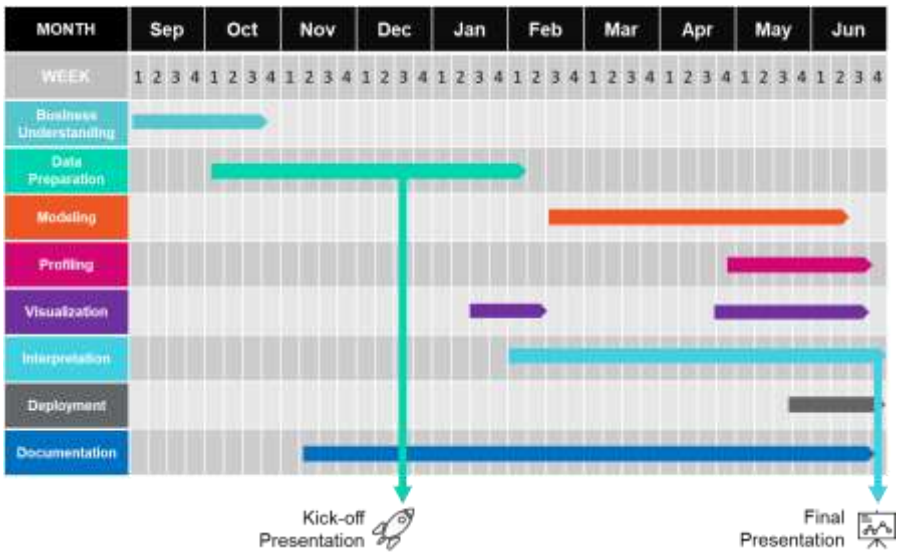


Figure 4.1 – Project Timeline

The first step included also defining the possible impact of the project on different departments highlighted in Figure 4.2. It is important for the business to know how the project can be used in the

future and bring value to other departments in the company. However, bringing to life this step depends on the implementation phase that was taken out-of-scope of this project. Therefore, the timeline presented above was more an indication at the beginning of the project than a strict plan that cannot be modified. The implementation is not described in this report but is one of the next steps to be done in the future.

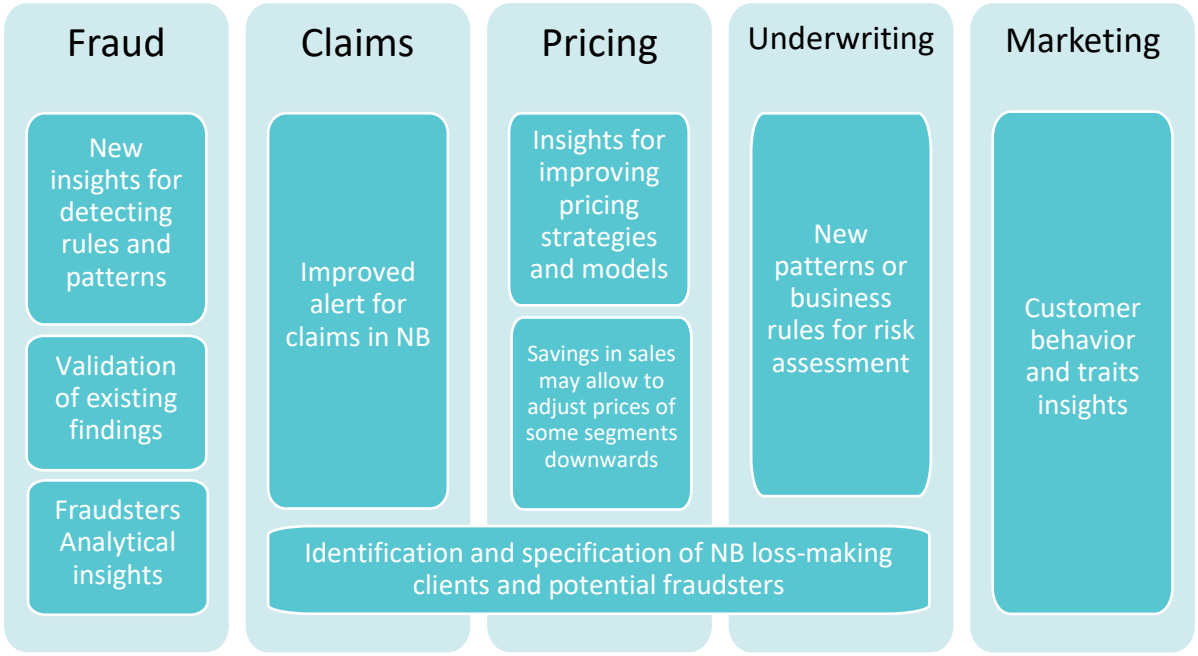


Figure 4.2 - Possible Impact of the Project on Different Departments

### 4.1. Data Understanding and Preparation

The main objective of this subsection is to describe existing data, as well as, present how it had to be joined and prepared for further analysis. To do so, all the variables had to be pre-analyzed, joined correctly, and validated. Then, it was possible to do the visualization and calculate the basic metrics.

For the purpose of this work, the main set containing around 304k observations and around 120 variables was prepared using SAS EG. Every observation in this dataset represents one policy of NB individual client that was signed between January 2015 and September 2018. Joining the data from different sources (e.g. tables based on different information systems) was the most time-consuming part of this project, as the tables were built in many different manners. There were some idiosyncrasies in the database of the company, as well as, some errors and missing or nonsense values that had to be studied in detail. This phase was also demanding as the information came from 41 tables. The full list is available in Appendix A.

Moreover, most of the data analysis in insurance is done for a snapshot or a certain timeframe, not for a different date for every policy (in this case starting point date). The solutions of the abovementioned problems are presented in the posterior subchapters.

As the claims have a different specification, the decision was made to analyze 3 topmost frequent types that are, presented in Figure 4.3, RCV, QIV, and CCC. That means there were created separate models for each type of claim, having in mind that the dataset for each model contains only policies

that have relevant cover included as presented in Table 4.1. In consequence, there will be 3 datasets in the further analysis but some of the data treatment will be conducted on the joined dataset to facilitate the process. However, a variable selection will be done for every target separately.

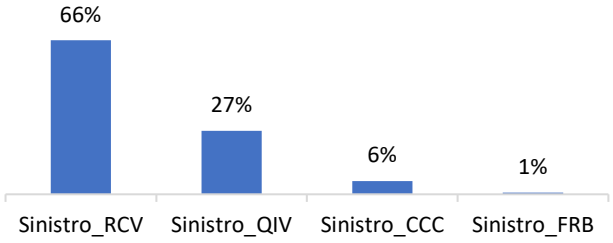


Figure 4.3 – Percentage of Total Claims for Most Common Types

| All claims | Claim frequency for all the policies | # of claims | # of claims within policies having the cover | # of policies with this cover | Claim frequency of policies with this cover |
|---|---|---|---|---|---|
| **Sinistro_RCV** | 14.20% | 41816 | 41816 | 294495 | **14.20%** |
| **Sinistro_QIV** | 6.05% | 17815 | 17499 | 253535 | **6.90%** |
| **Sinistro_CCC** | 1.26% | 3703 | 3603 | 41509 | **8.68%** |

Table 4.1 - Frequency of RCV, QIV, and CCC Claims

### 4.2. DETAILS ON VARIABLES

This part describes the data that was considered as potentially valuable for the model. Descriptive analysis and preliminary exploration of the features revealed some data problems that had to be resolved before using the modeling techniques. In some cases, the decision was taken to exclude some information. Before choosing variables, the data scope was determined as presented below. Later, the example variables describing the given categories were presented in Table 4.2 to facilitate understanding of the data treating process. However, all the variables had to be reviewed to check their quality and possible values. The values presented are before data treatment. The whole list of variables is not shared due to business decision. This part of the analysis was made jointly for all 3 datasets and the common target 'Within_90days' to simplify the process and avoid repeatability on highly alike data.

Data Scope:

- New Business clients
- Policies subscribed within January 2015 – September 2018
- Validation data for profiling
- Backup test data October 2018 – March 2019 for model testing
- Tow trucks (pt. *reboques*) and fleets (pt. *frotas*) are excluded
- Motor Personal Lines only
- Passenger cars, motorcycles, and scooters

| Variable Name | Description | Type | Values |
|---|---|---|---|
| Driver Characteristics | | | |
| **IDADE_CARTA_CONDUTOR** | Number of Years of Driving License (based on date) | interval | From -4 to 79 |
| **IDADE_CONDUTOR** | Driver's Age (based on birthdate) | interval | From 2 to 218 |
| **Sexo_ch** | Driver's Gender | nominal | H, M, J |
| Policy Holder Characteristics | | | |
| **FOV_JUBILADO** | Flag retired | binary | 0/1 |
| **EST_CIVIL** | Civil state | nominal | CA, DI, SO, UF, VI |
| **Sexo_tom** | Policy Holder's Gender | nominal | H, M, J |
| Vehicle Characteristics | | | |
| **CVE_NACIONALIDADE** | Vehicle's Nationality | nominal | E, O, P |
| **IMPORTED_CAR** | Flag Imported | binary | 0/1 |
| **NUM_LUGARES** | Number of Seats | interval | From 0 to 9 |
| Policy Details | | | |
| **Any_protocol** | Exclusive offer | binary | 0/1 |
| **CVE_BONUS_MALUS** | Bonus Malus score | nominal | Rating 1-25 |
| **CVE_PACK** | Pack | nominal | E.g. CA, DC, DD |
| Historical Data | | | |
| **No_anulado_before** | Number of canceled policies | interval | From 0 to 26 |
| **No_anulado_before_auto** | Number of cancelled motor policies | interval | From 0 to 23 |
| **COND_BLOCKED_EVER** | Flag Driver Ever Blocked | binary | 0/1 |
| Demographic Data | | | |
| **ZONA_RC** | District | nominal | 26 values e.g. |
| **URBAN_RURAL** | Urbal/Rural Area | nominal | RURAL, URBAN |
| **NACIONALIDADE** | Nationality | nominal | e.g. AO, AU, BE, BG |

Table 4.2 - Examples of the Variables

The distribution of both continuous (interval) and categorical (class) variables should be reviewed to determine the extent of missing data and transformation that might be required. However, minor missing values and categorical variables with too many values can be maintained within the datasets meant to be used by robust algorithms such as random forests, gradient boosting and decision trees ensembles. In the case of the remaining models, various transformations can be applied to the relevant data, through techniques such as normalization and imputation of missing values. For logistic regression and neural network modeling, data replacement node is required to impute some input variables with around 14% missing. All the variables that have 50% or more missing values were excluded from the analysis.

Further data transformation was performed for two purposes: to enable distribution of the transformed variable approximate normality and to discretize the interval variables by creating “bins” or “buckets” based on quantiles before feeding the variable to the logistic regression or neural network as an input. Various transformations of the variables might help to best expose their information content. Consequently, the model can be trained on the data with limited noise.

Although there are many variables available, some of them are quite useless as their discrimination ability is not strong enough. This might be a consequence of highly skewed continuous variables or heavily loaded on 1 class nominal variables. Hence, the usefulness of those variables should be reviewed before final modeling. Some of them could be then transformed to increase predictive power or removed to reduce dimensionality.

Extremely skewed variables can be problematic because the number of records available to predict the target varies greatly across the range of the input features. Those extreme values in the points of the tails of the distributions might have a great impact on the model that is being trained. Consequently, that might cause the resulting model fit to be suboptimal for many predictor values as most of the predictions are not being made at these extreme values. This problem may also result in making the feature appear more (or less) important than it actually is (Wielenga, 2007, p. 3).

### 4.2.1. Missing values diagnosis – data incompleteness

Decision tree and random forest modeling can be performed without missing values treatment. However, it is not possible in the case of logistic regression and neural networks.

Moreover, when comparing predictive models, it is important to bear in mind that all models should use the same observations. If an entity is omitted from scoring for one model but not from another you get invalid model comparisons. Therefore, the missing input variables should be treated first (SAS Institute Inc., 2018, p. 158).

The missing values can be counted both ways, per variable and per observations. Consequently, they can be treated in the following manner. Exclusion of incomplete observations should be considered, as well as, the variables with many missing values should be excluded. The ones that can be useful for improving models’ performance might be further imputed by using appropriate techniques.

Some of the missing values were present as the result of the data table structure and values the variable could take. In other words, some of the events are coded only for the event occurrence while for lack of occurrence they indicate by missing values (Wielenga, 2007, p. 3). In this case, it was possible to impute the value of 0 or create a new category according to the data type of the feature. For instance, the missing value for variable CVE_TRANSFERENCIA means that there was no transfer of policy, so these cases can be grouped as another category “no transfer”. The variables indicating additional safety equipment in the car had also many missing records but as the company treats them as optional, so the missing data can be treated as 0s.

Before removing variables from the analysis their predictive power had to be checked (excluding missing observations). This was done in order to choose the incomplete variables that have some predictive potential and might be imputed. The initial predictive power analysis (for clients with common target ‘Within_90days’) by Variables Selection and Decision Tree Node and Random Forest Node indicated that none of the highly missing variables had high importance.

| Missing Variables per observation | | |
|---|---|---|
| Limit | # observations | % observations |
| missVar > 5 | 23382 | 7.74% |
| missVar > 6 | 9870 | 3.27% |
| missVar > 7 | 9780 | 3.24% |
| missVar > 8 | **8560** | **2.83%** |
| missVar > 9 | 19 | 0.01% |

Table 4.3 - Incomplete Observation Counts vs. # Missing Variables

Then, the simple multivariate approach was used. As mentioned before, the non-informative observations were removed (Shapiro, 2016). The missing values were counted per observation and plotted in Figure 4.4. In this and further analysis, the variables with missing values rate higher than 30% were excluded. The decision was made to remove less than 3% of this data (see Table 4.3) while being careful so as not to remove too many target records. As presented below, in case of removing observations with 9 or more missing features for individual clients, it was only 102 policies with a claim within 90 days out of 8560 observations to be removed, which accounts for 2.83% of the data. Also, Table 4.4 of missing patterns was created where the counts of each group represent the number of observations that have the same variables' values missing (marked as X). The code for the MI procedure (SAS Institute Inc., 2015; Wicklin, 2016) is presented in the Appendix D. Below, there are highlighted 10 most frequent groups from 189 variables combinations. The analysis indicates that group 67 has many missing values for all 9 variables while having a low 'Within_90days' initial target frequency, which is desirable as removing this group does not imply that the target observations are being diminished.

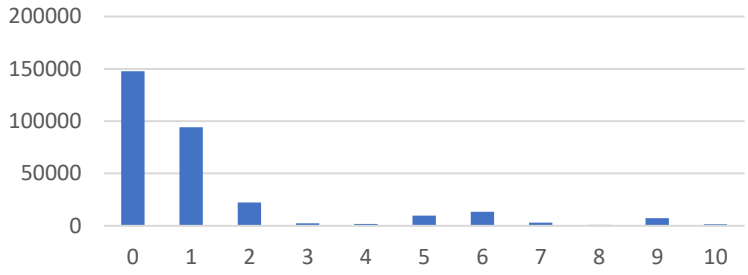


Figure 4.4 - Counts of rows that contain missing values (individual)

| Group | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | Count Observations | % of Total Observations | Claims Frequency | Initial Target Frequency |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | | | | | | | | | | 209645 | 69.36 | 0.21 | 0.05 |
| **33** | X | | | | | | | | | 39194 | **12.97** | 0.17 | 0.04 |
| **58** | X | X | X | | | X | X | | X | 13462 | **4.45** | 0.18 | 0.04 |
| **51** | X | | | | | | | | X | 10556 | **3.49** | 0.24 | 0.04 |
| **67** | X | X | X | X | X | X | X | X | X | 8540 | 2.83 | 0.06 | **0.01** |
| **41** | X | X | X | | | X | X | | | 7058 | 2.34 | 0.17 | 0.03 |
| **23** | | X | X | | | X | X | | X | 4473 | 1.48 | 0.26 | 0.04 |
| **36** | X | X | | | | | | | | 2336 | 0.77 | 0.21 | 0.04 |
| **16** | | | | | | | | | X | 2041 | 0.68 | 0.35 | 0.05 |
| **9** | | X | X | | | X | X | | | 1524 | 0.50 | 0.25 | 0.05 |
| **27** | | X | X | X | X | X | X | X | X | 805 | 0.27 | 0.06 | 0.01 |
| Total Groups | **6** | **7** | **6** | 2 | 2 | **6** | **6** | 2 | **6** | | | | |

Table 4.4 - Removing Uninformative Observations

69% of the records do not have any missing values. The most incomplete observations were removed, and other missing values are going to be imputed or treated as a separate group in case of nominal variables. However, if the missing bin is relatively small, it will be joined with another bin of similar target proportions.

### 4.2.2. Outliers, Incorrect values, and Missing Values Imputation - Noise

An outlier is an observation that is significantly distant from other observations. It might be a consequence of incorrect measurements, errors in data collection or unusual but correct cases. Having a lot of these kinds of records might have a great impact on the interpretation of the results or modeling itself. That is why if those observations' values are not justified or considered as important in the analysis, they should be treated.

In this project, outliers were treated in one way that was based on common sense and understanding of the data. An automatic method done in SAS EG was not used as most of the variables had correct values and the goal was to model all the not so extreme records.

The distributions and dispersion statistics of every variable were analyzed. The main conclusions were as follows. Variable Diff_cov_issue, presented in Figure 4.5, has probably outliers that should be investigated. The variable indicates the difference between a conversion and issue date of the policy. Looking at the histogram we can conclude that both dates are usually similar but there are some cases with both positive and negative extreme values. There are 90 policies for which the value was not within [-90;90] days. Only one of these policies had a claim within 90 days, so the decision was to set these values as missing and then impute them.

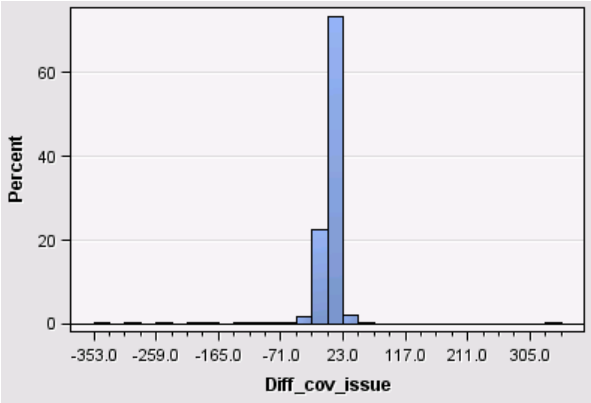


Figure 4.5 - Distribution of difference between issue and conversion date

Similar treatment was applied to IDADE_CARTA_CONDUTOR_NEW for values not within [0;85]. In the case of driver's or policy holder's age, IDADE_CONDUTOR_NEW, the impossible or incorrect values below 15 or above 100 were replaced by missing values. The same approach was for PESOBRUT values equal to 0, which is a weight of a vehicle.

The next step was to impute the missing data. As presented in Figure 4.6 the most missing values (around 13%) were imputed for QUA_VALOR_NOVO and PESOBRUT. We can see from the histograms that the 'Tree Surrogate' method attempted not to change the distributions too much. It was calibrated in a way that imputed only the variables with less than 30% of missing data. All the other imputed missing values were scarce, so to allow distribution corrections without introducing high bias.

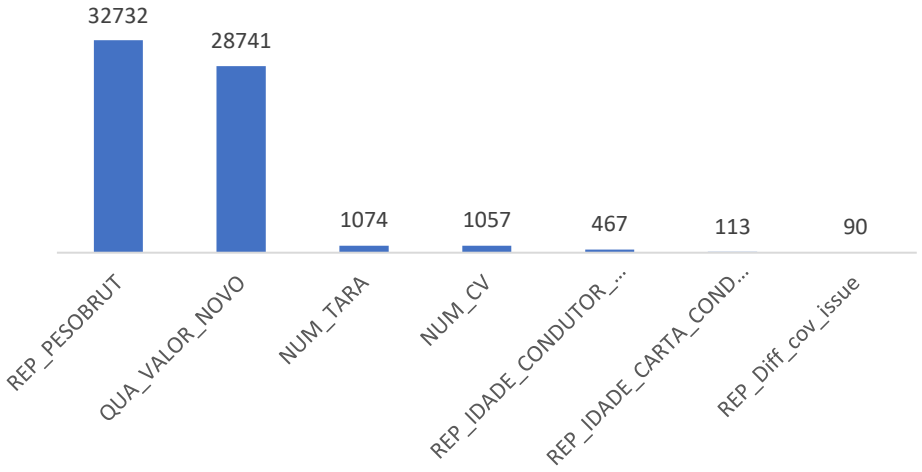


Figure 4.6 - Imputed Values

The figures below show the distributions of variables QUA_VALOR_NOVO and PESOBRUT before and after imputation. The idea is to retain information without intervening in variables' distribution.

