NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Optimization of firefighter response with predictive analytics

Practical application to Lisbon, Portugal

Leonor Braz Teixeira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

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

OPTIMIZATION OF FIREFIGHTER RESPONSE WITH PREDICTIVE ANALYTICS – PRACTICAL APPLICATION TO LISBON, PORTUGAL

by

Leonor Pimentel Perestrelo Braz Teixeira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Co Advisor: Leid Zejnilović

Co Advisor*:* Miguel Castro e Neto

ACKNOWLEDGEMENTS

I would like to thank my advisors, Leid Zejnilović and Miguel Castro e Neto, for their help and guidance throughout this thesis.

CML for making this project possible, with the contribution of Microsoft, SAP, ISEL, and Altice.

Flávio Pinheiro for all the technical guidance and feedback.

Pedro Sarmento for the help with maps and geographical insights.

Alexandre Dias (from Microsoft) for setting the Databricks environment and helping the implementation of the cluster and nodes.

All my fellow firefighters at the volunteer firefighters for the debates regarding occurrences and relevant variables and the knowledge taught me throughout the years, especially Diogo Barata and Mariana Guerreiro for the care and words of encouragement.

My friends Inês Pereira and João Veiga for their support and encouragement in both the masters and thesis.

All my friends and family for the support and understanding in the toughest moments and all the motivation to continue.

January 2020

ABSTRACT

Time is a crucial factor for the outcome of emergencies, especially those that involve human lives. This paper looks at Lisbon's firefighter's occurrences and presents a model, based on city characteristics and climacteric data, to predict whether there will be an occurrence at a certain location, according to the weather forecasts. In this study three algorithms were considered, Logistic Regression, Decision Tree and Random Forest, as well as four techniques to balance the data – random over-sampling, SMOTE, random under-sampling and Near Miss –, which were compared to the baseline, the imbalanced data.

Measured by the AUC, the best performant model was a random forest with random under-sampling at 0.68. This model was well adjusted across the city and showed that precipitation and size of the subsection are the most relevant features in predicting firefighter's occurrences.

The work presented here has clear implications on the firefighter's decision-making regarding vehicle allocation, as now they can make an informed decision considering the predicted occurrences.

KEYWORDS

Emergency response; Firefighters; Predictive modeling; Smart city

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADASYN	ADAptive SYNthetic sampling
AUC	Area Under the Curve
CML	Câmara Municipal de Lisboa (Lisbon's Municipality)
INE	Instituto Nacional de Estatística (Statistics National Institute)
IPMA	Instituto Português do Mar e da Atmosfera (Portuguese Weather Institute)
LIME	Local Interpretable Model-agnostic Explanations
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
ROC	Receiver Operating Characteristics
RSB	Regimento Sapador de Bombeiros (Lisbon's Municipal Firefighters)
SHAP	SHapley Addictive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
VUCI	Veículo de Combate a Incêncio Urbano (Vehicle to Fight Urban Fires)

1. INTRODUCTION

In 1395, the first official response to fires in Lisbon was created by D. João I, which grew and changed through time to the current Lisbon's Municipal Firefighters (Regimento Sapador de Bombeiros, from now on RSB). Its scope of action grew from only fires to accidents, rescues and other events, having responded to over 9 000 occurrences in 2018. As human life was more valued and societies grew, more investment and research were made into responding to fires, accidents and medical predicaments. Under all of these emergent events, time plays a critical role.

Sampalis, et al. (1993) concluded that response time impacted chances of survival in cases of severe injury, increasing the odds of dying significantly when it took more than 60 minutes between the event and getting hospital care. Feero, et al. (1995) examined trauma cases with unexpected outcomes, that is, cases in which patients that were expected to die survived and vice-versa, concluding that shorter out-of-hospital EMS time intervals could represent an important factor in survival. A study by Lerner, et al. (2003) concluded that the total out-of-hospital time was actually not associated with mortality, although it observed differences in these times according to severity: more critical patients that survived had longer times than those that died, while more stable patients that survived had smaller times than those that died, contributing to the idea of the importance of prehospital care. Harmsen, et al. (2015) evaluated the out-of-hospital time with greater granularity concluding that in undifferentiated trauma patients shorter response times (from call to arrival on scene) and transfer times (from scene to hospital) had a positive influence in mortality. However, higher times on scene, where prehospital care is provided, increase chances of survival. Being that most of the out-of-hospital time is on scene, higher total prehospital times translated to smaller odds of dying. Although there is not a scientifical consensus on whether out-of-hospital time impacts mortality, the common practice is for emergency response vehicles to try to bring the patients to the hospitals as quickly as possible, while providing the needed care on scene and in route.