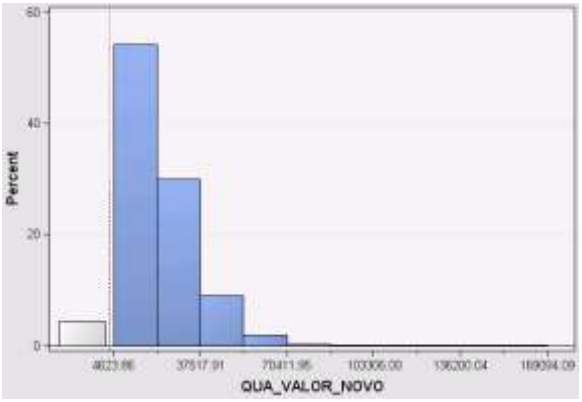


Figure 4.7 - Original distribution of value of a new car

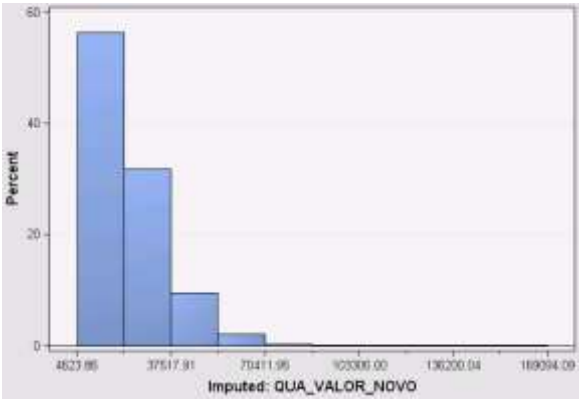


Figure 4.9 - Distribution of value of a new car after imputation

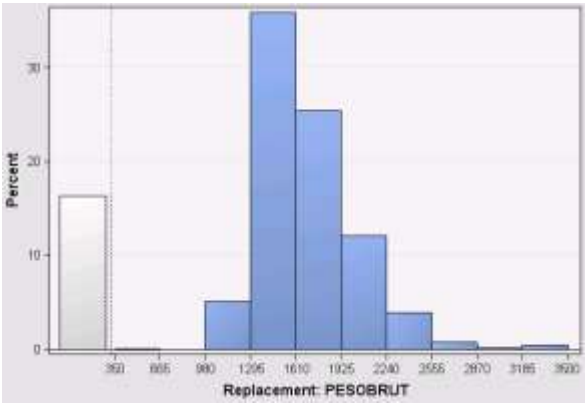


Figure 4.8 - Original distribution of car weight

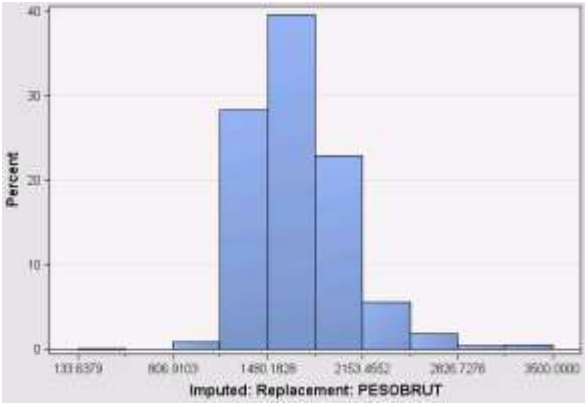


Figure 4.10 - Distribution of car weight after imputation

## 4.3. Feature Engineering and Selection

Many of the variables were created based on the other features found in the database. The best examples are the variables like age or tenure, based on the available dates. Some were just counting of previous policies or changes e.g. of plate numbers. Others were encoded as binary flags. All of them were created during the data joining process to avoid adding redundant variables to the final ABT. Next, some transformations were added to find the variables with the most predictive power. After that, there were multiple selection methods applied to pick the final set of variables for modeling. Moreover, nominal variables that are heavily loaded on 1 class were considered for removal as they usually have not high enough discrimination ability.

### 4.3.1. Transformations

Continuous inputs were standardized and added to the dataset as new variables. To increase the predictive power of initially weak variables, a few transformations were applied. Then, the ones that had had the best Chi-Squared test for the target remained (Sarma, 2007). If interval variables had less than 20 different values, there were transformed to nominal type. One of the further transformations is optimal binning which split the input into several bins, and the splits are placed to make the distribution of the target levels (for example, claimed and non-claimed) in each bin significantly different from the distribution in the other bins. Also, if the nominal variable had more than 20

different values, it was rejected. For nominal variables grouping method was applied. Levels with a very small occurrence, less than 0.01, were grouped in an _OTHER_ category. Moreover, if there still existed some missing values, there were treated as a new level.

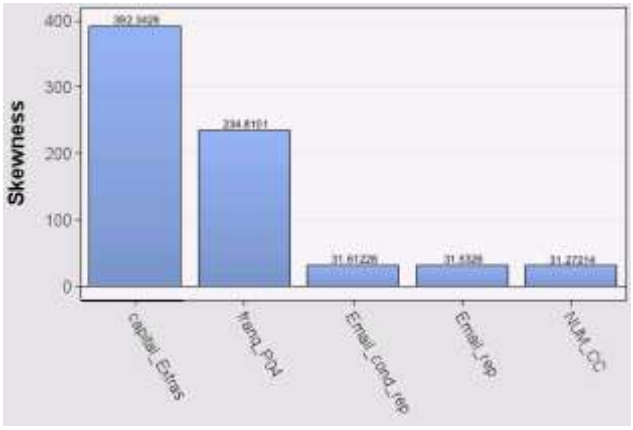


Figure 4.11 - Variables with the highest skewness

Figure 4.11 shows the top 5 variables with the highest values of skewness that might indicate that most of the values are concentrated on one side of the distribution only. Moreover, there are some variables with high values of kurtosis or too many levels. As mentioned before the following transformations were applied to continuous variables:

- Inversion
- Exponential Logarithm
- Logarithm
- Optimal Binning
- Square
- Square Root

Below, there are presented some examples of transformations:

1. EMAIL_COND_REP

This variable was transformed by applying inversion, which gave the best R-square of all transformations for this feature. The original variable indicates a number of driver's email repetition.

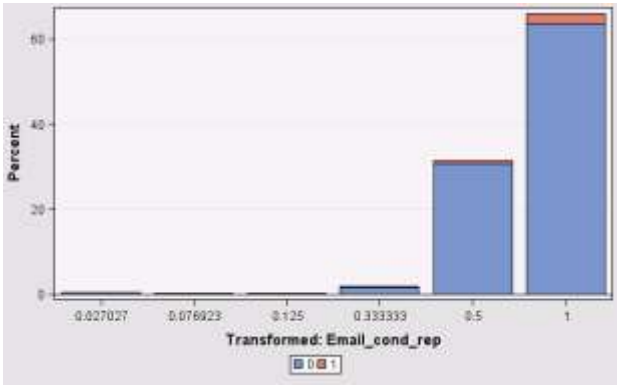


Figure 4.12 - Distribution of transformed Email_cond_rep

2. CVE_PACK

The variable had 65 levels, so the plot was hard to read. However, there were peaks for 3 levels. It means that the other ones can be somehow combined. Transformation node reduced the number of levels to 6 by binning together Packs with a frequency lower than 0.01.

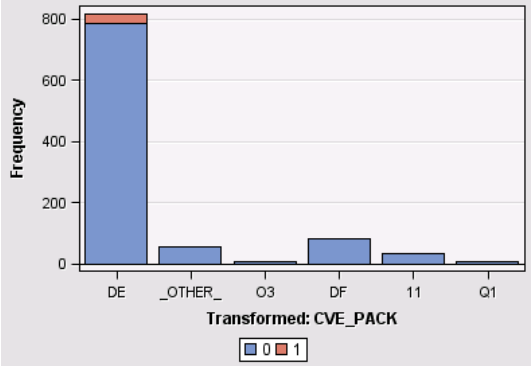


Figure 4.13 - Distribution of transformed CVE_PACK

3. PHONE_REP

This transformation is an example of an Optimal Binning with 3 resulting bins. The variable indicated the number of phone number repetitions in the database.

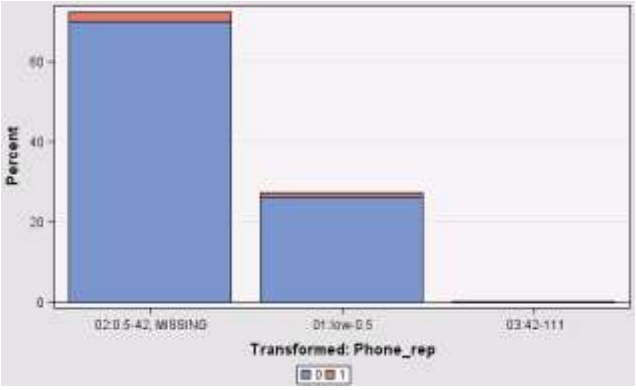


Figure 4.14 - Distribution of transformed Phone_rep

4. DIFF_COV_ISSUE

This transformation is an example of a square of previously replaced and imputed variable. It shows that for a single variable there might be several transformations applied in order to increase its predictive power.

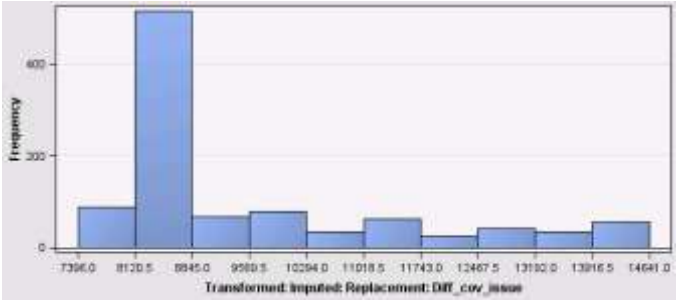


Figure 4.15 - Distribution of transformed Diss_cov_issue

5. CVE_COMBUSTIVEL

This variable indicates the type of fuel the car uses. The levels other than gasoline and diesel were grouped as their fraction was minor.

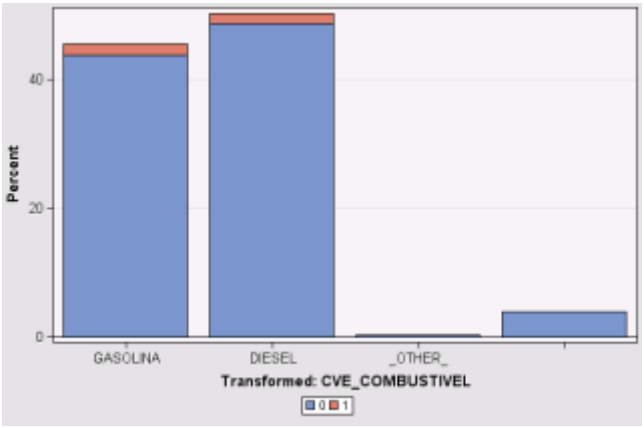


Figure 4.16 - Distribution of transformed CVE_COMBUSTIVEL

6. CVE_CARROCARIA

This variable signifies the bodywork/carrosserie of the car. The original feature had 15 distinct values. As shown below, 2 main types are limousine and combi. The less frequent that were not grouped are mono volume, van, coupé, off-road, cabrio, and transporter. All the bins <1% were grouped with other levels: CCD, CCL, CHC, PCD, PCL, PIC, ROA, TAR. The transformed variable had 9 levels.

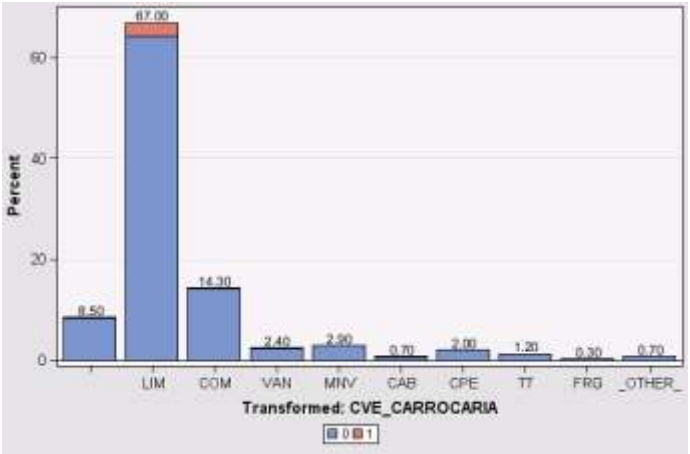


Figure 4.17 - Distribution of transformed CVE_CARROCARIA

### 4.3.2. Features Selection

In this step, only variables containing valuable information for the model have remained. Variables Selection was performed for every target separately. All further steps were repeated for each of the target variables.

Instead of choosing just one variable selection method multiple techniques were used jointly as the results can be combined using different selection properties in a SAS node. The options are as the following:

- None (no changes made from original metadata),
- Any (reject a variable if any previous variable selection nodes reject it),
- All (reject a variable if all the previous variable selection nodes reject it),
- A majority (reject a variable if the majority of the variable selection nodes reject it) (Lexi, 2018).

In this project, the majority method was initially chosen.

For deciding which features to prioritize on testing the models, we have combined features evaluation from the following methods:

- Highest variable Gini index reductions in Random Forest method (SAS HP Forest node),
- Highest variable importance in Decision Trees method (SAS HP Variable Selection node for DT),
- $R^2$ for numerical variables and $\chi 2$ statistics (SAS Variable Selection node),
- Variable worth (SAS StatExplore node),
- Variance Explained – forward selection (SAS HP Variable Selection - Fast),
- Variable clusters (SAS Variable Clustering node),
- Pearson or Spearman correlation index (SAS Variable Clustering node),
- Regularization regression technique based on input lambda parameters (SAS HP Variable Selection),
    - Ridge regression will penalize large coefficients but constraining the sum of the absolute coefficients. The least angle regression (LAR) method starts with no effects in the model and adds effects. The estimates at any step are reduced when compared to the corresponding least squares estimates,
    - Lasso regression will drop a variable completely by setting its coefficient to zero. This method adds and deletes variables based on a version of ordinary least squares where the sum of the absolute regression coefficients is constrained (SAS Institute Inc., 2018, pp. 1403-1411).

Some of the techniques above have important advantages:

1. Reduction of multicollinearity
2. Redundancy reduction with low information loss
3. Identification of underlying data structure
4. The interpretation of original input variables can be kept in successor nodes
5. Reduction of overfitting

The previously mentioned majority method seems to be reasonable but to make sure that the chosen set is not biased by the methods combined, the manual selection was performed. It included:

i. Majority selection of all the methods
ii. Majority selection of High-Performance methods only
iii. Majority selection of methods other than High Performance

Then, the variable selection procedure was repeated for data that was previously standardized. Both times under-sampled equal proportion training set was used, as well as, stratified sample.

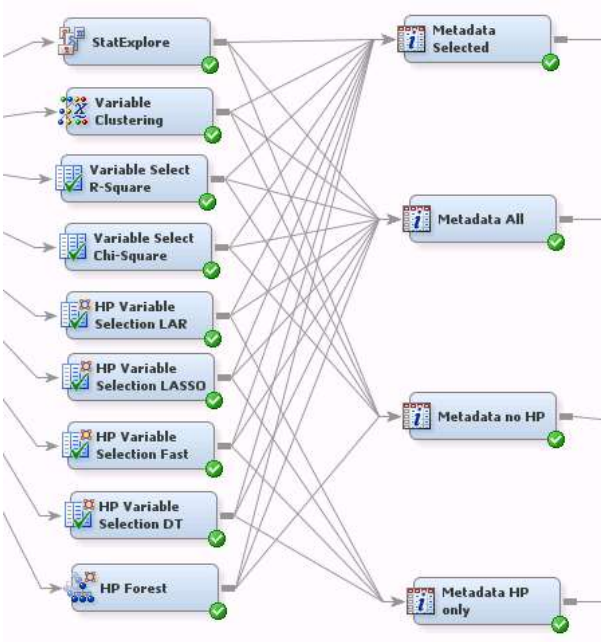


Figure 4.18 - Variables Selection Nodes

The first step to use the methods above jointly was to eliminate non-informative variables as presented in Figure 4.18. The results were combined by a voting system that allowed to reject some variables that were frequently on the bottom of the ranking. The full tables can be found in the Appendix B. Below, there is a Table 4.5 presented for the first 12 variables from the dataset and target RCV Claim within 90 days without standardized variables.

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **no HP** | **HP** | **ALL** | **no HP** | **HP** |
| CANAL | * | * | * | * | * | * |
| CDSUBRAMO | | * | | | | |
| COND_BLOCKED_EVER | | * | | | * | |
| CVE_BONUS_MALUS | * | * | | * | * | |
| CVE_CAIXA | | * | | | | |
| CVE_CAIXA_VELOC | | * | | | | |
| CVE_CARROCARIA | * | * | | | | |
| CVE_CATEGORIA | * | * | | | * | |
| CVE_CLASSE_DP | | * | | | | |
| CVE_COD_RC | * | * | * | | | |
| CVE_COMBUSTIVEL | | * | | | | |
| CVE_PACK | * | * | | * | * | |

Table 4.5 - Example of Feature Selection Table

## 4.4. DATA IMBALANCE

A dataset is unbalanced when the number of records belonging to one class is significantly lower than those belonging to other classes. That may affect the development of high quality and robust predictive models. This situation is problematic, as the objective of a classification algorithm is to learn a classifier that can discriminate the classes. A small number of records from one of the classes makes learning tasks more challenging (Douzas, Bação, & Last, 2018). Another point is that in the case of an imbalanced dataset the performance measures for standard algorithms are biased and not valid anymore.

The main motivation behind the need to preprocess imbalanced data before we feed it into a classifier is that typically classifiers are more sensitive to detecting the majority class and less sensitive to the minority class. Thus, if the issue is not properly addressed, the classification output might be biased, in many cases resulting in always predicting the majority class. One of the reasons is that the classifiers attempt to minimize a given cost, and if errors in both classes are treated equally, a classifier might develop this kind of bias towards the majority class (Boardman, Biron, & Rimbey, 2018; Guzman, 2015). Below, there are presented two basic ways of how the problem might be addressed in Figure 4.19, as well as, Table 4.6 with target frequency for each dataset in the study.

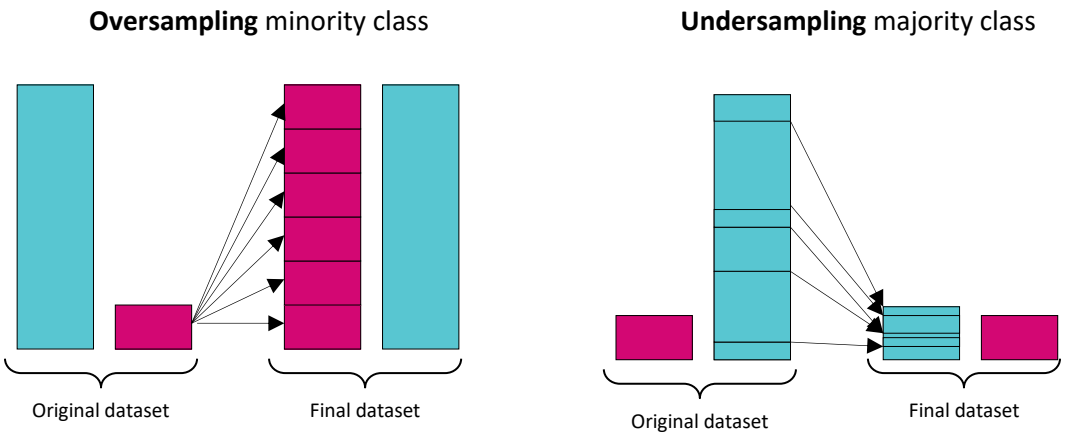


Figure 4.19 - Oversampling and Undersampling (Source: Fawcett, 2016)

| Claim | 90 days | 275 days | More 1y | General |
|---|---|---|---|---|
| **RCV** | 2.94% | 5.99% | 5.27% | 14.20% |
| **QIV** | 1.59% | 2.96% | 2.35% | 6.90% |
| **CCC** | 1.45% | 3.71% | 3.52% | 8.68% |

Table 4.6 - Claims Frequency, Imbalanced Data

### 4.4.1. Profit Matrix – Cost-sensitive Approach

This method does not change the proportions of the target levels in the dataset itself, but it adjusts prior probability and decision cost. Applying the profit matrix to a classification model can affect its decisions to maximize the expected profit. However, it won't improve a model's overall performance or discriminating power. This method is particularly useful when the goal is to maximize the profit or

minimize the cost while knowing the revenue and cost of each situation (Wang, Lee, & Wei, 2015, pp. 3-5).

### 4.4.2. Random Undersampling

This method allows to artificially balance the data by randomly dropping observations of the majority event. That approach is very simple, but it requires a big enough original dataset. The big disadvantage is that useful information might be lost easily. That is why the method should not be used exclusively in case of extremely unbalanced datasets (Wielenga, 2007, p. 5). Moreover, it often gives unstable results of the models applied to datasets balanced by undersampling.

### 4.4.3. Random Oversampling

It might be also called re-sampling (Guzman, 2015). This method is in some way similar to the one above, but it balances the data by randomly replicating the observations of the minority class, instead of dropping those from the majority class. In other words, it is resampling of rare events with replacement (Wielenga, 2007, p. 5). However, this approach does not add any information to the model and has the risk of duplicating noise samples, as well, as it impacts the variance of the target levels.

### 4.4.4. SMOTE

Synthetic minority oversampling technique (SMOTE) allows to artificially balance a dataset by generating synthetic examples in the minority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). New samples are generated in a way that they are close to existing rare cases. There exist many versions and modifications of the SMOTE algorithm. The methodology aims to reduce the effect of having few observations in the minority class.

The implementation used in this paper is the modified version of the one used by Boardman et al. (2018) with an adjustment for nominal variables and Gower distance presented by Guzman (2015, pp. 3-4), instead of Euclidean distance. The SMOTE code applied in SAS is given in Appendix E.

The method used in measuring distance that accepts nominal values is called Gower distance and is defined as a measure of similarity which is calculated between multivariate samples in case of mixed types of the variables such as nominal, categorical or binary, and continuous. Guzman (2015) has provided the following equation to calculate it:

$$d_{ij}^2 = 1 - s_{ij},$$

Where $s_{ij}$ is the Gower's similarity and is calculated as:

$$s_{ij} = \frac{\sum_{h=1}^{p_1}\left(1 - \frac{|x_{ih} - x_{jh}|}{G_h}\right) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

$p_1$: Number of continuous variables,
$p_2$: Number of binary variables,
$p_3$: Number of nominal variables,
$a$: Number of matches (1,1) in the binary variables,

$d$: Number of matches (0, 0) in the binary variables,
$\alpha$: Number of matches in the nominal variables,
$G_h$: Rank of the $h^{th}$ continuous variable.