Fires start by a small and controllable flame, which can easily be put out. As time goes by, it grows becoming harder to control and spreading to adjacent rooms and building, while worsening the conditions for combat due to smoke and temperatures, besides structural damage. A critical factor in containing the spread and eliminating a fire is the response time from fire brigades, pivotal to save lives and minimize loss of property (Xin, J., & Huang, C. (2013)).

Considering the high importance given to response time in emergencies, that is, unexpected situations that require immediate action, there have been several studies on how to shorten it. As early as the 1970s, studies were made to evaluate resource allocation for fire engines (Kolesar, P., & Blum, E. H. (1973)) and rearranging temporarily the locations to allow better coverage when a brigade is deployed (Kolesar, P., & Blum, E. H. (1974)).

van Buuren, et al. (2015) modeled EMS call centers, allowing to understand the impact of adding an additional dispatcher and call taker, dependent on the number of requests and the priority of the calls. It is also important to take in consideration the factors that tend to influence the number of calls: overall trend, weekday, public holiday, the incidence of influenza in the previous week and of gastroenteritis (Viglino, et al. (2017)). Bandara, D., Mayorga, M. E., & McLay, L. A. (2014) found that sending ambulances according to the call priority could reduce the average response time by dispatching the closest available unit to the most critical calls and the less busy ambulance to non-life

threatening events, improving patient survival without increasing costs. Although this project refers to firefighters' occurrences, the call centers are very similar to the medical ones. Additionally, the literature regarding ambulance dispatch presents valuable knowledge, as the essence is the same as fire truck dispatch, in the sense that a fire truck can be redirected in case of a more severe call.

An algorithm (Nordin, et al. (2012)) was developed in C# to find the best route from any location (where the ambulance might be) to the incident site, in order to improve EMS response times. The location of emergency vehicles has been studied from several perspectives: initially it was attempted to get cover the maximum area, constrained on the number of ambulances (Church, R., & Velle, C. R. (1974); White, J. A., & Case, K. E. (1974)), then new factors were included, such as maximizing additional coverers and weighting the call frequency or population demand (Hogan, K., & ReVelle, C. (1986)). Later research made by the same pair included an estimation of busy ambulances determined by call frequency and the number of vehicles within the area, as well as duration of the call (Revelle, C., & Hogan, K. (1989)).

However, it is important to note that ambulances and fire truck allocation differ as an ambulance can respond to all medical emergencies, while fire engines and fire trucks might not respond to the same events and have different standard response times (ReVelle, C. (1991)). In that sense, it has been studied the allocation of engines, trucks and fire trucks in order to maximize the calls with both vehicles within the covering distance (ReVelle, C., & Marianov, V. (1991)). Specifically for the city of Lisbon, there are three main vehicles to consider: first response truck (VUCI), the tank and the stair vehicle. For any occurrence, the VUCI must arrive within 5 minutes of dispatch, as it can provide an initial response while the other vehicles are still en route.

Understanding the importance of a quick arrival on scene, a project was developed with CML to suggest the best locations for vehicles according to the prediction of occurrences, the expected traffic and any disturbance in the city's normal functioning (closed streets, construction work, …). This thesis focuses on the first part of the project, answering the question *What are the main factors influencing the occurrence of events that require firefighter rapid response in Lisbon?* and predicting those occurrences. The final aim of the project is to allow for data-driven decision making and improve the firefighter's response time and resource allocation, being therefore crucial for the model to be easily interpretable for a smoother adoption.

2. LITERATURE REVIEW

RSB aims to improve its response to fires, accidents, rescues, and calls regarding infrastructural issues, which can be achieved via two strategies, by better understanding the causes of each event in order to predict when it is more likely to occur or by improving the response time in dispatch and route. This thesis focuses on the prediction of occurrences.

2.1. PREDICTIVE ANALYTICS IN CITY MANAGEMENT

The capacity to store more data cheaply as well as the digitization of services have been two important drivers of Big Data, defined by De Mauro, A., Greco, M., & Grimaldi, M. (2016) as *the information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value*. Thanks to big data and the development of computational capacity, the area of predictive analytics has grown, aiming at discovering patterns and relationships in data through more realistic models. Emergent events and catastrophes are no exception, having been the purpose of such analysis. This study has allowed for a better understanding of the underlying causes of these events and to mitigate their consequences or even avoid them – a model of criminal activity was in the origin of a 47% decrease in citizen complaints of random gunfire, leading to an increase in public safety (McCue, C. (2006)). This section looks at the different sources of risk on which RSB must act, as well as the main studies for each source.