The SMOTE process used in this project includes the following steps:

1. Start with an input dataset D.
2. Generate a dataset F with only minority class observations from the input dataset D.
3. Set the number k of nearest minority class neighbors to use for SMOTE-ing.
4. Set the SMOTE multiplier m, which is the number of additional minority class instances desired for each of the original minority class examples.
5. Generate a dataset N with the k nearest minority class neighbors selected inside the subset using Gower distance for each minority class example.
6. Generate a dataset L of observation-nearest neighbor pairs by randomly sampling from one of the k nearest minority class neighbors for each observation in N and repeating this procedure m times.
7. For each observation-nearest neighbor pair in L:
    a. Take the difference between the feature vectors of the neighbor and the observation under consideration.
    b. Multiply this difference by a uniform random number between 0 and 1.
    c. Copy values for nominal variables and for interval add the scaled difference vector to the feature vector of the observation under consideration to obtain a new, synthetic minority class example.
    d. Store the new, synthetic observations in a dataset R.
8. Add the new, synthetic minority class cases in R to the input dataset D.
9. Stop. The observations in D comprise the output dataset (Boardman, Biron, & Rimbey, 2018).

In order to avoid bias on the distances calculated, the data was standardized before applying the algorithm.

## 4.5. MODELS

In the following subsections, there are described three modeling techniques used in that project. The description is quite general and does not go into details as the given models are well known techniques. Therefore, only the important aspects are highlighted.

### 4.5.1. Logistic Regression

Logistic Regression is a type of regression used for binary target variable and belongs to a class of models called Generalized Linear Models (GML). GML is a class of statistical models that are a generalization of the classical linear ones: they use a link function to model a relationship between a response binary variable and a set of independent variables (Hosmer & Lameshow, 1989). There are many link functions, and in this case, the logit function is used.

The logit function is given by the formula:

$$logit(x) = log\left(\frac{x}{1-x}\right)$$

A binary logistic model has a dependent variable with two possible values: in this case, 1 if the given claim occurred, and 0 if it did not occur.
It is possible to define the probability of the presence of the event for each considered object $i$:

$$p_i = P(Y\ i = 1)$$

Also, it is possible to define the odds ratio of the occurrence of the event.

$$Odds\ Ratio = \frac{P(Y\ i\ =\ 1)}{1\ -\ P(Y\ i\ =\ 1)} = \frac{p_i}{1\ -\ p_i}$$

The logistic regression generates the coefficients of a formula to predict the logit transformation of the probability of the considered event's presence (claim) considering the predictors' linear combination:

$$logit(p_i)\ =\ log\ (\frac{p_i}{1\ -\ p_i})\ =\ \beta 0\ +\ \beta 1X1\ +\ \cdots +\ \beta nXn$$

Then, using logistic regression it is possible to calculate the probability for each object $i$ that the event occurs:

$$p_i\ =\ P(Y\ i\ =\ 1)\ =\ \lambda(\beta 0\ +\ \beta 1X1\ +\ \cdots +\ \beta nXn)$$

with $\lambda$ logistic function: $\lambda(x)\ =\ \frac{1}{1+e^{-(x)}}$

In each claim model, $pi$ is the probability of having a claim within 90 days or the rest of the first year for each policy and the predictors $X_1$, $X_2$ …$X_n$ are previously selected features.

Logistic regression is a very popular modeling technique for different reasons: it is well known and frequently used, it has an easy interpretation of the logit and generally provides good and robust results that may even outperform more sophisticated methods. However, the logit models are vulnerable to overconfidence which means in many cases they might overfit.

### 4.5.2. Decision Trees

Decision Tree Model is a predictive modeling approach that maps observations to conclusions about the target value. In other words, it can be used as a model for a sequential decision problem under uncertainty. It graphically represents the decision to be made, the events that may occur, and the resulting combinations of the decisions and events with associated terminal values (Quinlan, 1986).

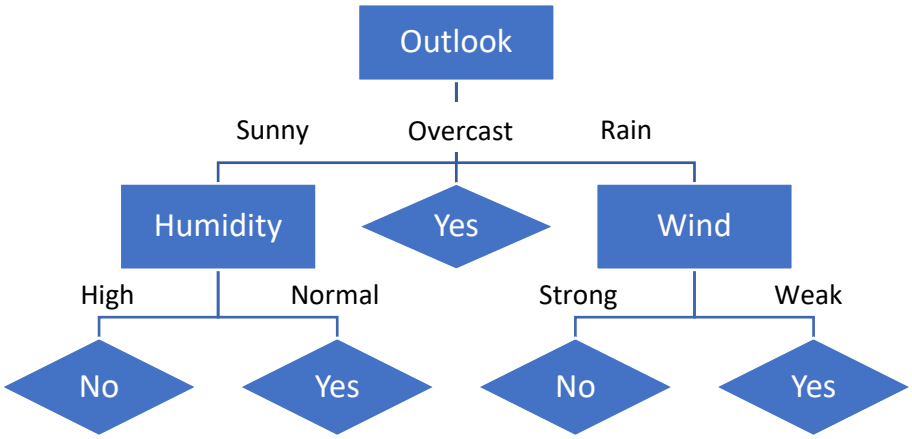


Figure 4.20 - Example of Decision Tree (Source: Mitchell, 1997)

A decision tree divides data into groups by applying a series of simple rules. The rules are organized hierarchically in a tree resembling structure with nodes connected by branches. The first rule at the top of the tree is called the root node. Each rule assigns a record to a group based on the value of one input. One rule is applied after another, resulting in a hierarchy of groups. The hierarchy is called a tree, and each group is called a node. The root node contains the entire data set. A node with all its successors forms a branch of the nodes called a subtree. The terminal nodes are called leaves. For each leaf, a decision is made and applied to all observations in that resulting final node. The paths from the root to the leaf represent classification rules (Mitchell, 1997, pp. 52-64) like in example in Figure 4.20.

This top-down approach is called the ID3 algorithm in the literature. The algorithm selects which feature to test at each node in the tree starting from the top. It is done by evaluating for each variable a statistical measure, called information gain, to determine how well it classifies (discriminates) the training set. The information gain of a feature is the expected reduction in entropy caused by partitioning the examples according to the values of a given feature. The entropy measures the level of "disorder" of the dataset and is calculated as:

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

where given a feature that has c possible values, and a given set S that contains some observations of this feature. $p_i$ is the proportion of instances in S in which the value of the feature is equal to i.

It means that the lower the value of entropy, the dataset has less different and unorganized values. Thus, this is what the algorithm is supposed to achieve by adding a new splitting rule. Therefore, the reduction in entropy is calculated as follows:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entropy(S_v)$$

where S is the dataset, Values(A) is the set of all possible values for feature A, and $S_v$ is the subset of S for which feature A has value v.

An advantage of this method over other modeling approaches, such as neural networks, is that it gives as an output easily interpretable rules that describe the logic of the scoring model in detail (SAS Institute Inc., 2018, pp. 737-738). That is why decision trees are often the most preferred method by the business. However, in many cases, this method tends to give a problem of overfitting and requires applying some approaches that help to avoid it, like grow and prune strategy or CHAID. Also, when a relationship between predictors and target is linear, decision tree tends to overfit

### 4.5.3. Random Forests

Random Forests or random decision forests are an ensemble learning method for classification, regression, and other tasks, that operate by constructing a several of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

This method makes use of a random selection of features in splitting the decision trees, hence the classifier built from this model is made up of a set of tree-structured classifiers. Its major advantage is that it can be used to judge variable importance by ranking the performance of each feature. The model “memorizes it” by estimating the predictive value of variables ant then scrambling the variables to examine how much the performance of the model drops (Akindaini, 2017).

### 4.5.4. Neural Networks

Artificial neural networks are computational techniques inspired by the brain structure and they represent a simplified model of this organ. Neural networks might be composed of billions of interconnected neurons that send and receive signals between one another. These techniques are a class of flexible nonlinear models used for supervised prediction problems. The single neuron is able to perform a small task only, but a collection of many neurons can solve more challenging problems such as learning to interpret complex real-world sensor data or learning to recognize handwritten characters (Mitchell, 1997, pp. 81-82).

The most widely used type of neural network in data analysis is the multilayer perceptron. These type of NN models were originally inspired by neurophysiology and the interconnections between neurons, and they are often represented by a network diagram instead of an equation. The basic building blocks of a multilayer perceptron are called hidden units. Hidden units are modeled after the neuron. Each hidden unit receives a linear combination of input variables. The coefficients are called the synaptic weights. An activation function transforms the linear combinations and then outputs them to another unit that can then use them as inputs. The basic network with one unit called perceptron is illustrated in Figure 4.21.

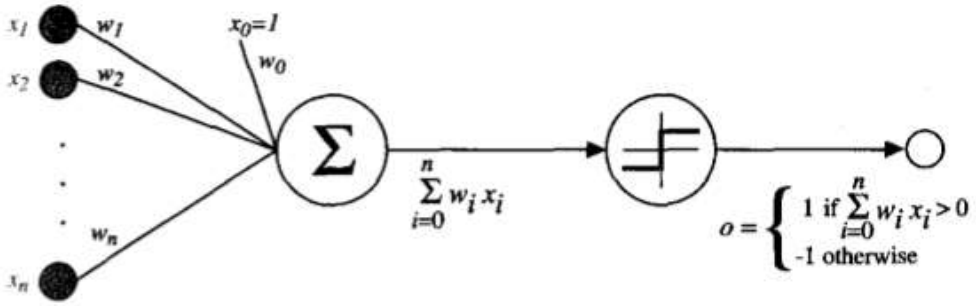


Figure 4.21 - Perceptron (Source: Mitchell, 1997, p. 87)

As networks of threshold units can represent a rich variety of functions while a single unit alone cannot, it is generally more useful to train multilayer networks of units (Mitchell, 1997, p. 88). One of the most common algorithms that can train this kind of multilayer networks is called backpropagation. It allows representing nonlinear functions when units other than linear activation functions are used for training e.g. sigmoid units.

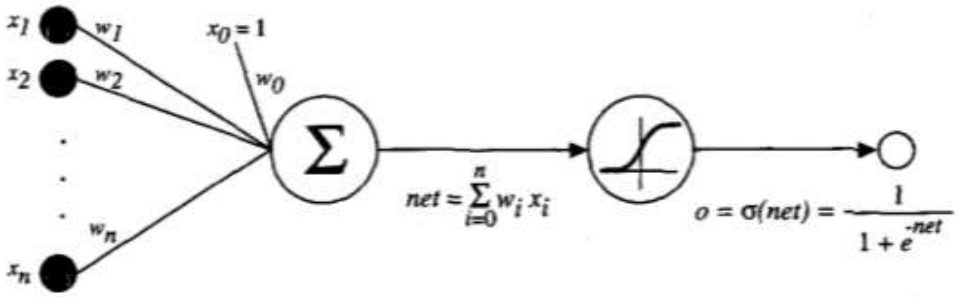


Figure 4.22 - Sigmoid threshold unit (Source: Mitchell, 1997, p. 96)

The sigmoid unit first computes a linear combination of its inputs, like the perceptron, and then applies an activation function to the output but in this case, it is a continuous function. More precisely it uses the logistic function a threshold, and it is presented in Figure 4.22. There exist more functions that could be applied like tenh, rectified linear function or maxout.

The backpropagation algorithm searches the space of possible results using gradient descent to iteratively reduce the error in the network fit to the observations used for training. Gradient descent converges to a local minimum in the error concerning the weights of the network.

Neural networks just like other modeling approaches have a learning issue of overfitting. It might happen that despite excellent performing results on training data, the algorithm is not able to generalize well to new data. One of the solutions is using cross-validation methods that can help to estimate an appropriate stopping point for gradient descent search and minimize the risk of overfitting (Mitchell, 1997, p. 123). However, having a limited time to conclude the project, this method is out of scope.

## 4.6. Evaluation Metrics

Model validation is a process to score the candidate models on the validation data set to select the best model with a good balance of accuracy and stability. This can be achieved by comparing standard metrics like precision or recall ratios or by more complex measures that allow confronting actual values against predicted values from the model. There exist many model validations methods including Gini Ratios, Lift Charts, Confusion Matrices, Receiver Operating Characteristic (ROC), Gain Charts, Bootstrap Sampling, and Cross-Validation, etc. (Hossin & Sulaiman, 2015, pp. 1-11). However, the choice of the right measures should be done having in mind the problem to be solved, objectives, time available and characteristics of the data (Sokolova & Lapalme, 2009, pp. 429-431).

For example, regular algorithms are frequently biased towards the majority class because their loss functions attempt to optimize measures such as error rate, not considering the imbalanced data distribution. That might result in creating a trivial classifier that classifies every observation as a majority class (Fawcett, 2016).

### 4.6.1. Confusion Matrix and Accuracy

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | TN | FP |
| Actual Positive | FN | TP |

Table 4.7 - Confusion Matrix

True Positives (TP) – Number of positive examples correctly classified
True Negatives (TN) - Number of negative examples correctly classified
False Positives (FP) - Number of negative examples incorrectly classified as positive
False Negatives (FN) - Number of positive examples incorrectly classified as negative

Having an imbalanced dataset, the learning process significantly impacts the solution performance, as most standard algorithms initially assume that the class distribution is balance and the misclassification costs are equal. Consequently, the accuracy measure as the loss function is not reliable and cannot be used anymore (Chawla N. V., 2005, p. 855). Accuracy is a measure of the ratio between correctly classified examples and the total number of examples, calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

SAS EM provides also a measure that is the reverse of accuracy, misclassification rate, which is a ratio of incorrectly classified examples and the total number of instances. It can be computed as follows:

$$Misclassification\ rate = 1 - Accuracy$$

### 4.6.2. Other Performance Metrics

In the case of imbalanced data problems, the following metrics are advised to be used:

- Sensitivity

It is a measure of the model capability to correctly classify the event observations and it is calculated as:

$$Sensitivity = \frac{TP}{FN + TP}$$

- Specificity

It measures the capability of the model to correctly classify the non-event observations. It is calculated as follows:

$$Specificity = \frac{TN}{TN + FP}$$

- The area under the ROC curve (AUC)

ROC curve represents the family of best decision boundaries for the relative tradeoff of TP and FP. The ROC curve is plotted on axis X that represents FP rate, and an axis Y that represents TP rate and its examples is presented in Figure 4.23. It is calculated as:

$$False\ positive\ rate = \frac{FP}{TN + FP} = 1 - Specificity$$

$$True\ positive\ rate = \frac{TP}{TP + FN} = Sensitivity$$

A good classifier should be characterized by a curve with high sensitivity and specificity values. Models with the characteristics above might be found by calculating the area under the curve and considering the models with the highest values. AUC is a useful metric for measuring the classifier performance as it is not dependent on the decision criterion or prior probabilities (Chawla N. V., 2005, pp. 855-856). The perfect classifier has AUC equal to 1, while a model with AUC less than 0.5 perform worse than random guessing.

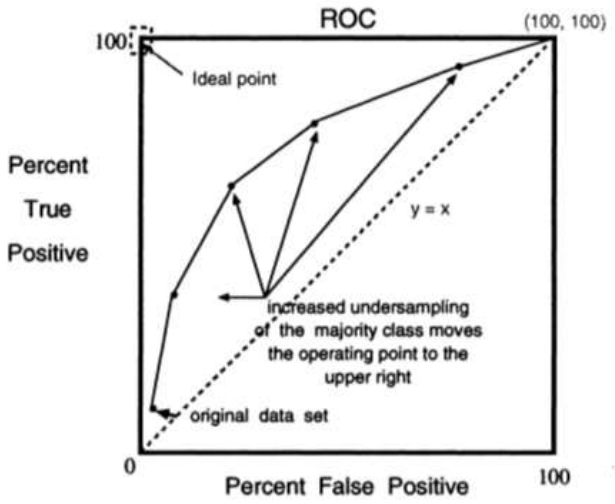


Figure 4.23 - Roc Curve (Source: Chawla N. V., 2005, p. 856)

- K-S Statistics

Kolmogorov Smirnov (K-S) statistic measures the distance between two distribution functions, in this case, the two levels of the target variable. The greater the value is, the better the model is in discriminating two classes. However, it requires a fixed reference point, meaning threshold. $F_{Ch}(s)$ and $F_{NCh}(s)$ are the cumulative distributions for each rank of definite probability of events and non-events respectively. It is calculated as follows (Guzman, 2015):

$$KS = max_s|F_{Ch}(s) - F_{NCh}(s)|$$

- Geometric mean score

This score is defined as a geometric mean of sensitivity and specificity. It aggregates those two metrics into a single value that has a range 0 to 1. It is calculated as follows:

$$g - mean = \sqrt{sensitivity * specificity}$$

- Precision (in top percentiles)

It is a rate of correct predictions among all examples predicted to belong to the event class and it is calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

- Recall (in top percentiles)

It measures how many event examples were correctly classified and is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

- The F1 Score

It is the harmonic mean of precision and recall that combines the trade-offs of those metrics that can be often conflicting. It might happen when increasing the TP for the minority class, the number of FP increases and consequently the precision is being reduced.

$$F1 - Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$

- Lift

It allows assessing a discriminatory performance of the model by measuring if the cases that are predicted to be events are at the top of the descending posterior probability ranking. The range of the posterior probability might be divided into deciles or demi-deciles. A good predictive model should predict a higher number of events in the top decile, called a lift ratio. It is calculated between the number of predicted and real events. In the following deciles, the lift ratio should monotonically decrease.

The obtained lift is often compared with overall event proportion to understand how much better the model performs in a particular decile or demi-decile

$$Lift = \frac{Success\ proportion\ in\ the\ top\ x\%\ group\ with\ the\ highest\ posterior\ probability}{Success\ proportion\ in\ the\ whole\ dataset}$$

- Gain

Shows how much better the model is performing compared to no model

$$Gain = (\%\ of\ events\ in\ decile\ /\ random\ \%\ events\ in\ decile) - 1 = Lift - 1$$

Some of the abovementioned measures are not available in SAS EM but could be easily computed. Others like G-mean and F1 score are not used in further analysis due to the lack of information in the software.

### 4.6.3. Evaluation Phases

The evaluation of the models must consider the business aspects of what information is the most valuable from the results. Consequently, the evaluation of the models had two phases.

The first phase considered the development of the models and basic configuration of the parameters to achieve a good performance. During this phase, the choice of the models was different for each method. For logistic regression the selection criterion was AIC (Akaike Information Criterion). For decision tree it was entropy in a single process or AUC between different configurations. AUC was also used to compare different configurations of random forest. Finally, for neural network, the model selection criterion was misclassification in a single process and AUC between different configurations.

The second phase picks the best model based on AUC and lift of the model. High values mean that the model not only correctly classifies the events but is also able to rank them correctly based on the posterior probabilities. Gain is used as one of the final evaluation criteria as the business interest is just a group of clients with the highest probability of having a claim. The further profiling also focuses on this high-risk group.

### 4.7. Testing Process

The general model performance on validation data might be overstated because this dataset has been used to pick the best model. Consequently, model testing needs to be performed to further evaluate how the model performs and obtain a final unbiased assessment of overall model performance. Model testing is resembling the model validation process but uses a backup testing dataset that was not used in any previous steps (Najim, 2018, p. 17). It also enables assessing if the model has good generalization ability, thus if it is overfitting. Figure 4.24 presents the complete scheme of the modeling process including backup testing.

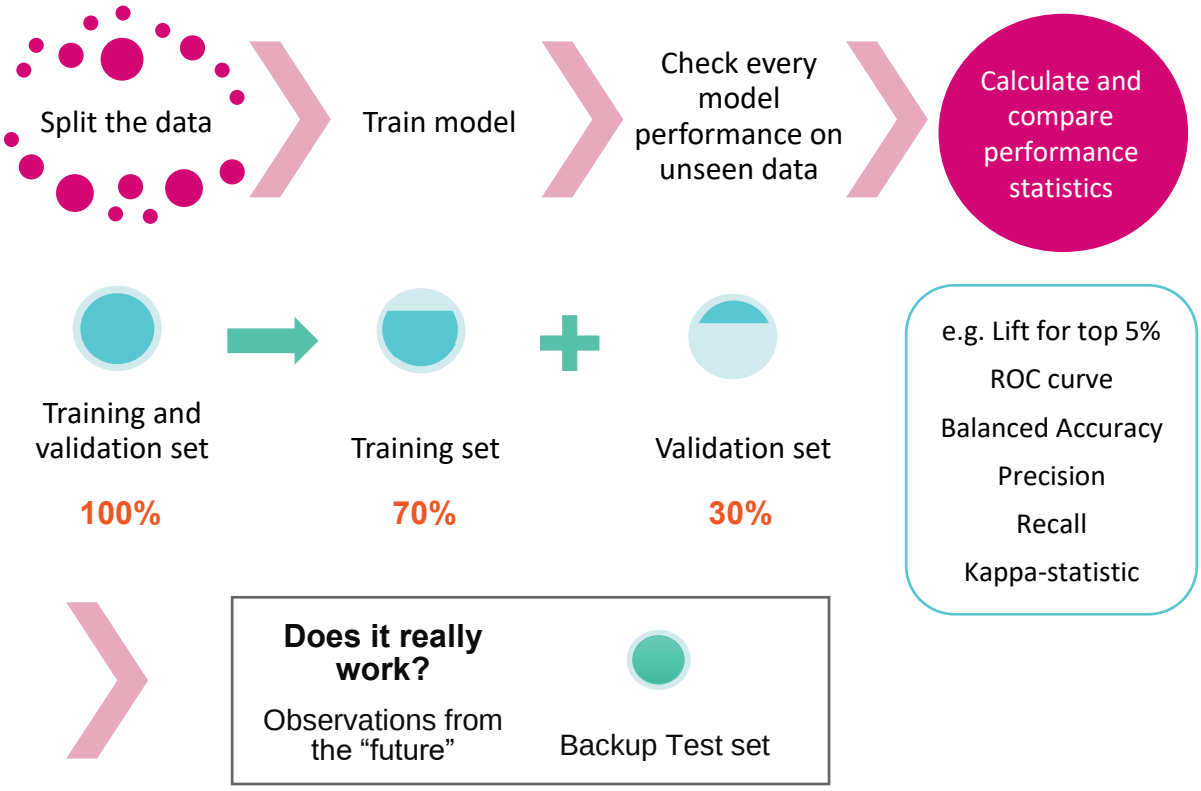


Figure 4.24 - Modeling Process Scheme

As stated before, all the presented modeling methods are exposed to the risk of overfitting but there exist many methods to overcome this problem and improve the modeling. However, this process requires deeper analysis and time-consuming investigation of the possible solutions that could not have been conducted during this project due to time constraints and the required simplicity of the processes.

# 5. RESULTS AND DISCUSSION

Given all the methodology presented, the previously mentioned supervised learning methods were applied. This chapter presents the general results of the models and discusses their performance in predicting the target. It was important for the business to compare the modeling results and insights between the claims that happened in the first 90 days of the policy with those that happened in the rest of the first annuity.

The modeling phase was carried in the same way for all 3 datasets. The following proportions in the training data were tested:

- Original proportions
- Undersampling 50/50
- Undersampling 30/70
- SMOTE 30/70
- SMOTE 30/70 + underbalanced 50/50

The best results were obtained for the original proportions and undersampling. SMOTE method did not improve model training and discriminative ability as it was expected before.

Models were trained on pre-selected variables based on the voting presented in Appendix B., as well as, on all the variables for decision tree and random forest, and chi-square selection for regression and neural networks. That is because first two methods theoretically do not need variable selection. These 3 variables sets were tested to make sure that variable selection described in chapter 4.3.2. not only restricted the dimensionality but also did not impact negatively the performance of the models.

Four algorithms were applied, namely logistic regression, decision tree, random forest, and neural networks. Some configurations of these algorithms were tested and compared. However, SAS EM software has some limitations in comparing the models that make setting the results together inflexible and sometimes time-consuming. The tested configurations with the final choice are presented below.

A linear regression node in SAS EM allows easily tuning the model. The choice was to use logistic regression with logit link function without two-factor interactions and polynomial terms. There were not included as the number of interactions would require long training time due to the high number of inputs. Additionally, the best model for each dataset was selected by the node using stepwise feature selection with the AIC.

Decision tree node in SAS allows picking different configurations for the model such as target criterion and size limitations. However, the focus was to compare different models for each target and different target proportions, so not all the possible configurations were tested as they gave similar initial results for the main targets. The configuration that was used for further comparisons is a decision tree with entropy as a target criterion, missing values to be used in the split search, maximum branch 2, maximum depth 10 and minimum categorical size 4. Then the other configurations chosen were leaf size of minimum 4, assessment method for a subtree with a lift as an assessment measure (pruning strategy), and Bonferroni adjustment before the split (unbiased selection of predictor at a split regardless to the number of possible values).

Having some promising initial results of random forest method, there were 3 main configurations tested:

- Maximum number of trees equals 100, a maximum depth of 25 for each tree within the forest,
- Maximum number of trees equals 80, maximum depth of 50 for each tree within the forest,
- Maximum number of trees equals 80, maximum depth of 25 for each tree within the forest.

The rest of the configurations was the same as the following. 60% of observations used to train a tree with, missing values to be used in a search, maximum of 30 categories in a split search, minimum category size of 5 and loss reduction as variable importance method. The best results were obtained having maximum depth of 25.

Neural Networks were not explored during this project as they could be if this phase had been given more time. The node in SAS EM gives a possibility to tune many different properties of the network or optimization process. The main method used in this project was a multilayer perceptron with default configurations. Ripley (as cited in Wielenga, 2007, p. 16) states that "Theoretically, a multilayer perceptron with one hidden layer and a sufficient number of hidden units is a universal approximator, which means that an MLP can approximate any given surface within a prescribed error range." That is why the abovementioned propertes selection was kept in the process. There have been also AutoNeural SAS node initially tested for comparison but did not give any promising results.

The final models were chosen by picking the best model in each sample of different target proportions. The most important criterium was AUC and gain for validation set. The same measurements for test set were considered in case of the models for claims within 90 days. The other criterium was a difference between AUC in training and validation set or test set if available. Thus, it was possible to capture extreme overfitting. Then, having winning model of each sample, the ones with the best performance on validation and test sets were picked as final models. They are presented below. The full table of all models taken into consideration, as well as, metrics calculated in SAS EM are presented in Appendix C.

| Cover | Days | Best Model | AUC Train | AUC Validation | AUC Test | KS Validation | Gain Validation | Gain Test | Training Dataset |
|---|---|---|---|---|---|---|---|---|---|
| **RCV** | <90 | Random Forest | 0.659 | **0.61** | 0.638 | 0.165 | **120.06** | 182.44 | original |
| | >90 | | 0.652 | 0.651 | - | 0.215 | 100.74 | - | original |
| **QIV** | <90 | Logistic Regression | 0.612 | **0.611** | 0.615 | 0.167 | **97.80** | 91.12 | level 30/70 |
| | >90 | | 0.646 | 0.631 | - | 0.184 | 116.36 | - | level 50/50 |
| **CCC** | <90 | Neural Network | 0.651 | **0.570** | 0.577 | 0.017 | **120.88** | 83.10 | original |
| | >90 | Random Forest | 0.695 | 0.620 | - | 0.176 | 115.88 | - | original |

Table 5.1 - Final Models Selection

Table 5.1 shows the final models selected for each cover with target variables of claim within 90 days and the rest of the first annuity. In case of RCV random forest method performed the best for both targets. There was a big difference in AUC between training and validation for the model of target within 90 days. However, there could be observed high gain for top 5% probabilities group on the test data. In case of QIV logistic regression was the method of both winning models. Models for both RCV and QIV have performed well on the test dataset. Neural network was the best model only for CCC claim within 90 days target but had quite poor performance on both validation and test sets. There was an evidence of overfitting. In case of CCC target for claim after more than 90 days within the first annuity, the winning model was random forest. Backup testing was performed only for the targets of having a claim within 90 days due to time constraint. The results for all the groups were quite similar regardless the method, so the winning models were not very distinct from the rest. The plots below represent how each model performs on the validation data among all the semi-deciles in the descending probability ranking of predictions.

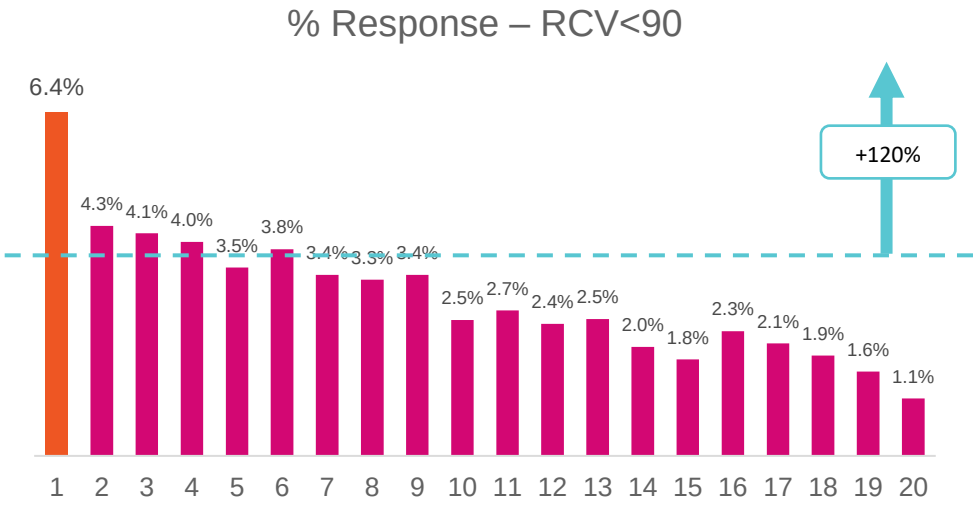


Figure 5.1 - Overall vs. predicted probability for RCV<90 days

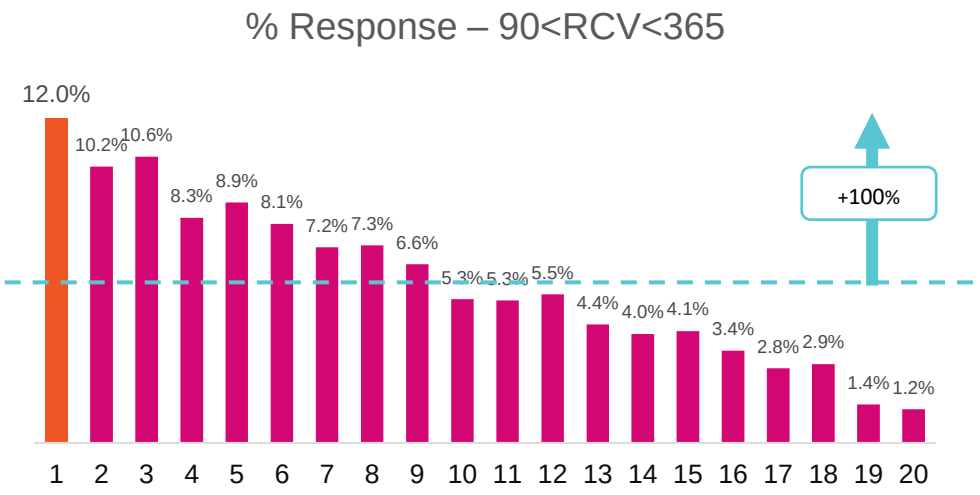


Figure 5.2 - Overall vs. predicted probability for 90<RCV<365 days

The groups in Figure 5.1 and 5.2 are divided by descending probability of RCV claim in the first 90 days. Each semi-decile in the graph has the same number of policies that is equal to 4418.

The overall probability of having a claim within 90 days is 2,9%. The winning model increases the chance of identifying a client who has claimed by more than double, having a gain of 120% in the top 5% of highest probabilities. In this group there were 6.4% claims of interest identified.

The top 10 most important variables for this model included payment frequency, transformed time with driving license, transformed age of the driver, district, phone flag, transformed number of different license plates before, transformed count of phone number repetition, grouped sales channel, grouped bonus-malus rating, and 2nd vehicle flag.

The overall probability of having a claim in the first year, but after 90 days is 6%, while the model increases this chance by 100% in the top 5% of highest probabilities. There were 12.0% claims of interest identified in this top semi-decile.

The top 10 most important variables for this model were tariff, transformed vehicle age, transformed number of different license plates, transformed time with driving license, type of bonus-malus, transformed phone number repetitions, grouped district, transformed driver's age, type of manual bonus-malus, transformed number of different cars before.

Both models' predictions vs. real events are distributed as expected. The models predict a higher number of the events in the top decile and in the following deciles, the lift ratio monotonically decreases.

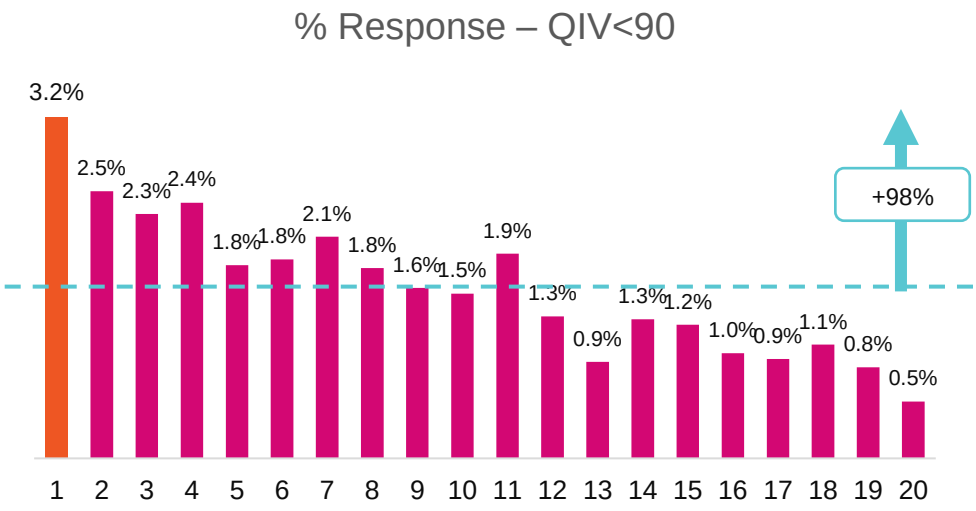


Figure 5.3 - Overall vs. predicted probability for QIV<90 days

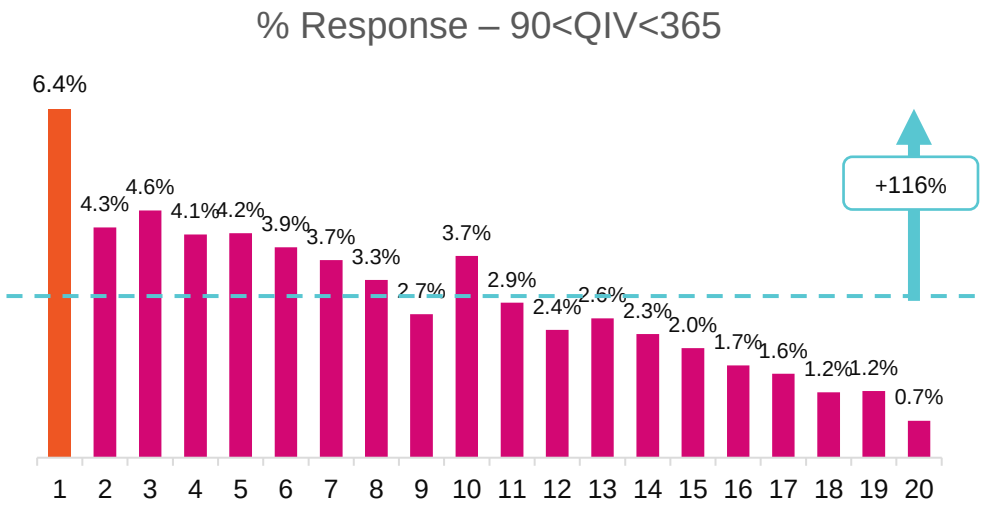


Figure 5.4 - Overall vs. predicted probability for 90<QIV<365 days

The groups in Figure 5.3 and 5.4 are divided by descending probability of QIV claim. Each semi-decile in the graph has the same number of policies that is equal to 3803.

The overall probability of having a claim within 90 days is 1,6%. By having the model, the chance of identifying a client who had claimed is almost double, having a gain of 98% in the top 5% of risky policies. This group has 3.2% claims of interest identified.

There were 17 variables used in this model, e.g. some Class DP levels, sex of driver, transformed number of motor policies canceled before, transformed age of the driver and frequency of payment.

Without using the model, the chance of finding a client with a claim in the first year, but after 90 days is 3%, while the model increases this chance by 116% in the top 5% of risky policies. There were 6.4% claims of interest identified in this top semi-decile.

There were 25 variables used in this model, e.g. some Class DP levels, 3 main levels for Tariff, type of bonus malus, driver's sex, transformed value of new car, transformed age of the driver and 8 different bodywork types.

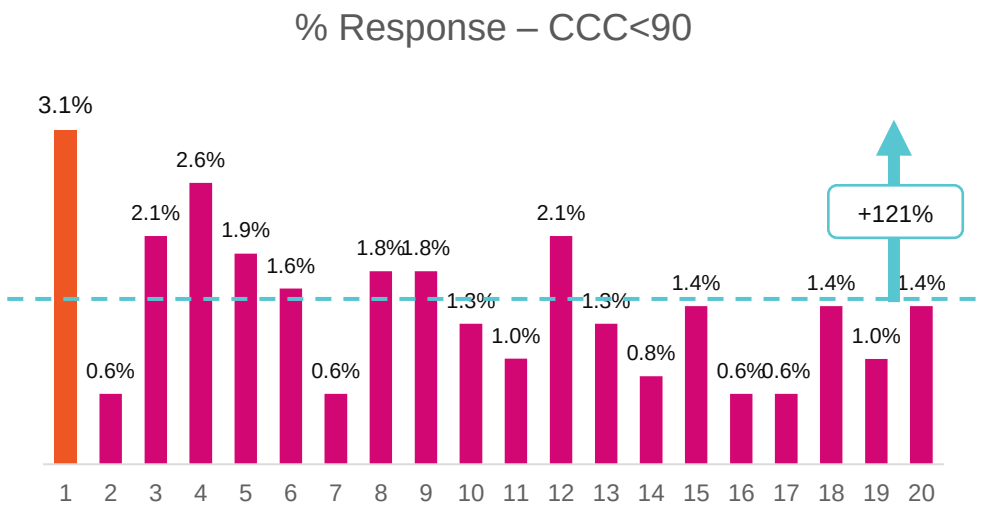


Figure 5.5 - Overall vs. predicted probability for CCC<90 days

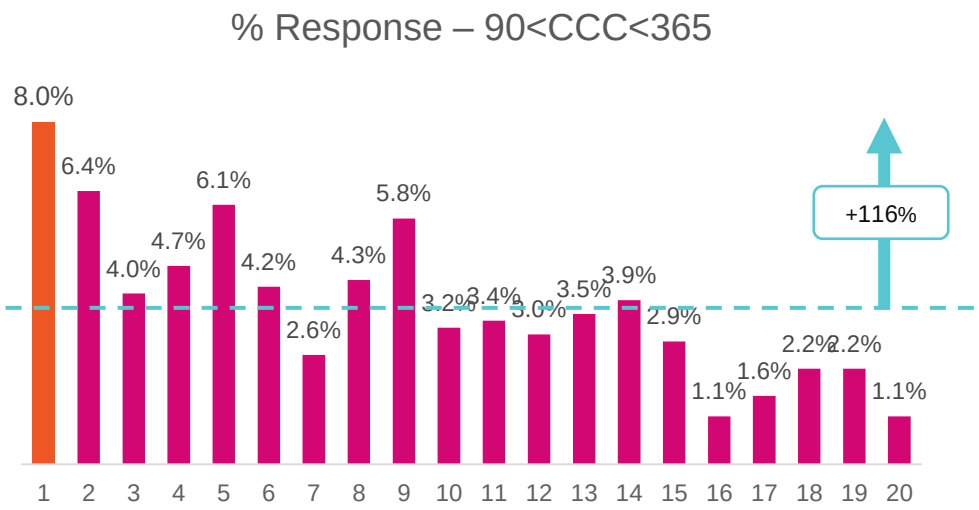


Figure 5.6 - Overall vs. predicted probability for 90<CCC<365 days

The groups in Figure 5.5 and 5.6 are divided by descending probability of CCC claim. Each semi-decile in the graph has the same number of policies that is equal to 3803.

If we pick a client without using our model, then the chances of finding a client that had claimed within 90 days is 1,45%. By having the model, the chance of identifying a client who has claimed is more than double, having a gain of 121% in the top semi-decile. In this group there were 3.1% claims of interest identified. However, the model does not perform so well in all the risk groups.

The variables used in this model include district, type of the brand, nationality of policyholder, pack, bodywork types, and bonus malus.

Without using the model, the chance of finding a client with a claim in the first year, but after 90 days is 3,7%, while the model increases this chance by 116%. In the top 5% of risky policies, there were 8% of claims of interest identified.

The top 10 most important variables for this model included Tariff, transformed age of the car, transformed number of phone repetitions, transformed deductible of one of the covers, PO option, type of bonus malus, channel, pack, and 2$^{nd}$ vehicle indicator.

The models for CCC claim are not performing well. It is expected that the models should predict a higher number of the events in the top decile and in the following deciles, the lift ratio should monotonically decrease. However, this behavior is not observed for the abovementioned models.

# 6. CONCLUSIONS

The primary goal of this project was to provide predictive models of 3 different types of claims, as well as, profiling of the riskiest groups. The models were performed for both targets of having a claim within first 90 days of the policy or after 90 days but within first year. The idea was to compare if there are significant differences between those risk groups.

The machine learning methods had quite low performance. It is probably caused by imbalanced data problem. Another issue might be overlapping of claim characteristics. The dataset was split by indicator if the client had the given cover, but the claims of other type were not excluded from the given dataset. In case the claiming clients had similar characteristics, the noise was introduced. It is suggested to test a general model in the future with target of any claim. It would be also worth to test the stability of the models as the winning choices were not clearly distinct from the rest regarding their performance.

In order to answer the question, if there are differences between the model results and consequently profiling of each type of claimers the most important variables were plotted on the given semi-deciles to enable interpretation. Some of the examples are presented below. The rest was not shared due to confidentiality reasons.

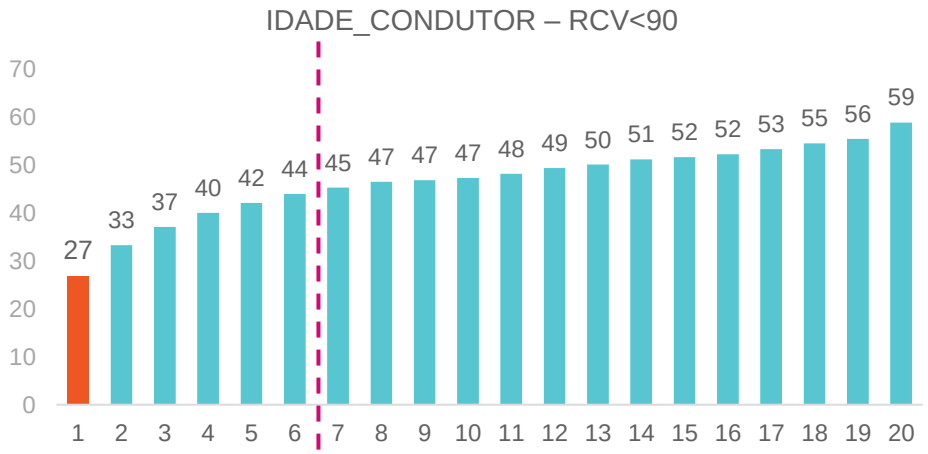


Figure 6.1 – Average Driver's Age (RCV<90 days)

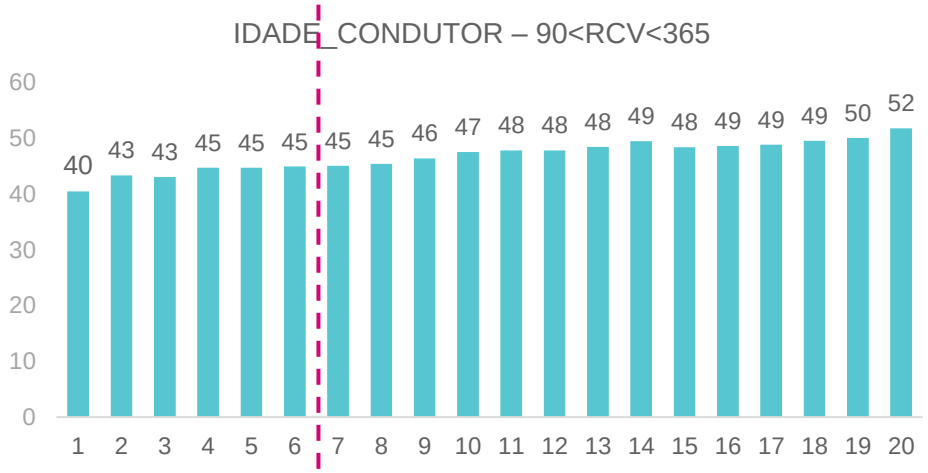


Figure 6.2 – Average Driver's Age (90<RCV<365 days)

Figure 6.1 shows that the risk of RCV claim within 90 days is higher when the driver is younger. Drivers having higher risk of claim within 90 days are on average younger than those of high risk of having claim after 90 days presented in Figure 6.2.

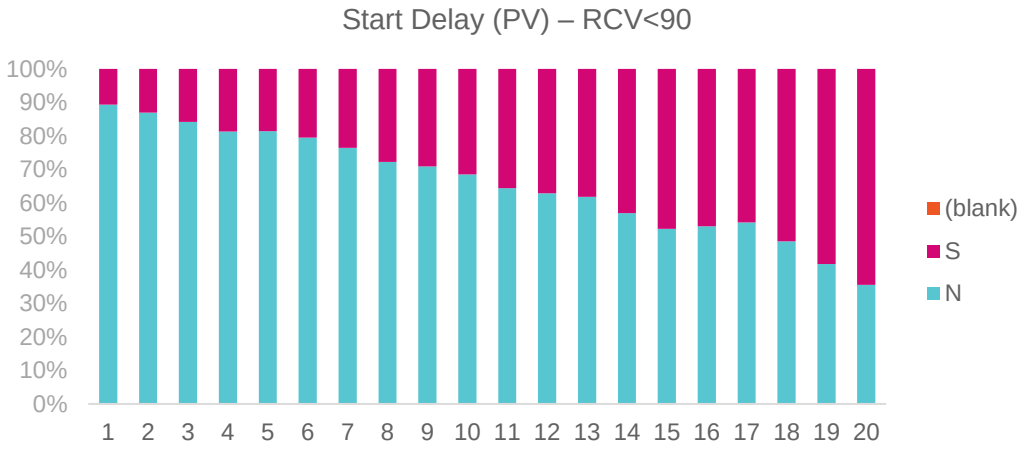


Figure 6.3 - Delay in start of the policy (RCV<90 days)

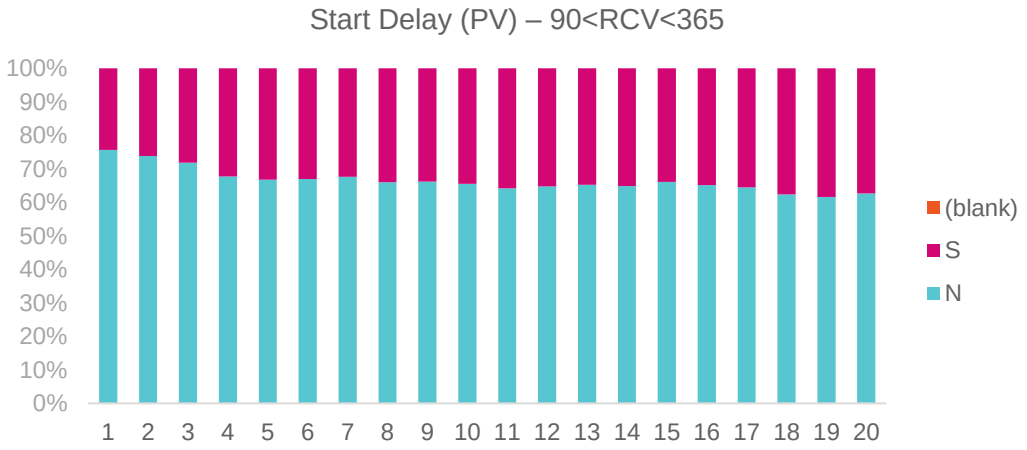


Figure 6.4 - Delay in start of the policy (90<RCV<365 days)

The risk of claim increases with no delay between start and issue date. Comparing the high risk of claims within the first 90 days and after 90 days, there are more policies with the delay in the latter.

This type of analsysis was performed for all the most important variables for each claim. Some of the variables brought the information about the difference between the groups. Other were not so clear in interpretation. The insights were the easiest to find for RCV, and the hardest for CCC.

Additionaly, the results were presented together with fraud data. There is some relation as presented in Figure 6.5 but it might be explained in a way that without having a claim there is no possibility for fraud. Thus, the separate model should be performed on fraud, as initially planned in the project. The conclusion cannot be done without confirming it.

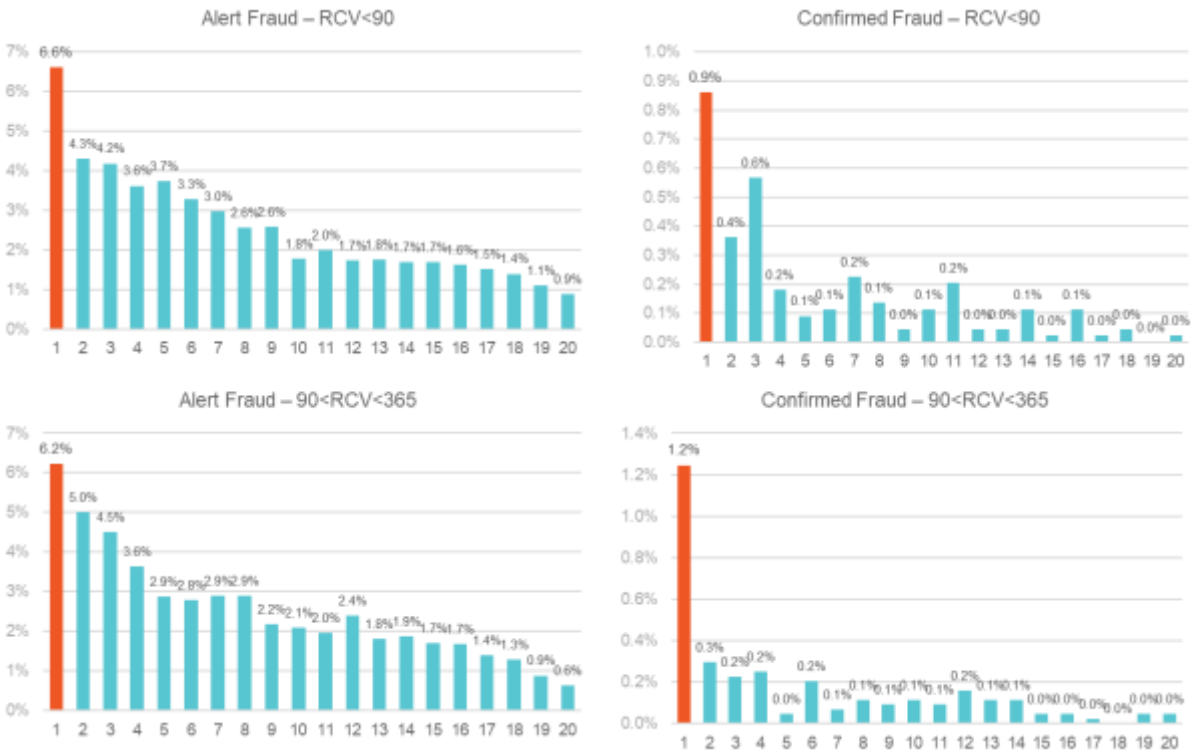


Figure 6.5 - RCV relation to fraud

Dealing with an unbalanced dataset makes the classification task difficult to perform, therefore an under-sampling technique and SMOTE needed to be explored. Unfortunately, SMOTE did not bring better results in any of the datasets. The performance of the models was rather disappointing. However, it was possible to obtain some results and focus the analysis on the group of the highest probabilities in the ranking.

Since the implementation was not performed and designed, it was not possible to measure directly the value of the project that was brought to the company. However, the project gave interesting insights and guidelines for further investigation and highlighted some differences for the riskiest groups. Moreover, it can be used as an inspiration for other projects and an example in the future to improve the project planning. The limitations and recommendations for future work have been described in greater detail in the next chapter.

In conclusion, this project was a good opportunity to apply the theory in a company scenario. However, the problem should be better defined at the beginning of the project. Also, the planning should get more attention to avoid fundamental changes during the work.

# 7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This chapter outlines the limitations met during the development of the project and proposes analytical recommendations for further work. It also describes the next steps that should be taken.

The analysis was limited in several ways. First, as the data extraction and preparation were the longest parts of the project, it is highly recommended that the database will be improved, and more historical variables will be added. Also including data from the external sources should be considered. The other main limitation was a time constraint.

This research has raised many questions in need of further investigation that should be done by the appropriate teams. Future steps should aim at the validation of the existing fraud detection models, as well as, reviewing the existing rules for alerts and blockages. One of the most important steps that need to be done is the decision on adjusting the portfolio to reduce losses based on the riskiest profiles and calculated KPIs. That is because some of the conventional KPIs are calculated for the whole portfolio in the rolling year and this analysis takes into consideration only the moment when the policy starts.

From an analytical point of view, it would be recommended to try some more advanced models, as well as, more configurations of the parameters to find a more reliable solution. One possible improvement could be done by testing more methods that deal with unbalanced data. It would be worthy to try methods that deal with nominal variables like SMOTE-NC, SMOTE-N, etc.

Furthermore, the same analysis could be done for commercial lines and the results compared. This would allow investigating if there are some major differences between those two portfolios. Eventually, claims without the cover could be addressed as they were excluded from this analysis. That also involves the claims that were not covered in the policy at the beginning but were included later. That suggests some more dynamic analysis could be considered.

# 8. BIBLIOGRAPHY


Akindaini, B. (2017). Machine learning applications in mortgage default prediction. Retrieved from https://trepo.tuni.fi/bitstream/handle/10024/102533/1513083673.pdf?sequence=1&isAllowed=y

Associação Portugesa de Seguradores. (2016). *Contributos para a Análise da Percepção da Fraude [Contribution to Analysis of Fraud Perception].* Observatório de Economia e Gestão de Fraude.

Boardman, J., Biron, K., & Rimbey, R. (2018). Mitigating The Effects Of Class Imbalance Using SMOTE and TOMEK Link Undersampling In SAS®. Retrieved from https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/3604-2018.pdf

Bond, A., & Stone, M. (2004). How the automotive insurance claims experience affects customer retention. *Journal of Financial Services Marketing, 9*(2), 160-171.

Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. *Journal of American society for information science*, 12-19.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626-4636.

Burri, R. D., Burri, R., Bojja, R. R., & Buruga, S. R. (2019). Insurance Claim Analysis Using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering, 8*(6S4), 577-582.

Chang, Y. C., & Nelson, H. (2017). Data Opportunities in Insurance. *Silicon Valley Data Science*.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon, & L. Rokach, *The Data Mining and Knowledge Discovery Handbook* (pp. 853-867).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence, 16*, 321-357. doi:https://doi.org/10.1613/jair.953

David, M., & Jemna, D. (2015). Modeling the Frequency of Auto Insurance Claims By Means of Poisson and Negative Binomial Models.

De Jong, P., & Heller, G. Z. (2008). Generalized Linear Models for Insurance Data. doi:https://doi.org/10.1017/CBO9780511755408

Douzas, G., & Bação, F. (2017). Geometric SMOTE: Effective oversampling for imbalanced learning through a geometric extension of SMOTE.

Douzas, G., Bação, F., & Last, F. (2018). Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-means and SMOTE. *Inf. Sci., 465*, 1-20.

Fawcett, T. (2016). Learning from Imbalanced Classes. *Silicon Valley Data Science*. Retrieved from https://www.svds.com/learning-imbalanced-classes/

Gilenko, E. V., & Mironova, E. A. (2017). Modern claim frequency and claim severity models: An application to the Russian motor own damage insurance market. *Cogent Economics & Finance, 5*(1). doi:10.1080/23322039.2017.1311097

Guzman, L. (2015). Data sampling improvement by developing SMOTE technique in SAS. *SAS Global Forum 2015 Conference*, (p. 19). Retrieved from http://support.sas.com/resources/papers/proceedings15/3483-2015.pdf

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 220-239.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284.

Hosmer, D. W., & Lameshow, S. (1989). Applied Logistic Regression. The Multiple Logistic Regression Model. 25-37.

Hossin, M. B., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 5*(2), 1-11. Retrieved from https://pdfs.semanticscholar.org/6174/3124c2a4b4e550731ac39508c7d18e520979.pdf

Last, F., Douzas, G., & Bação, F. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE. Retrieved from https://arxiv.org/pdf/1711.00837v2.pdf

Lexi, H. (2018). Top 10 tips for SAS Enterprise Miner. Retrieved from https://blogs.sas.com/content/sgf/2018/05/29/top-ten-tips-for-sas-enterprise-miner-based-on-20-years-experience/

Mitchell, T. (1997). *Machine Learning.*

Najim, M. (2018). Claim Analytics: A Litigation Prediction Case Study. *SAS Global Forum 2018.* Retrieved from https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2504-2018.pdf

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106. doi:https://doi.org/10.1007/BF00116251

Rejda, G. E., & McNamara, M. J. (2017). Principles of Risk Management and Insurance.

Sarma, K. S. (2007). Variable Selection and Transformation of Variables in SAS® Enterprise Miner™ 5.2. Retrieved from https://www.lexjansen.com/nesug/nesug07/sa/sa17.pdf

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide: The MI Procedure. Retrieved from https://support.sas.com/documentation/onlinedoc/stat/141/mi.pdf

SAS Institute Inc. (2018). *SAS® Enterprise Miner™ 15.1: Reference Help.* Retrieved from https://documentation.sas.com/api/docsets/emref/15.1/content/emref.pdf?locale=en#nameddest=p0qiq0a4vnebuzn16v8fzossk4gp

Shapiro, M. (2016). CMISS® the SAS® Function You May Have Been "MISSING". Retrieved from https://analytics.ncsu.edu/sesug/2016/RV-201_Final_PDF.pdf

Society of Actuaries. (2019). Machine-Learning Methods for Insurance Applications. Retrieved from https://www.soa.org/globalassets/assets/files/resources/research-report/2019/machine-learning-methods.pdf

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*, 427-437.

Wang, R., Lee, N., & Wei, Y. (2015). A Case Study: Improve Classification of Rare Events with SAS® Enterprise Miner™. 1-12. Retrieved from https://support.sas.com/resources/papers/proceedings15/3282-2015.pdf

Werner, G., & Modlin, C. (2016). *Basic Ratemaking.* Willis Towers Watson. Retrieved from https://www.casact.org/library/studynotes/Werner_Modlin_Ratemaking.pdf

Wicklin, R. (2016). Examine patterns of missing data in SAS. Retrieved from https://blogs.sas.com/content/iml/2016/04/18/patterns-of-missing-data-in-sas.html

Wielenga, D. (2007). Identifying and Overcoming Common Data Mining Mistakes. *SAS Global Forum*, (p. 20).

# 9. APPENDIX

**A.** List of all data tables

The data came from 41 tables including:

- HIS_AUT_CUER_CICLO – all the historical data of motor policies including some variables about the vehicle insured;
- HIS_CUERPOLIZA – all the basic historical data of all the policies;
- HIS_AUT_m002 – financial data about the policies, e.g. premiums and capital amounts;
- Age_clasificaciones – information about the agents and sales channels;
- Age_tipos_clasificaciones – codes of the agent types and sales channels;
- Eurotax 0-4 – excel files with the car codes from national Eurotax system (the codes are required to obtain car standard characteristics);
- 3 excels of tylacodes (the same as from Eurotax) from the company
- AUT_VEICULOS – technical characteristics of the car;
- CUERPOLIZA_COMUN – information about the partnership protocols;
- HIS_PERSONAS – nationality, sex and birth date of the main driver and policyholder (historical data);
- PERSONAS - nationality, sex and birth date of the main driver and policyholder (actual data);
- HIS_AUT_CUERPOLIZA – retirement flag, transfer flag, years with insurance, aggregated data of claims, main discounts;
- HIS_PERSONAS_PROD – ID of the policyholder and main driver and home address (historical data);
- PESSOAS_FUSIONADAS – IDs of the clients that should be merged because of entity repetitions;
- HIS_PERSONAS_DOCUMENTOS – tax identification number (historical data);
- PERSONAS_DOCUMENTOS – tax identification number (actual data);
- PERSONAS_CARTA_CONDUCAO – driving license data;
- HIS_PERSONAS_DOMICILIOS – postal code of the policyholder and main driver (historical data);
- PERSONAS_DOMICILIOS - postal code of the policyholder and main driver (actual data);
- 1 excel with identification of district and urban/rural location by postal code;
- AUT_AXA_CUERPOLIZA_ATRIBUTOS – bonus malus levels (discounts);
- MO_PESSOA_AXA – type of entity (individual/company), nationality, marital status, phone number, email;
- PERSONAS_PROD - ID of the policyholder and main driver (actual data);
- CUERPOLIZA – policy number, start and end date (actual data);
- AUT_M002_DES_AGRAV – information about the second car insured;

- Ry_12m_final1118 – costs, exposure, number of claims during the rolling year;
- SIN_AUTOCOB_ACTUAL – all information about the claims;
- AGE_AGENTES – ID of the agent;
- SIN_EXPEDIENTES – basic info about fraud;
- SIN_INTERVINIENTES – entities involved in the claim;
- HIS_PERSONAS_BLOQUEIO_SNV – blocked entities;
- SIN_EXPEDIENTES_MOTIVOS_FRAUDE – fraud alerts;
- SIN_EXPEDIENTES_VEIC_AUT – third part car info;
- SIN_EXPEDIENTES_PERIT_AUT – claim handlers' info, estimated value;
- AUT_CERTIFICADO_PROVISORIO – provision plate certificate;
-

**B.** Variables Selection Tables

ORIGINAL - RCV

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **no HP** | **HP** | **ALL** | **NO HP** | **HP** |
| CANAL | * | * | | | * | |
| CDSUBRAMO | | * | | | | |
| COND_BLOCKED_EVER | | * | | | * | |
| CVE_BONUS_MALUS | * | * | | * | * | |
| CVE_CAIXA | | * | | | | |
| CVE_CAIXA_VELOC | | * | | | * | |
| CVE_CARROCARIA | * | * | | * | * | |
| CVE_CATEGORIA | | * | | | * | |
| CVE_CLASSE_DP | | * | | | | |
| CVE_COD_RC | | * | | | | |
| CVE_COMBUSTIVEL | | * | | | | |
| CVE_EXTENSAO | | * | | | | |
| CVE_NACIONALIDADE | | * | | | | |
| CVE_PACK | * | * | | | * | |
| CVE_PROTOCOLO_nomiss | * | * | | | | |
| CVE_TIPO_VEICULO | | * | | | | |
| CVE_TRANSFERENCIA | | * | | | * | |
| EMAIL_binary | | * | | | | |
| EST_CIVIL | | * | | | * | |
| Email_cond_rep | | | | | | |
| FOV_ALARME_SEGURANCA | | | | | | |
| FOV_CONTROLO_TRAVAGEM | | * | | | | |
| FOV_JUBILADO | | * | | | * | |
| FPROCESO_time | * | * | | * | * | |
| Franquia_euros_qiv | | * | | | | |
| IDADE_VEICULO | | * | | | | |
| IMP_NUM_CV | * | * | | | | |
| IMP_NUM_TARA | | * | | | | |
| IMP_QUA_VALOR_NOVO | | * | | | * | |
| IMP_REP_Diff_cov_issue | * | * | | * | * | |
| IMP_REP_IDADE_CARTA_CONDUTOR_NEW | * | * | | * | * | |
| IMP_REP_IDADE_CONDUTOR_NEW | * | * | | | * | |
| IMP_REP_PESOBRUT | | | | | | |
| INV_Email_cond_rep | | * | | | | |
| INV_Email_rep | | | | | | |
| INV_FPROCESO_time | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| INV_IMP_REP_IDADE_CARTA_CONDUTOR | | | | | | |
| INV_IMP_REP_IDADE_CONDUTOR_NEW | | * | | | * | |
| INV_No_MATRICULA | * | * | | | * | |
| INV_No_NUM_VEICUL | * | * | | | * | |
| INV_No_anulado_before | | | | | | |
| INV_Phone_cond_rep | | * | | | * | |
| INV_Phone_rep | | * | | | * | |
| INV_franq_P02 | | * | | | | |
| LG10_capital_CCC | | * | | | | |
| LG10_capital_FN | | * | | | | |
| LOG_IMP_QUA_VALOR_NOVO | * | * | | | * | |
| LOG_IMP_REP_IDADE_CARTA_CONDUTOR | * | * | * | * | * | * |
| LOG_No_CP | * | * | | | * | |
| LOG_capital_Extras | | * | | | | |
| LOG_capital_IRE | | * | | | | |
| NACIONALIDADE | | * | | | | |
| NUM_CC | | | | | | |
| NUM_LUGARES | | * | | | | |
| No_CP | | * | | | * | |
| No_MATRICULA | | * | | | * | |
| No_NUM_VEICUL | | * | | | * | |
| No_anulado_before | | | | | * | |
| No_anulado_before_auto | | * | | | | |
| No_cond_changes | | | | | | |
| No_tomador_changes | | * | | | | |
| OPCAO_BG | | * | | | | |
| OPCAO_BG_binary | | | | | | |
| OPCAO_EXTRAS | * | * | * | | | |
| OPCAO_PO | * | * | | | * | |
| OPCAO_PO_binary | | * | | | | |
| OPCAO_VEIC | | * | | | | |
| OPCAO_VEIC_binary | | * | | | | |
| OPT_IMP_REP_Diff_cov_issue | | | | | | |
| PHONE_binary | * | * | * | * | * | |
| PV | * | * | * | * | * | * |
| Phone_cond_rep | | * | | | * | |
| Phone_rep | | * | | | * | |
| SQRT_IMP_QUA_VALOR_NOVO | | | | | | |
| SQRT_IMP_NUM_CV | * | * | | | | |
| SQR_IMP_REP_Diff_cov_issue | * | | | | * | |
| SQRT_No_CP | | | | | | |
| SQRT_capital_CCC | | | | | | |
| SQRT_capital_FN | | | | | | |
| SQRT_capital_FR | | | | | | |
| SQRT_capital_IRE | | | | | | |
| SQRT_capital_RS | | | | | | |
| SQR_IDADE_VEICULO__IDADE_VEICULO | | * | | | | |
| SQR_IMP_REP_Diff_cov_issue | | * | | | | |
| Sexo_ch__Sexo_ch_ | | | | | | |
| Sexo_tom | | | | | | |
| TG_CANAL | | * | | | * | |
| TG_CDSUBRAMO | | * | | | | |
| TG_CVE_BONUS_MALUS | * | * | | | * | |
| TG_CVE_CARROCARIA | | * | | | * | |
| TG_CVE_CATEGORIA | * | * | | | * | |
| TG_CVE_CLASSE_DP | | * | | | | |
| TG_CVE_COD_RC | | * | | | | |
| TG_CVE_COMBUSTIVEL | | * | | | | |
| TG_CVE_PACK | | * | | | | |
| TG_CVE_TIPO_SEGURO | | * | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| TG_CVE_TIPO_VEICULO | | * | | | | |
| TG_CVE_TRANSFERENCIA | | | | | * | |
| TG_NUM_LUGARES | | * | | | | |
| TG_No_cond_changes | | * | | | | |
| TG_No_tomador_changes | | * | | | | |
| TG_OPCAO_EXTRAS | | * | | | | |
| TG_OPCAO_PO | | * | | | * | |
| TG_OPCAO_VEIC | | * | | | | |
| TG_Tipo_Marca | * | * | | | * | |
| TG_ZONA_RC | * | * | | * | * | |
| TG_capital_qiv | | * | | | | |
| TG_formacob | | | | | | |
| TG_opcao_P11 | | * | | | | |
| TOM_BLOCKED_EVER | * | * | * | * | * | * |
| TOM_COND | | | | | | |
| Tariff | * | * | | | | |
| Tipo_Marca | * | * | | * | * | |
| URBAN_RURAL | * | * | * | | * | |
| ZONA_RC | * | * | | * | * | |
| capital_CCC | | | | | * | |
| capital_Extras | | * | | | | |
| capital_FN | | * | | | | |
| capital_FR | | * | | | | |
| capital_IRE | | * | | | | |
| capital_P30 | | * | | | | |
| capital_RS | | * | | | | |
| capital_qiv | | * | | | | |
| flag_2a_viatura | | * | | | * | |
| flag_com | | | | | | |
| formacob | | * | | | | |
| formapa | * | * | * | * | * | * |
| franq_P02 | * | * | | | | |
| franq_P04 | | | | | | |
| opcao_P11 | | * | | | | |
| tipo_manual | | * | | | | |
| vip | | * | | | | |
| TOTAL | 32 | 104 | 7 | 13 | 48 | 4 |

STANDARDIZED - RCV

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **NO HP** | **HP** | **ALL** | **NO HP** | **HP** |
| CANAL | * | * | | | * | |
| CDSUBRAMO | | * | | | | |
| CODIGO | | | | | | |
| COND_BLOCKED_EVER | | * | | | * | |
| CVE_BONUS_MALUS | * | * | | * | * | |
| CVE_CAIXA | | * | | | | |
| CVE_CAIXA_VELOC | | * | | | | |
| CVE_CARROCARIA | | * | | * | * | |
| CVE_CATEGORIA | * | * | | | | |
| CVE_CLASSE_DP | | * | | | | |
| CVE_COD_RC | | * | | | | |
| CVE_COMBUSTIVEL | | * | | | | |
| CVE_NACIONALIDADE | | * | | | | |
| CVE_PACK | * | * | | * | * | |
| CVE_PROTOCOLO_nomiss | * | * | | | | |
| CVE_TIPO_VEICULO | | * | | | | |
| CVE_TRANSFERENCIA | | * | | | * | |
| EMAIL_binary | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| EST_CIVIL | | * | | | * | |
| EXP_STD_FPROCESO_time | * | * | | | * | |
| EXP_STD_IMP_REP_PESOBRUT | | | | | | |
| EXP_STD_NUM_CC | | | | | | |
| FOV_CONTROLO_CONDUCAO | | * | | | | |
| FOV_CONTROLO_TRAVAGEM | | * | | | | |
| FOV_JUBILADO | | * | | | * | |
| Franquia_euros_qiv | | * | | | | |
| INV_STD_Email_rep | | * | | | | |
| INV_STD_IMP_REP_IDADE_CARTA_COND | * | * | * | * | * | * |
| INV_STD_IMP_REP_IDADE_CONDUTOR_N | * | * | | | * | |
| INV_STD_IMP_QUA_VALOR_NOVO | * | * | | | * | |
| INV_STD_No_MATRICULA | * | * | | | * | |
| INV_STD_No_NUM_VEICUL | * | * | | | * | |
| INV_STD_No_anulado_before | | | | | | |
| INV_STD_Phone_cond_rep | | * | | | * | |
| INV_STD_Phone_rep | | * | | | * | |
| INV_STD_capital_FN | | * | | | | |
| INV_STD_capital_FR | | * | | | * | |
| INV_STD_capital_IRE | | * | | | | |
| INV_STD_capital_RS | | * | | | | |
| LOG_STD_capital_CCC | | * | | | | |
| NACIONALIDADE | | * | | | | |
| NUM_LUGARES | | * | | | | |
| No_cond_changes | | * | | | | |
| No_tomador_changes | | * | | | | |
| OPCAO_BG_binary | | * | | | | |
| OPCAO_EXTRAS | * | * | * | | | |
| OPCAO_PO | * | * | | * | * | |
| OPCAO_VEIC | | * | | | | |
| OPCAO_VEIC_binary | | * | | | | |
| OPT_STD_IMP_REP_Diff_cov_issue | | | | | | |
| PHONE_binary | * | * | * | | * | |
| PV | * | * | * | * | * | * |
| Retro | | | | | | |
| SQRT_STD_IMP_NUM_CV | * | * | | | | |
| SQRT_STD_IMP_QUA_VALOR_NOVO | * | | | | | |
| SQRT_STD_No_CP | | * | | | * | |
| SQR_STD_FPROCESO_time | | | | | | |
| SQR_STD_IDADE_VEICULO__IDADE_VEI | | * | | | | |
| SQR_STD_IMP_REP_Diff_cov_issue | * | * | | * | * | |
| STD_FPROCESO_time | | | | | * | |
| STD_IDADE_VEICULO__IDADE_VEICULO | | | | | * | |
| STD_IMP_NUM_CV | | * | | | | |
| STD_IMP_NUM_TARA | | * | | | | |
| STD_IMP_QUA_VALOR_NOVO | | * | | | * | |
| STD_IMP_REP_Diff_cov_issue | * | * | | | * | |
| STD_IMP_REP_IDADE_CARTA_CONDUTOR | | * | | | * | |
| STD_IMP_REP_IDADE_CONDUTOR_NEW | | * | | * | * | |
| STD_IMP_REP_PESOBRUT | | | | | | |
| STD_No_CP | | * | | | * | |
| STD_No_MATRICULA | * | * | | | * | |
| STD_No_NUM_VEICUL | | * | | | * | |
| STD_No_anulado_before_auto | | * | | | | |
| STD_Phone_cond_rep | | * | | | * | |
| STD_Phone_rep | | * | | | * | |
| STD_capital_CCC | | | | | * | |
| STD_capital_Extras | * | * | | | | |
| STD_capital_FN | | * | | | | |
| STD_capital_FR | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| STD_capital_IRE | | | | | | |
| STD_capital_RS | | * | | | | |
| STD_franq_P02 | | | | | | |
| STD_franq_P04 | | | | | | |
| Sexo_ch__Sexo_ch_ | | * | | | | |
| Sexo_tom | | | | | | |
| TG_CANAL | | * | | | * | |
| TG_CDSUBRAMO | | * | | | | |
| TG_CVE_BONUS_MALUS | * | * | | | * | |
| TG_CVE_CARROCARIA | | * | | | | |
| TG_CVE_CATEGORIA | * | * | | | | |
| TG_CVE_CLASSE_DP | | * | | | | |
| TG_CVE_COD_RC | | * | | | | |
| TG_CVE_COMBUSTIVEL | | * | | | | |
| TG_CVE_NACIONALIDADE | | * | | | | |
| TG_CVE_PROTOCOLO_nomiss | | | | | | |
| TG_CVE_PACK | | * | | | | |
| TG_CVE_TRANSFERENCIA | | * | | | * | |
| TG_NACIONALIDADE | | * | | | | |
| TG_NUM_LUGARES | | * | | | | |
| TG_No_cond_changes | | * | | | | |
| TG_OPCAO_BG | | * | | | | |
| TG_OPCAO_PO | | * | | | | |
| TG_OPCAO_VEIC | | | | | | |
| TG_Tipo_Marca | | * | | | * | |
| TG_ZONA_RC | * | * | | * | * | |
| TG_capital_qiv | | * | | | | |
| TG_formacob | | | | | | |
| TOM_BLOCKED_EVER | * | * | * | * | * | * |
| TOM_COND | | * | | | | |
| Tariff | * | * | | | | |
| Tipo_Marca | * | * | | * | * | |
| URBAN_RURAL | * | * | * | | * | |
| ZONA_RC | * | * | | * | * | |
| capital_qiv | | * | | | | |
| flag_2a_viatura | * | * | | | * | |
| formacob | | * | | | | |
| formapa | * | * | * | * | * | * |
| opcao_P11 | | * | | | | |
| opcao_P11_binary | | | | | | |
| tipo_bm | | * | | | | |
| tipo_manual | | * | | | | |
| TOTAL | 31 | 98 | 7 | 13 | 45 | 4 |

ORIGINAL - QIV

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **NO HP** | **HP** | **ALL** | **NO HP** | **HP** |
| CANAL | | | | | | |
| CVE_BONUS_MALUS | | | | * | * | |
| CVE_CARROCARIA | | * | | | | |
| CVE_CATEGORIA | | * | | | | |
| CVE_CLASSE_DP | * | * | | * | * | * |
| CVE_COD_RC | * | | | * | * | |
| CVE_COMBUSTIVEL | | | | * | * | * |
| CVE_PACK | * | | | | | |
| CVE_PROTOCOLO_nomiss | * | * | * | | | |
| CVE_TIPO_VEICULO | | * | | | | |
| EST_CIVIL | * | | | | | |
| Email_cond_rep | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| FOV_JUBILADO | | | | * | | * |
| FPROCESO_time | | * | | | | |
| IDADE_VEICULO__IDADE_VEICULO_ | | | | | | |
| IMP_NUM_CV | * | * | | | | |
| IMP_NUM_TARA | | | | | | |
| IMP_QUA_VALOR_NOVO | * | | | | | |
| IMP_REP_Diff_cov_issue | * | * | | | | |
| IMP_REP_IDADE_CARTA_CONDUTOR_NEW | * | | | | | |
| IMP_REP_IDADE_CONDUTOR_NEW | * | * | * | * | * | * |
| IMP_REP_PESOBRUT | * | | | | | |
| LG10_IMP_QUA_VALOR_NOVO | * | | | | | |
| LOG_Email_rep | | | | | | |
| LOG_IMP_REP_PESOBRUT | | | | | | |
| LOG_No_MATRICULA | | | | | | |
| NACIONALIDADE | | | | | | |
| NUM_CC | | | | | | |
| No_CP | * | * | | | | |
| No_MATRICULA | | | | * | * | |
| No_NUM_VEICUL | * | | | * | | * |
| No_anulado_before | * | * | | | | |
| No_anulado_before_auto | * | | | | | |
| NUM_LUGARES | | | | | | |
| OPCAO_PO | * | | | | * | |
| OPT_NUM_CC | * | | * | * | | * |
| PV | | | | | | |
| SQRT_IMP_NUM_CV | * | | | | | |
| SQRT_IMP_NUM_TARA | | | | | | |
| SQRT_IMP_REP_IDADE_CARTA_CONDUTO | | | | | | |
| SQR_IMP_REP_Diff_cov_issue | * | | | | | |
| Sexo_tom | | | | | | |
| TG_CVE_BONUS_MALUS | | | | | | |
| TG_CVE_COD_RC | | | | | | |
| TG_Tipo_Marca | * | | | | | |
| TG_ZONA_RC | | | | * | * | |
| Tariff | | | | | | |
| Tipo_Marca | * | * | | * | * | |
| ZONA_RC | * | * | | | * | |
| capital_FR | | | | | | |
| formacob | | | | | * | |
| formapa | * | | * | * | * | * |
| tipo_bm | | | | | | |
| tipo_manual | | | | | | |
| TOTAL | 24 | 13 | 4 | 12 | 12 | 7 |

STANDARDIZED - QIV

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **NO HP** | **HP** | **ALL** | **NO HP** | **HP** |
| CANAL | | | | | | |
| CVE_BONUS_MALUS | | | | * | * | |
| CVE_CARROCARIA | * | * | | | | |
| CVE_CATEGORIA | * | * | | | | |
| CVE_CLASSE_DP | * | * | | * | * | * |
| CVE_COD_RC | * | | | * | * | |
| CVE_COMBUSTIVEL | | | | * | * | * |
| CVE_PACK | | | | | * | |
| CVE_PROTOCOLO_nomiss | * | * | * | | | |
| CVE_TIPO_VEICULO | | * | | | | |
| EST_CIVIL | * | | | | | |
| FOV_JUBILADO | | | | * | | * |

| | | | | | | |
|---|---|---|---|---|---|---|
| INV_STD_Email_cond_rep | | | | | | |
| INV_STD_IDADE_VEICULO__IDADE_VEI | | | | | | |
| LG10_STD_IMP_NUM_CV | | | | | | |
| LG10_STD_IMP_NUM_TARA | * | | | | | |
| LG10_STD_IMP_QUA_VALOR_NOVO | * | | * | | | |
| LG10_STD_IMP_REP_PESOBRUT | | | | | | |
| NACIONALIDADE | | | | | | |
| NUM_LUGARES | | | | | | |
| OPCAO_PO | * | | | | | |
| OPT_STD_NUM_CC | * | | | * | | * |
| SQRT_STD_FPROCESO_time | * | * | * | | | |
| SQRT_STD_IMP_NUM_CV | * | * | | | | |
| SQRT_STD_IMP_REP_IDADE_CARTA_CON | * | | | | | |
| SQRT_STD_No_anulado_before | * | * | | | | |
| SQRT_STD_No_anulado_before_auto | * | | | | | |
| SQRT_STD_No_MATRICULA | | | | * | * | |
| SQRT_STD_No_NUM_VEICUL | | | | | | |
| SQR_STD_IDADE_VEICULO__IDADE_VEI | * | | | | | |
| SQR_STD_IMP_REP_Diff_cov_issue | * | * | | | | |
| STD_IDADE_VEICULO__IDADE_VEICULO | | | | | | |
| STD_IMP_REP_Diff_cov_issue | * | * | | | | |
| STD_IMP_REP_IDADE_CONDUTOR_NEW | * | * | * | * | * | * |
| STD_IMP_REP_PESOBRUT | * | | | | | |
| STD_No_CP | | * | | | | |
| STD_No_NUM_VEICUL | | | | * | | * |
| TG_CVE_BONUS_MALUS | | | | | | |
| TG_CVE_COD_RC | | | | | | |
| TG_Tipo_Marca | * | | | | | |
| TG_ZONA_RC | | | | * | * | |
| Tipo_Marca | * | * | | * | * | |
| ZONA_RC | * | * | | * | * | |
| CVE_PACK | | | | | * | |
| formapa | * | | * | * | * | * |
| tipo_manual | | | | | | |
| TOTAL | 24 | 14 | 5 | 13 | 12 | 7 |

ORIGINAL - CCC

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **NO HP** | **HP** | **ALL** | **NO HP** | **HP** |
| CANAL | | | | | | * |
| CVE_BONUS_MALUS | * | * | | * | * | * |
| COND_BLOCKED_EVER | | | | | | |
| CVE_CAIXA_VELOC | | | | | | |
| CVE_CARROCARIA | * | * | | | | |
| CVE_CLASSE_DP | | | | | | |
| CVE_COD_RC | | | | | | |
| CVE_PACK | * | | | | * | |
| CVE_PROTOCOLO_nomiss | * | * | | | | |
| EST_CIVIL | * | | | * | * | |
| EXP_FPROCESO_time | | | | | | |
| FPROCESO_time | * | * | | * | * | |
| IMP_NUM_CV | * | * | | | | |
| IMP_NUM_TARA | * | | | | | |
| IMP_QUA_VALOR_NOVO | | | | | | * |
| IMP_REP_Diff_cov_issue | * | | | | | |
| IMP_REP_IDADE_CARTA_CONDUTOR_NEW | * | | | | | |
| IMP_REP_IDADE_CONDUTOR_NEW | | | | | * | |
| IMP_REP_PESOBRUT | * | | | | | |
| INV_FPROCESO_time | | * | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| INV_Phone_cond_rep | | | | | | |
| INV_Phone_rep | * | | | * | | |
| LG10_franq_P02 | | | | | | |
| LG10_IMP_NUM_CV | * | | | | | |
| NACIONALIDADE | | * | | | | |
| No_MATRICULA | | * | | | | |
| LOG_NUM_CC | | | | | | |
| No_CP | | | | | | |
| No_anulado_before_auto | | | | | | |
| OPCAO_BG | | | | * | | |
| OPCAO_PO | | | | * | * | * |
| OPCAO_VEIC | | | | | * | |
| OPT_IMP_REP_IDADE_CARTA_CONDUTOR | * | * | | | | |
| OPT_Phone_cond_rep | * | | | | | |
| SQRT_capital_CCC | | | | * | * | |
| SQRT_capital_FR | * | | | | | |
| SQRT_IMP_QUA_VALOR_NOVO | * | | * | | | |
| SQR_IMP_NUM_TARA | * | | | | | |
| TG_CVE_BONUS_MALUS | * | | | | | |
| TG_OPCAO_PO | * | | | | | |
| TG_Tipo_Marca | * | | | * | * | * |
| TG_ZONA_RC | | | | * | | * |
| TOM_BLOCKED_EVER | | * | | | | |
| Tipo_Marca | * | * | * | | * | |
| ZONA_RC | * | * | | | * | |
| capital_FN | | | | * | | |
| capital_RS | * | | | | | |
| formacob | | | | | * | |
| formapa | * | | | * | * | * |
| franq_P02 | * | * | | | | |
| TOTAL | 26 | 13 | 2 | 11 | 13 | 7 |

STANDARDIZED - CCC

| | Stratified | | | Equal | | |
|---|---|---|---|---|---|---|
| **Effect** | **ALL** | **NO HP** | **HP** | **ALL** | **NO HP** | **HP** |
| CANAL | * | * | | | | * |
| CDSUBRAMO | | | | | | |
| COND_BLOCKED_EVER | | | | | | |
| CVE_BONUS_MALUS | * | * | | * | * | * |
| CVE_CAIXA_VELOC | | | | | | |
| CVE_CLASSE_DP | | | | | | |
| CVE_CARROCARIA | * | * | | | | |
| CVE_COD_RC | | | | | | |
| CVE_PACK | * | | | | * | |
| CVE_PROTOCOLO_nomiss | * | * | | | | |
| EST_CIVIL | | * | | * | * | |
| INV_STD_FPROCESO_time | * | * | | | | |
| INV_STD_IMP_NUM_CV | * | | | | | |
| INV_STD_IMP_REP_Diff_cov_issue | * | | | | | |
| INV_STD_IMP_REP_IDADE_CONDUTOR_N | * | * | | | * | |
| INV_STD_NUM_CC | * | | | | | |
| INV_STD_Phone_cond_rep | | | | | | |
| INV_STD_Phone_rep | * | | | * | | |
| INV_STD_capital_Extras | | | | | | |
| INV_STD_franq_P02 | * | | | | | |
| LG10_STD_IMP_QUA_VALOR_NOVO | * | | * | | | |
| NACIONALIDADE | | * | | | | |
| OPCAO_BG | | | | * | | |
| OPCAO_PO | * | * | | * | * | * |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OPCAO_VEIC | | | * | | | | | | * | | | | |
| OPT_STD_IMP_REP_IDADE_CARTA_COND | * | | * | | | | | | | | | | |
| OPT_STD_Phone_cond_rep | * | | | | | | * | | | | | | |
| SQRT_STD_IMP_QUA_VALOR_NOVO | | | | | | | | | | | | | |
| SQRT_STD_capital_CCC | * | | | | | | * | | * | | | | |
| SQRT_STD_capital_FR | * | | * | | | | | | | | | | |
| SQR_STD_IMP_NUM_TARA | * | | | | | | | | | | | | |
| STD_FPROCESO_time | * | | * | | | | | | * | | | | |
| STD_IMP_NUM_TARA | | | | | | | | | | | | | |
| STD_IMP_REP_Diff_cov_issue | | | | | | | | | | | | | |
| STD_IMP_QUA_VALOR_NOVO | * | | | | | | | | | | | | |
| STD_IMP_REP_IDADE_CARTA_CONDUTOR | * | | * | | | | * | | | | | | |
| STD_IMP_REP_PESOBRUT | * | | * | | | | | | | | | | |
| STD_No_anulado_before | | | | | | | | | | | | | |
| STD_capital_FN | | | | | | | | | | | | | |
| STD_capital_RS | * | | | | | | | | | | | | |
| STD_franq_P02 | | | * | | | | | | | | | | |
| TG_CVE_BONUS_MALUS | * | | | | | | | | | | | | |
| TG_OPCAO_PO | * | | | | | | | | | | | | |
| TG_Tipo_Marca | * | | * | | | | * | | * | | * | | |
| TG_ZONA_RC | | | | | | | * | | | | * | | |
| TOM_BLOCKED_EVER | | | | | | | | | | | | | |
| Tipo_Marca | * | | * | | * | | | | * | | | | |
| ZONA_RC | * | | * | | | | | | * | | | | |
| formacob | | | | | | | | | * | | | | |
| formapa | * | | | | | | * | | * | | * | | |
| TOTAL | | 30 | | 19 | | 2 | | 11 | | 13 | | 6 | |

### C. Tested models

RCV < 90 days

| SAMPLING | Model Description | Valid: Roc Index | Train:Avg Sq Error | Train: Misclass Rate | Valid:Avg Sq Error | Valid: Misclass Rate | Valid: Gain | Train: Roc Index | Diff in ROC |
|---|---|---|---|---|---|---|---|---|---|
| SMOTE RCV 30-70 | Regression | 0.59 | 0.19 | 0.29 | 0.10 | 0.06 | 83.13 | 0.66 | 0.07 |
| SMOTE RCV 30-70 | Neural Network | 0.59 | 0.19 | 0.29 | 0.10 | 0.06 | 80.05 | 0.67 | 0.08 |
| SMOTE RCV 30-70 | HP RF max depth 25 | 0.58 | 0.15 | 0.23 | 0.09 | 0.05 | 76.20 | 0.87 | 0.29 |
| SMOTE RCV 30-70 | HP Forest | 0.58 | 0.15 | 0.22 | 0.09 | 0.05 | 74.66 | 0.87 | 0.29 |
| SMOTE RCV 30-70 | Decision Tree | 0.55 | 0.16 | 0.22 | 0.08 | 0.03 | 40.88 | 0.77 | 0.22 |
| SMOTE RCV 30-70 | AutoNeural | 0.50 | 0.21 | 0.30 | 0.10 | 0.03 | 0.11 | 0.50 | 0.00 |
| underbalanced RCV 50-50 | Regression | 0.61 | 0.24 | 0.42 | 0.24 | 0.41 | 87.20 | 0.62 | 0.01 |
| underbalanced RCV 50-50 | HP RF max depth 25 | 0.61 | 0.24 | 0.39 | 0.24 | 0.42 | 103.13 | 0.65 | 0.04 |
| underbalanced RCV 50-50 | HP Forest | 0.61 | 0.24 | 0.40 | 0.24 | 0.43 | 94.67 | 0.65 | 0.05 |
| underbalanced RCV 50-50 | Neural Network | 0.60 | 0.24 | 0.40 | 0.23 | 0.38 | 85.44 | 0.63 | 0.03 |
| underbalanced RCV 50-50 | Decision Tree | 0.59 | 0.24 | 0.42 | 0.24 | 0.52 | 83.67 | 0.62 | 0.02 |
| sel underbalanced RCV 50-50 sel var | Regression | 0.61 | 0.24 | 0.41 | 0.24 | 0.42 | 85.90 | 0.62 | 0.01 |
| sel underbalanced RCV 50-50 sel var | HP Forest | 0.61 | 0.24 | 0.41 | 0.24 | 0.43 | 82.36 | 0.63 | 0.02 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| sel underbalanced RCV 50-50 sel var | Neural Network | 0.61 | 0.24 | 0.41 | 0.23 | 0.38 | 90.05 | 0.64 | 0.03 |
| sel underbalanced RCV 50-50 sel var | Decision Tree | 0.58 | 0.23 | 0.38 | 0.25 | 0.34 | 59.17 | 0.66 | 0.07 |
| proportional RCV 2.9-97.1 | HP Forest | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 120.06 | 0.66 | 0.05 |
| proportional RCV 2.9-97.1 | Regression | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 91.26 | 0.61 | 0.00 |
| proportional RCV 2.9-97.1 | HP RF max depth 25 | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 116.98 | 0.66 | 0.05 |
| proportional RCV 2.9-97.1 | Decision Tree | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 125.22 | 0.64 | 0.03 |
| proportional RCV 2.9-97.1 | Neural Network | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 93.13 | 0.63 | 0.02 |
| proportional RCV 2.9-97.1 | Decision Tree | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 120.36 | 0.66 | 0.05 |
| sel proportional RCV 2.9-97.1 | Regression | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 93.13 | 0.62 | 0.01 |
| sel proportional RCV 2.9-97.1 | Neural Network | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 110.06 | 0.62 | 0.01 |
| sel proportional RCV 2.9-97.1 | HP Forest | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 90.05 | 0.63 | 0.02 |
| sel proportional RCV 2.9-97.1 | Decision Tree | 0.61 | 0.03 | 0.03 | 0.03 | 0.03 | 116.24 | 0.63 | 0.02 |
| underbalanced SMOTE (30-70) RCV 50-50 | Regression | 0.59 | 0.23 | 0.38 | 0.23 | 0.41 | 86.21 | 0.66 | 0.07 |
| underbalanced SMOTE (30-70) RCV 50-50 | Neural Network | 0.59 | 0.23 | 0.38 | 0.22 | 0.38 | 79.28 | 0.67 | 0.09 |
| underbalanced SMOTE (30-70) RCV 50-50 | AutoNeural | 0.58 | 0.23 | 0.37 | 0.23 | 0.38 | 70.82 | 0.67 | 0.09 |
| underbalanced SMOTE (30-70) RCV 50-50 | HP RF max depth 25 | 0.58 | 0.18 | 0.24 | 0.20 | 0.29 | 75.43 | 0.84 | 0.26 |
| underbalanced SMOTE (30-70) RCV 50-50 | HP Forest | 0.58 | 0.18 | 0.25 | 0.20 | 0.29 | 70.05 | 0.83 | 0.26 |
| underbalanced SMOTE (30-70) RCV 50-50 | Decision Tree | 0.55 | 0.19 | 0.32 | 0.18 | 0.26 | 51.78 | 0.77 | 0.22 |
| level RCV 30-70 | HP Forest | 0.61 | 0.20 | 0.30 | 0.10 | 0.03 | 109.29 | 0.66 | 0.04 |
| level RCV 30-70 | HP RF max depth 25 | 0.61 | 0.20 | 0.30 | 0.10 | 0.03 | 113.14 | 0.66 | 0.05 |
| level RCV 30-70 | Regression | 0.61 | 0.20 | 0.30 | 0.10 | 0.05 | 105.44 | 0.62 | 0.02 |
| level RCV 30-70 | Neural Network | 0.61 | 0.20 | 0.30 | 0.10 | 0.05 | 94.67 | 0.63 | 0.02 |
| level RCV 30-70 | Decision Tree | 0.57 | 0.20 | 0.29 | 0.10 | 0.05 | 90.52 | 0.58 | 0.01 |
| sel level RCV 30-70 | Regression | 0.61 | 0.20 | 0.30 | 0.10 | 0.04 | 88.19 | 0.62 | 0.01 |
| sel level RCV 30-70 | Neural Network | 0.61 | 0.20 | 0.29 | 0.10 | 0.05 | 86.98 | 0.63 | 0.02 |
| sel level RCV 30-70 | HP Forest | 0.61 | 0.20 | 0.30 | 0.10 | 0.03 | 78.51 | 0.63 | 0.02 |
| sel level RCV 30-70 | Decision Tree | 0.57 | 0.20 | 0.29 | 0.10 | 0.05 | 91.32 | 0.59 | 0.02 |

RCV > 90 days

| SAMPLING | Model Description | Valid: Roc Index | Train: Avg Sq Error | Train: Miscla ss Rate | Valid: Avg Sq Error | Valid: Miscla ss Rate | Valid: Gain | Train: Roc Index | Diff in ROC |
|---|---|---|---|---|---|---|---|---|---|
| underbalanced RCV 50-50 | HP RF max depth 25 | 0.65 | 0.23 | 0.38 | 0.24 | 0.47 | 91.29 | 0.67 | 0.02 |
| underbalanced RCV 50-50 | HP Forest 80 trees | 0.64 | 0.23 | 0.38 | 0.24 | 0.47 | 89.77 | 0.67 | 0.03 |
| underbalanced RCV 50-50 | Regression | 0.64 | 0.23 | 0.39 | 0.24 | 0.45 | 94.69 | 0.65 | 0.00 |
| underbalanced RCV 50-50 | Neural Network | 0.64 | 0.23 | 0.40 | 0.23 | 0.51 | 111.70 | 0.65 | 0.01 |
| underbalanced RCV 50-50 | Decision Tree | 0.62 | 0.23 | 0.39 | 0.24 | 0.44 | 58.86 | 0.65 | 0.03 |
| proportional RCV 2.9-97.1 | HP Forest 80 trees | 0.65 | 0.06 | 0.06 | 0.06 | 0.06 | 99.22 | 0.68 | 0.03 |
| proportional RCV 2.9-97.1 | HP RF max depth 25 | 0.65 | 0.06 | 0.06 | 0.06 | 0.06 | 100.74 | 0.68 | 0.03 |
| proportional RCV 2.9-97.1 | Neural Network | 0.65 | 0.06 | 0.06 | 0.06 | 0.06 | 100.02 | 0.66 | 0.01 |
| proportional RCV 2.9-97.1 | Regression | 0.65 | 0.06 | 0.06 | 0.06 | 0.06 | 113.59 | 0.65 | 0.01 |
| proportional RCV 2.9-97.1 | Decision Tree | 0.64 | 0.06 | 0.06 | 0.06 | 0.06 | 101.36 | 0.66 | 0.02 |
| level RCV 30-70 | Regression | 0.65 | 0.20 | 0.30 | 0.11 | 0.08 | 108.67 | 0.65 | 0.01 |
| level RCV 30-70 | HP RF max depth 25 | 0.65 | 0.20 | 0.30 | 0.11 | 0.06 | 92.04 | 0.67 | 0.02 |
| level RCV 30-70 | HP Forest 80 trees | 0.65 | 0.20 | 0.30 | 0.11 | 0.06 | 95.82 | 0.67 | 0.02 |
| level RCV 30-70 | Neural Network | 0.65 | 0.20 | 0.30 | 0.11 | 0.07 | 96.58 | 0.66 | 0.01 |
| level RCV 30-70 | Decision Tree | 0.61 | 0.20 | 0.30 | 0.11 | 0.07 | 76.43 | 0.63 | 0.01 |

QIV < 90 days

| SAMPLING | Model Description | Valid: Roc Index | Train :Avg Sq Error | Train : Miscl ass Rate | Valid: Avg Sq Error | Valid: Miscl ass Rate | Valid: Gain | Train : Roc Index | Diff in ROC | Test: Roc Index | Test: Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SMOTE QIV 30-70 | Neural Network | 0.58 | 0.20 | 0.30 | 0.10 | 0.02 | 62.35 | 0.66 | 0.07 | - | - |
| SMOTE QIV 30-70 | Regression | 0.58 | 0.20 | 0.30 | 0.10 | 0.05 | 55.72 | 0.67 | 0.09 | - | - |
| SMOTE QIV 30-70 | HP Forest | 0.57 | 0.12 | 0.16 | 0.08 | 0.05 | 59.03 | 0.94 | 0.37 | - | - |
| SMOTE QIV 30-70 | HP RF max depth 25 | 0.57 | 0.12 | 0.16 | 0.08 | 0.05 | 47.44 | 0.94 | 0.37 | - | - |
| SMOTE QIV 30-70 | Decision Tree | 0.55 | 0.19 | 0.28 | 0.09 | 0.02 | 45.39 | 0.69 | 0.14 | - | - |
| underbalanced QIV 50-50 | HP Forest | 0.62 | 0.24 | 0.39 | 0.24 | 0.45 | 125.30 | 0.66 | 0.04 | 0.61 | 81.15 |
| underbalanced QIV 50-50 | HP RF max depth 25 | 0.62 | 0.24 | 0.39 | 0.24 | 0.46 | 126.95 | 0.66 | 0.04 | 0.61 | 48.55 |
| underbalanced QIV 50-50 | Regression | 0.61 | 0.24 | 0.43 | 0.25 | 0.48 | 73.52 | 0.60 | -0.01 | 0.61 | 73.91 |
| underbalanced QIV 50-50 | Decision Tree | 0.59 | 0.23 | 0.40 | 0.24 | 0.46 | 56.17 | 0.64 | 0.05 | 0.58 | 18.23 |
| underbalanced QIV 50-50 | Neural Network | 0.56 | 0.23 | 0.41 | 0.24 | 0.62 | 55.72 | 0.64 | 0.08 | 0.56 | 23.19 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| sel underbalanced QIV 50-50 | HP Forest | 0.61 | 0.24 | 0.42 | 0.24 | 0.48 | 50.75 | 0.62 | 0.01 | 0.61 | 41.30 |
| sel underbalanced QIV 50-50 | Regression | 0.61 | 0.24 | 0.43 | 0.24 | 0.45 | 82.84 | 0.60 | -0.01 | 0.61 | 55.79 |
| sel underbalanced QIV 50-50 | Neural Network | 0.59 | 0.23 | 0.40 | 0.24 | 0.51 | 29.21 | 0.63 | 0.05 | 0.59 | 12.32 |
| sel underbalanced QIV 50-50 | Decision Tree | 0.57 | 0.22 | 0.38 | 0.25 | 0.51 | 39.77 | 0.67 | 0.10 | 0.57 | 25.58 |
| proportional RCV 2.9-97.1 | HP Forest | 0.62 | 0.02 | 0.02 | 0.02 | 0.02 | 103.76 | 0.69 | 0.07 | 0.61 | 23.19 |
| proportional RCV 2.9-97.1 | HP RF max depth 25 | 0.62 | 0.02 | 0.02 | 0.02 | 0.02 | 108.73 | 0.69 | 0.07 | 0.61 | 34.05 |
| proportional RCV 2.9-97.1 | Neural Network | 0.62 | 0.02 | 0.02 | 0.02 | 0.02 | 85.54 | 0.61 | 0.00 | 0.61 | 73.91 |
| proportional RCV 2.9-97.1 | Regression | 0.61 | 0.02 | 0.02 | 0.02 | 0.02 | 76.47 | 0.61 | -0.01 | 0.62 | 73.24 |
| proportional RCV 2.9-97.1 | Decision Tree | 0.60 | 0.02 | 0.02 | 0.02 | 0.02 | 58.27 | 0.64 | 0.03 | 0.59 | 49.22 |
| sel proportional RCV 2.9-97.1 | Reg6 | 0.61 | 0.24 | 0.43 | 0.24 | 0.45 | 75.60 | 0.61 | 0.00 | 0.61 | 77.53 |
| sel proportional RCV 2.9-97.1 | HPDMForest24 | 0.61 | 0.23 | 0.50 | 0.24 | 0.02 | 72.29 | 0.63 | 0.02 | 0.61 | 52.17 |
| sel proportional RCV 2.9-97.1 | HPDMForest22 | 0.61 | 0.23 | 0.40 | 0.24 | 0.51 | 81.67 | 0.63 | 0.02 | 0.61 | 59.42 |
| sel proportional RCV 2.9-97.1 | Neural6 | 0.61 | 0.22 | 0.38 | 0.25 | 0.51 | 78.91 | 0.61 | 0.00 | 0.62 | 102.89 |
| sel proportional RCV 2.9-97.1 | Tree11 | 0.60 | 0.21 | 0.38 | 0.26 | 0.39 | 66.18 | 0.64 | 0.04 | 0.57 | 46.94 |
| underbalanced SMOTE (30-70) RCV 50-50 | Neural Network | 0.58 | 0.20 | 0.30 | 0.10 | 0.02 | 62.35 | 0.66 | 0.07 | - | - |
| underbalanced SMOTE (30-70) RCV 50-50 | Regression | 0.58 | 0.20 | 0.30 | 0.10 | 0.05 | 55.72 | 0.67 | 0.09 | - | - |
| underbalanced SMOTE (30-70) RCV 50-50 | HP Forest | 0.57 | 0.12 | 0.15 | 0.08 | 0.05 | 55.72 | 0.94 | 0.37 | - | - |
| underbalanced SMOTE (30-70) RCV 50-50 | HP RF max depth 25 | 0.57 | 0.12 | 0.16 | 0.08 | 0.05 | 52.41 | 0.94 | 0.37 | - | - |
| underbalanced SMOTE (30-70) RCV 50-50 | Decision Tree | 0.55 | 0.19 | 0.28 | 0.09 | 0.02 | 45.39 | 0.69 | 0.14 | - | - |
| level RCV 30-70 | HP RF max depth 25 | 0.62 | 0.20 | 0.30 | 0.09 | 0.02 | 100.45 | 0.66 | 0.04 | 0.61 | 52.17 |
| level RCV 30-70 | HP Forest | 0.62 | 0.20 | 0.30 | 0.09 | 0.02 | 87.19 | 0.66 | 0.04 | 0.61 | 26.81 |
| level RCV 30-70 | Neural Network | 0.61 | 0.20 | 0.30 | 0.10 | 0.02 | 92.16 | 0.64 | 0.02 | 0.60 | 70.29 |
| level RCV 30-70 | Regression | 0.61 | 0.20 | 0.30 | 0.10 | 0.02 | 97.80 | 0.61 | 0.00 | 0.62 | 91.12 |
| level RCV 30-70 | Decision Tree | 0.59 | 0.20 | 0.30 | 0.10 | 0.02 | 55.85 | 0.63 | 0.04 | 0.58 | 26.92 |
| sel level RCV 30-70 | HP Forest | 0.61 | 0.20 | 0.30 | 0.10 | 0.02 | 78.91 | 0.62 | 0.01 | 0.61 | 59.42 |
| sel level RCV 30-70 | Regression | 0.61 | 0.20 | 0.30 | 0.10 | 0.02 | 73.94 | 0.61 | 0.00 | 0.61 | 48.55 |
| sel level RCV 30-70 | Neural Network | 0.61 | 0.20 | 0.30 | 0.10 | 0.02 | 70.63 | 0.61 | 0.00 | 0.61 | 84.78 |
| sel level RCV 30-70 | Decision Tree | 0.59 | 0.20 | 0.30 | 0.10 | 0.02 | 50.47 | 0.59 | 0.01 | 0.57 | 17.88 |

QIV > 90 days

| SAMPLING | Model Description | Valid: Roc Index | Train: Avg Sq Error | Train: Miscla ss Rate | Valid: Avg Sq Error | Valid: Miscla ss Rate | Valid: Gain | Train: Roc Index | Diff in ROC |
|---|---|---|---|---|---|---|---|---|---|
| underbalanced QIV 50-50 | HP Forest 80 trees | 0.64 | 0.23 | 0.38 | 0.24 | 0.44 | 101.28 | 0.68 | 0.03 |
| underbalanced QIV 50-50 | HP RF max depth 25 | 0.64 | 0.23 | 0.38 | 0.24 | 0.44 | 112.81 | 0.68 | 0.04 |
| underbalanced QIV 50-50 | Regression | 0.63 | 0.23 | 0.40 | 0.23 | 0.41 | 116.36 | 0.65 | 0.02 |
| underbalanced QIV 50-50 | Neural Network | 0.62 | 0.23 | 0.39 | 0.23 | 0.38 | 72.02 | 0.67 | 0.04 |
| underbalanced QIV 50-50 | Decision Tree | 0.62 | 0.23 | 0.39 | 0.24 | 0.35 | 55.79 | 0.65 | 0.04 |
| proportional QIV 2.9-97.1 | HP RF max depth 25 | 0.64 | 0.03 | 0.03 | 0.03 | 0.03 | 100.40 | 0.68 | 0.04 |
| proportional QIV 2.9-97.1 | HP Forest 80 trees | 0.64 | 0.03 | 0.03 | 0.03 | 0.03 | 95.08 | 0.68 | 0.04 |
| proportional QIV 2.9-97.1 | Decision Tree | 0.64 | 0.03 | 0.03 | 0.03 | 0.03 | 94.60 | 0.67 | 0.04 |
| proportional QIV 2.9-97.1 | Regression | 0.63 | 0.03 | 0.03 | 0.03 | 0.03 | 76.20 | 0.65 | 0.02 |
| proportional QIV 2.9-97.1 | Neural Network | 0.63 | 0.03 | 0.03 | 0.03 | 0.03 | 94.63 | 0.66 | 0.03 |
| level QIV 30-70 | HP RF max depth 25 | 0.64 | 0.20 | 0.30 | 0.10 | 0.03 | 118.13 | 0.68 | 0.03 |
| level QIV 30-70 | HP Forest 80 trees | 0.64 | 0.20 | 0.30 | 0.10 | 0.03 | 105.72 | 0.68 | 0.03 |
| level QIV 30-70 | Regression | 0.63 | 0.20 | 0.30 | 0.10 | 0.05 | 111.92 | 0.65 | 0.02 |
| level QIV 30-70 | Neural Network | 0.63 | 0.20 | 0.30 | 0.10 | 0.05 | 86.21 | 0.66 | 0.03 |
| level QIV 30-70 | Decision Tree | 0.62 | 0.20 | 0.29 | 0.10 | 0.04 | 62.41 | 0.65 | 0.03 |

CCC < 90 days

| SAMPLING | Model Description | Valid: Roc Index | Train: Avg Sq Error | Train: Miscla ss Rate | Valid: Avg Sq Error | Valid: Miscla ss Rate | Valid: Gain | Train: Roc Index | Diff in ROC | Test: Roc Index | Test: Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SMOTE CCC 30-70 | Regression | 0.57 | 0.19 | 0.27 | 0.09 | 0.07 | 43.58 | 0.69 | 0.13 | - | - |
| SMOTE CCC 30-70 | HP Forest | 0.55 | 0.04 | 0.02 | 0.04 | 0.03 | 32.53 | 1.00 | 0.44 | - | - |
| SMOTE CCC 30-70 | HP RF max depth 25 | 0.55 | 0.04 | 0.02 | 0.04 | 0.03 | 21.49 | 1.00 | 0.45 | - | - |
| SMOTE CCC 30-70 | Decision Tree | 0.55 | 0.16 | 0.23 | 0.09 | 0.11 | 52.41 | 0.80 | 0.25 | - | - |
| SMOTE CCC 30-70 | Neural Network | 0.54 | 0.19 | 0.28 | 0.08 | 0.04 | 54.62 | 0.69 | 0.15 | - | - |
| underbal. CCC 50-50 | HP Forest | 0.61 | 0.24 | 0.34 | 0.24 | 0.54 | 54.62 | 0.72 | 0.10 | 0.64 | 69.01 |
| underbal. CCC 50-50 | HP RF max depth 25 | 0.61 | 0.24 | 0.34 | 0.24 | 0.54 | 54.62 | 0.72 | 0.10 | 0.64 | 69.01 |
| underbal. CCC 50-50 | HP RF max depth 25 | 0.61 | 0.24 | 0.34 | 0.24 | 0.54 | 54.62 | 0.72 | 0.10 | 0.64 | 69.01 |
| underbal. CCC 50-50 | Neural Network | 0.60 | 0.20 | 0.34 | 0.25 | 0.41 | 87.75 | 0.75 | 0.15 | 0.56 | 26.76 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| underbal. CCC 50-50 | Decision Tree | 0.60 | 0.20 | 0.35 | 0.25 | 0.34 | 66.86 | 0.72 | 0.12 | 0.56 | 42.79 |
| underbal. CCC 50-50 | Regression | 0.54 | 0.24 | 0.46 | 0.25 | 0.90 | 8.90 | 0.54 | 0.00 | 0.51 | 1.39 |
| sel CCC 50-50 | Decision Tree | 0.59 | 0.20 | 0.35 | 0.27 | 0.27 | 97.87 | 0.73 | 0.13 | 0.49 | 27.95 |
| sel CCC 50-50 | HP RF max depth 25 | 0.55 | 0.24 | 0.43 | 0.25 | 0.34 | 10.44 | 0.62 | 0.07 | 0.58 | 26.76 |
| sel CCC 50-50 | HP Forest | 0.55 | 0.24 | 0.43 | 0.25 | 0.34 | 10.44 | 0.62 | 0.07 | 0.58 | 26.76 |
| sel CCC 50-50 | Regression | 0.54 | 0.25 | 0.45 | 0.25 | 0.25 | 29.66 | 0.56 | 0.03 | 0.55 | 58.35 |
| sel CCC 50-50 | Neural Network | 0.53 | 0.19 | 0.32 | 0.29 | 0.58 | 32.53 | 0.77 | 0.25 | 0.54 | 57.75 |
| proportional CCC 2.9-97.1 | HP RF max depth 25 | 0.64 | 0.01 | 0.01 | 0.01 | 0.01 | 87.75 | 0.73 | 0.09 | 0.64 | 1.41 |
| proportional CCC 2.9-97.1 | Decision Tree | 0.63 | 0.01 | 0.01 | 0.01 | 0.01 | 90.58 | 0.70 | 0.07 | 0.54 | 3.76 |
| proportional CCC 2.9-97.1 | HP RF max depth 25 | 0.62 | 0.01 | 0.01 | 0.01 | 0.01 | 65.67 | 0.73 | 0.10 | 0.64 | 15.49 |
| proportional CCC 2.9-97.1 | HP Forest | 0.62 | 0.01 | 0.01 | 0.01 | 0.01 | 76.71 | 0.73 | 0.11 | 0.64 | 1.41 |
| proportional CCC 2.9-97.1 | Neural Network | 0.60 | 0.01 | 0.01 | 0.01 | 0.01 | 109.84 | 0.67 | 0.07 | 0.59 | 83.10 |
| proportional CCC 2.9-97.1 | Regression | 0.54 | 0.01 | 0.01 | 0.01 | 0.01 | 8.90 | 0.54 | 0.00 | 0.51 | 1.39 |
| sel proportional CCC 2.9-97.1 | Decision Tree | 0.58 | 0.01 | 0.01 | 0.01 | 0.01 | 76.71 | 0.86 | 0.28 | 0.55 | 16.52 |
| sel proportional CCC 2.9-97.1 | HP RF max depth 25 | 0.57 | 0.01 | 0.01 | 0.01 | 0.01 | 43.58 | 0.66 | 0.08 | 0.61 | 97.18 |
| sel proportional CCC 2.9-97.1 | HP Forest | 0.57 | 0.01 | 0.01 | 0.01 | 0.01 | 43.58 | 0.66 | 0.08 | 0.61 | 97.18 |
| sel proportional CCC 2.9-97.1 | Neural Network | 0.56 | 0.01 | 0.01 | 0.01 | 0.01 | 65.67 | 0.65 | 0.09 | 0.59 | 97.18 |
| sel proportional CCC 2.9-97.1 | Regression | 0.50 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.50 | 0.00 | 0.50 | 0.00 |
| underbal. SMOTE (30-70) CCC 50-50 | Regression | 0.57 | 0.19 | 0.27 | 0.09 | 0.07 | 43.58 | 0.69 | 0.13 | | |
| underbal. SMOTE (30-70) CCC 50-50 | HP RF max depth 25 | 0.56 | 0.04 | 0.02 | 0.04 | 0.03 | 65.67 | 1.00 | 0.44 | | |
| underbal. SMOTE (30-70) CCC 50-50 | HP Forest | 0.55 | 0.04 | 0.02 | 0.04 | 0.03 | 32.53 | 1.00 | 0.45 | | |
| underbal. SMOTE (30-70) CCC 50-50 | Decision Tree | 0.55 | 0.16 | 0.23 | 0.09 | 0.11 | 52.41 | 0.80 | 0.25 | | |
| underbal. SMOTE (30-70) CCC 50-50 | Neural Network | 0.54 | 0.19 | 0.28 | 0.08 | 0.04 | 54.62 | 0.69 | 0.15 | | |
| level CCC 30-70 | HP RF max depth 25 | 0.60 | 0.20 | 0.30 | 0.09 | 0.01 | 109.84 | 0.70 | 0.10 | 0.63 | 26.76 |
| level CCC 30-70 | HP RF max depth 25 | 0.60 | 0.20 | 0.30 | 0.09 | 0.01 | 109.84 | 0.70 | 0.10 | 0.63 | 26.76 |
| level CCC 30-70 | HP Forest | 0.60 | 0.20 | 0.30 | 0.09 | 0.01 | 109.84 | 0.69 | 0.09 | 0.64 | 1.41 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| level CCC 30-70 | Decision Tree | 0.59 | 0.16 | 0.24 | 0.13 | 0.10 | 33.27 | 0.78 | 0.19 | 0.52 | 55.90 |
| level CCC 30-70 | Neural Network | 0.58 | 0.18 | 0.27 | 0.12 | 0.07 | 32.53 | 0.71 | 0.14 | 0.55 | 26.76 |
| level CCC 30-70 | Regression | 0.54 | 0.20 | 0.30 | 0.10 | 0.01 | 8.90 | 0.55 | 0.00 | 0.51 | 1.39 |
| sel level CCC 30-70 | HP RF max depth 25 | 0.56 | 0.21 | 0.30 | 0.09 | 0.01 | 41.74 | 0.62 | 0.06 | 0.60 | 12.68 |
| sel level CCC 30-70 | HP Forest | 0.56 | 0.21 | 0.30 | 0.09 | 0.01 | 41.74 | 0.62 | 0.06 | 0.60 | 12.68 |
| sel level CCC 30-70 | Decision Tree | 0.56 | 0.16 | 0.25 | 0.12 | 0.12 | 21.39 | 0.78 | 0.22 | 0.53 | 39.87 |
| sel level CCC 30-70 | Neural Network | 0.55 | 0.19 | 0.28 | 0.10 | 0.05 | 10.44 | 0.68 | 0.13 | 0.59 | 83.10 |
| sel level CCC 30-70 | Regression | 0.53 | 0.21 | 0.30 | 0.09 | 0.01 | 7.61 | 0.56 | 0.03 | 0.55 | 24.54 |

CCC > 90days

| SAMPLING | Model Description | Valid: Roc Index | Train: Avg Sq Error | Train: Miscla ss Rate | Valid: Avg Sq Error | Valid: Miscla ss Rate | Valid: Gain | Train: Roc Index | Diff in ROC |
|---|---|---|---|---|---|---|---|---|---|
| underbalanced CCC 50-50 | HP Forest | 0.62 | 0.23 | 0.39 | 0.24 | 0.63 | 115.88 | 0.70 | 0.08 |
| underbalanced CCC 50-50 | HP RF max depth 25 underbalanced | 0.62 | 0.23 | 0.39 | 0.24 | 0.63 | 115.88 | 0.70 | 0.08 |
| underbalanced CCC 50-50 | Regression | 0.61 | 0.23 | 0.41 | 0.25 | 0.53 | 94.29 | 0.64 | 0.03 |
| underbalanced CCC 50-50 | Decision Tree | 0.61 | 0.24 | 0.41 | 0.25 | 0.65 | 91.02 | 0.63 | 0.02 |
| underbalanced CCC 50-50 | Neural Network | 0.60 | 0.22 | 0.37 | 0.23 | 0.40 | 46.80 | 0.68 | 0.08 |
| proportional CCC 2.9-97.1 | HP RF max depth 25 underbalanced | 0.62 | 0.04 | 0.04 | 0.04 | 0.04 | 115.88 | 0.70 | 0.08 |
| proportional CCC 2.9-97.1 | HP Forest | 0.62 | 0.04 | 0.04 | 0.04 | 0.04 | 102.93 | 0.70 | 0.08 |
| proportional CCC 2.9-97.1 | Decision Tree | 0.61 | 0.04 | 0.04 | 0.04 | 0.04 | 116.53 | 0.65 | 0.03 |
| proportional CCC 2.9-97.1 | Regression | 0.60 | 0.04 | 0.04 | 0.04 | 0.04 | 111.56 | 0.61 | 0.01 |
| proportional CCC 2.9-97.1 | Neural Network | 0.60 | 0.04 | 0.04 | 0.04 | 0.04 | 42.48 | 0.62 | 0.03 |
| level CCC 30-70 | HP Forest | 0.62 | 0.20 | 0.30 | 0.10 | 0.04 | 94.29 | 0.69 | 0.07 |
| level CCC 30-70 | HP RF max depth 25 underbalanced | 0.62 | 0.20 | 0.30 | 0.10 | 0.04 | 94.29 | 0.69 | 0.07 |
| level CCC 30-70 | Regression | 0.60 | 0.20 | 0.30 | 0.11 | 0.06 | 100.82 | 0.62 | 0.02 |
| level CCC 30-70 | Neural Network | 0.59 | 0.20 | 0.29 | 0.11 | 0.06 | 42.48 | 0.66 | 0.07 |
| level CCC 30-70 | Decision Tree | 0.59 | 0.20 | 0.29 | 0.11 | 0.07 | 97.12 | 0.62 | 0.03 |

**D.** SAS CODE for finding the missing patterns

```
/*compute missings for all the variables*/
data A;
set try(drop=anulada CVE_MODELO CVE_PROTOCOLO NUM_SINISTROS_ULT_5_ANOS TIPO
flag_CCC flag_FRB flag_fusion_cond flag_fusion_tom flag_QIV flag_RCV
Sinistro_CCC Sinistro_CCCold Sinistro_FRB Sinistro_FRBold Sinistro_ODVold
Sinistro_QIV Sinistro_QIVold Sinistro_RCV Sinistro_RCV_IDS_Credorold
Sinistro_RCV_IDS_Devedorold Sinistro_RCV_sem_IDSold Sinistro_RCVold
Sinistro_CCC90 Sinistro_FRB90 Sinistro_QIV90 Sinistro_RCV90
NUM_ANOS_COM_SEGURO NUM_SINISTROS_ULT_5_ANOS NUM_SINISTROS_ULT_3_ANOS);
array varsN(*) _NUMERIC_;
nMissingNumeric = cmiss(of varsN[*]);
array varsCh(*) _CHARACTER_;
nMissingCharacter = cmiss(of varsCh[*]);
nMissingTotal=nMissingNumeric+nMissingCharacter;
nvars = dim(varsN) + dim(varsCh);
percent_missing = round((nMissingTotal/nvars)*100, .01);
keep NPOLIZA nMissingNumeric nMissingCharacter nMissingTotal
percent_missing/*nAll*/;
run;

/*Patterns in missing values*/
ods select MissPattern;
proc mi data=try1 nimpute=0;
var _ALL_;
CLASS _CHARACTER_;
FCS;
run;
```

**E.** SAS CODE for SMOTE

```
/*****************************************************************************
****************************/
/*/*/*/*/* SMOTE - data loading and parameters */*/*/*/*/*/
/*****************************************************************************
****************************/

*Macrovariable: number of k nearest neighbors to SMOTE +1;
%Let NN = 6;
*Macrovariable: the value of the SMOTE multiplier - the number of additional
minority class instances desired for each of the original minority class examples;
%Let Mult = 13; *it was 10;
*Macrovariable: dataset name to be SMOTED;
%Let ToSMOTE = GTNVAUTO.RCV_FORSMOTE_TRAIN;
*Macrovariable: target name;
%Let Target = Sinistro_RCV90;

options compress=yes;

*Loads initial dataset and creates ID used later;
DATA prep;
set &ToSMOTE
(drop=_WARN_);
ID = cats(_N_);
RUN;

ods _all_ close;

*Dataset that contains only the observations in input where target=1;
DATA claim;
set prep;
where &Target.=1;
drop NPOLIZA &Target.;
RUN;

*Dataset without ids;
data forvars;
set claim;
drop _dataobs_ id;
run;

/*****************************************************************************
****************************/
/*Macros for name lists*/
/*****************************************************************************
****************************/

/*creates a macro variable called &vlist that will contain the names of all the
variables in your dataset, separated by a space*/

*Macrovariables of all inputs, Nominal and Interval;
%MACRO getvars(dsn, type);
%IF &type=char %THEN %DO;
      %global vlistN;
       proc sql;
       select name into :vlistN separated by ' '
       from dictionary.columns
       where libname=upcase("work") and memname=upcase("&dsn") and type='char';
       quit;
%END;

%ELSE %IF &type=num %THEN %DO;
            %global vlistI;
             proc sql;
             select name into :vlistI separated by ' '
             from dictionary.columns
```

```
              where libname=upcase("work") and memname=upcase("&dsn") and
              type='num';
              quit;
%END;
%ELSE %DO;
      %global vlistAll;
              proc sql;
              select name into :vlistAll separated by ' '
              from dictionary.columns
              where libname=upcase("work") and memname=upcase("&dsn");
              quit;
%END;
%MEND getvars;
%getvars(forvars, char);
%getvars(forvars, num);
%getvars(forvars, all);

/*Macro for variable list separated by comma*/
%macro qcomma(list, delim=%str( ));
%local varlist i var;
%let varlist=&list.;
%let i=0;
%do %while(%qscan(&varlist.,%eval(&i.+1),%quote(&delim.)) ne );
      %if &i. gt 0 %then %do; %let var=%str(%trim(%quote(&var.))%str(, )); %end;
            %let i = %eval(&i.+1);
            %let
var=&var.%str(%')%qcmpres(%qscan(&varlist,&i.,%quote(&delim.)))%str(%');
%end;
%unquote(&var.) %* MUST be first and only thing on a line by itself *;
%mend;

/*Macrovariables for variables lists separated by comma*/
%let listAllcomma=%qcomma(&vlistAll., delim=%str( ));
%let listIcomma=%qcomma(&vlistI., delim=%str( ));
%let listNcomma=%qcomma(&vlistN., delim=%str( ));
/*************************************************************************/

/*****************************************************************************
*****************************/
/*/*/*/*/* SMOTE - method */*/*/*/*/*/
/*****************************************************************************
*****************************/

*Compute Gower distance (Euklidian doesn't accept nominal variables);
PROC DISTANCE Data=claim METHOD=dgower SHAPE=SQUARE out=distances;
VAR interval(&vlistI) nominal(&vlistN);
id ID;
RUN;

*Creates a dataset - nTest - with the dk-1 nearest neighbors(target=1) of each
observation where target=1.
The ID of and distance to the dk-1 nearest neighbors are given by the "Nbor" and
"Distance" variables;
PROC MODECLUS data=distances dk=&NN
neighbor out=modeclus_out;
id ID;
ods output Neighbor = nTest;
RUN;

ods listing;

*Creates a dataset of each pair of target=1 observations and dk-1 nearest neighbors
(also target=1);
*Essentially modifying nTest so that all blank IDs are filled with the appropriate
ID (now called center_id);
*Also, an index variable is added, and every variable except "center_id", "nbor",
and "index" is dropped;
data neighbor_id (keep=center_id nbor index);
```

```
set ntest;
retain center_id '000000000000';
if not missing(id) then center_ID = put(id,$12.);
else center_id = center_id;
index = _N_; /*prepare id for the new cases*/
run;

proc sort data= neighbor_id; by center_id; run;

*Transposes the neighbor_id;
*Essentially creates a list of center_ids with each of the 5 nearest neighbors
listed as a variable (in a column);
*The other superfluous columns/variables are dropped in the following data step;
proc transpose data = neighbor_id out= neighbor_trans;
by center_id;
var nbor;
run;

*Removes unwanted columns/variables from neighbor_trans;
data neighbor_trans (drop= _name_ _label_ col&NN);
set neighbor_trans;
run;

*Creates neighbor_id2 containing a list of observation/neighbor pairs to be used
for SMOTEing. For each observation
in claim, &Mult pairs of that observation with one of its &NN-1 nearest neighbors
will be generated. The neighbor
used for each pair for each observation is chosen randomly from the &NN-1 nearest
neighbors;
data neighbor_id2 (keep = center_id nbor);
set neighbor_trans;
do i = 1 to &Mult;
 random = rand("uniform");
 if random < .2 then nbor = col1;
 else if random < .4 then nbor = col2;
 else if random < .6 then nbor = col3;
 else if random < .8 then nbor = col4;
 else nbor = col5;  *Configuration for 5 nearest neighbors;
output;
end;
run;

*Adds an index variable to neighbor_id2;
data neighbor_id2;
set neighbor_id2;
index = _n_;
/*if index >3000 then delete; *here is temporary CONDITION only to test!!!!!!!!!;*/
run;

/*****************************************************************************
***************************/
/* NEW METHOD */
/*****************************************************************************
***************************/

/*Macro for intervals*/
%macro loopI (name_list);
       %local i next_name;
       %do i=1 %to %sysfunc(countw(&name_list));
          %let next_name = %scan(&name_list, &i);

              data &next_name._prep;
              set claim;
              keep id &next_name;
              run;

              PROC SQL;
              CREATE TABLE &next_name._aux AS
```

```
            SELECT a.*,b.&next_name. as &next_name._center
            FROM neighbor_id2 as a
            left join &next_name._prep as b
            on a.center_id=b.id;
            quit;

            PROC SQL;
            CREATE TABLE &next_name._new AS
            SELECT a.*,b.&next_name. as &next_name._nbor
            FROM &next_name._aux as a
            left join &next_name._prep as b
            on a.nbor=b.id;
            quit;

            data &next_name.;
            set &next_name._new;
            rand=rand('uniform');
            &next_name.=(&next_name._nbor-
            &next_name._center)*rand+&next_name._center;
            keep index &next_name.;
            run;

            proc sort data=&next_name.; by index; run;
            proc delete data= &next_name._prep &next_name._aux &next_name._new;
           /*comment out to see all the tables*/
            run;
      %end;

      data Interval;
      merge &name_list;
      by index;
      run;
%mend loopI;

%loopI(&vlistI);

/*Macro for nominals*/
%macro loopN (name_list);
      %local i next_name;
      %do i=1 %to %sysfunc(countw(&name_list));
         %let next_name = %scan(&name_list, &i);

            data &next_name._prep;
            set claim;
            keep id &next_name;
            run;

            PROC SQL;
            CREATE TABLE &next_name._new AS
            SELECT a.*,b.&next_name. as &next_name._nbor
            FROM neighbor_id2 as a
            left join &next_name._prep as b
            on a.nbor=b.id;
            quit;

            data &next_name.;
            set &next_name._new;
            &next_name.=&next_name._nbor;
            keep index &next_name.;
            run;

            proc sort data=&next_name.; by index; run;
            proc delete data= &next_name._prep &next_name._new; /*comment out to
see all the tables*/
      %end;

      data Character;
      merge &name_list;
```

```
      by index;
      run;
%mend loopN;

%loopN(&vlistN);


/*Here binary variables need to be modified by hand*/
data new_cases2;
merge Character Interval;
by index;
      if PHONE_binary <0.5 then PHONE_binary=0;
      if PHONE_binary >=0.5 then PHONE_binary=1;
      if flag_2a_viatura <0.5 then flag_2a_viatura=0;
      if flag_2a_viatura >=0.5 then flag_2a_viatura=1;
      if TOM_BLOCKED_EVER <0.5 then TOM_BLOCKED_EVER=0;
      if TOM_BLOCKED_EVER >=0.5 then TOM_BLOCKED_EVER=1;

      drop index;
run;


/*************************************************************************************
****************************/
/* COMBINING NEW CASES WITH ORIGINAL CLAIMS AND SAVING
OUTPUT */
/*************************************************************************************
****************************/
data new2 (keep=&vlistAll.);
set new_cases2;
run;
data old2 (keep=&vlistAll.);
set claim;
run;
data newclaim2 (keep=&Target. &vlistAll.);
set old2 new2;
&Target.=1;
run;
data noclaim2 (keep=&Target. &vlistAll.);
set prep (where=(&Target.=0));
run;
/*data RCV_TRAIN_SMOTED;*/
data GTNVAUTO.RCV_TRAIN_SMOTED;
set newclaim2 noclaim2;
run;
```