2.1.1. Fires

Most predictive models on the topic of fires concern wildfires fires and its propagation, as the potential economic loss is greater than in average structural fires (Madasseri Payyappalli, V. (2019)). A study done on 5 Portuguese municipalities found that land use is consistently not a relevant variable in predicting ignitions, while features as topography and distance to roads and to urban areas played a significant role in this model (De Vasconcelos, et al. (2001)). Once there is an ignition, it is crucial to understand how it develops and to where it will spread so that it can be contained and then extinguished. Also on this topic, Denham, et al. (2012) built a two-stage prediction model to decrease the uncertainty in input climacteric data and therefore improve predictions.

However, for the purpose of this thesis, it is more relevant to focus on structural fires, as RSB's area of action is the city of Lisbon, hence a smaller vegetated area. Facing over 3 000 major building fires yearly, New York felt the need to make an assessment of its city and find the characteristics which made a building more susceptible to fire. An algorithm was built with this purpose – FireCast is a Risk-Based Inspection Strategy which, based on data from over 5 years, evaluates 2 400 variables, weighted differently across the city, and provides a score for the risk of each building. From that, each fire station receives daily a list of the 15 buildings with higher risk to inspect, making those inspections 20% more accurate (Dwoskin, E. (2014, January 24)). More studies have been made regarding assessment of risk of buildings (Watts, Jr., J. M., & Kaplan, M. E. (2001); Watts, Jr., J. M., (2003); Lau, et al. (2015)), based on structural characteristics and violation history, but there hasn't been one aiming at predicting fires in a more holistic way (historical fires, weather, …) (Madasseri Payyappalli, V. (2019)).

2.1.2. Accidents

Road accident prediction models usually predict the total accident frequency, while some focus specifically on pedestrian accidents. Both Mountain, L., & Fawaz, B. (1996) and Greibe (2003) found motorized vehicle traffic flow as the most important explanatory variable for accident frequency, although the latter, for Denmark, also concluded that speed limit, road width, number of exits and of minor side roads, parking and land use were significant variables, which wasn't the case for the United Kingdom. Instead of traffic flow, a study for Shanghai, China, found that increased traffic volume was related to higher crash frequency (Wang, et al. (2015)). The usage of GPS data from taxis in this study, corresponding to around 20% of the overall traffic volume, also allowed to conclude that the average speed only had an impact on accidents during peak hours.

Regarding risk for pedestrians, Leden (2002) concluded that at high flow locations, right turns were safer than left turns, a difference non-existent in low flow locations, in Ontario, Canada. Additionally, this study also realized that the accident risk per pedestrian reduces with the pedestrian flow and increases with the vehicle flow. A study in Maine found that the risk faced by a pedestrian crossing in a high speed environment is almost 50 times higher than crossing in a low speed one, concluding that high speed not only increases the chances of an accident but is also related to the severity of such crash (Gårder (2004)).

Due to its history and circumstance, each city has grown in a different way, resulting in distinct urban plans, which means that the conclusions regarding road accidents on a specific location might not be true for all locations. A study for Lisbon's urban area found that traffic, lane balance, average lane width, the presence of right turn lanes, the traffic control devices, a high number of lanes and the number of legs of the intersection were associated with increases in accident frequency, while the number of legs with traffic in only one direction and median presence on major direction decreased it (Vieira Gomes (2013)). Its results were also in line with lower accident risk in areas with high pedestrian traffic, although the small sample size lead to low quality in terms of fitting of the models.

2.1.3. Infrastructural issues

Under this category, RSB responds mostly to trees falling, buildings collapsing and floods, so this literature review will focus on these three subtopics.

In 2009, an analysis was conducted in Lisbon on the tree falls from 1990 to 2005 during a windstorm (defined by having more than 3 trees falling in the same day) to which RSB responded, with the majority of the events occurring when wind velocity was greater than 7 m/s in the 6 hours preceding the fall (Lopes, et al. (2009)).

Over 30 years of building collapses in Nigeria were assessed by Ayodeji, O. (2011), who found that the main reasons for this were poor maintenance, design error, poor quality of materials and workmanship, natural phenomena and excessive loading. Inspections and building requirements are used to avoid the human error factor in these catastrophes. To mitigate the loss in case of natural phenomena, extensive research on the collapse risk of buildings under seismic forces has been conducted, finding structural damping, concrete strength and joint cracking strain as key fragilities in Memphis, USA (Celik, O. C., & Ellingwood, B. R. (2010)).

Regarding floods, a study in Australia found that catchment area and design rainfall intensity to be the two best predictors, being that the artificial neural network yielded better results with more stations, that is, when a larger dataset was available (Aziz, K., et al. (2014)). While that study made use of a quantile regression technique, other strategies have been employed, as Bayesian forecasting, which has shown promising results within a short time frame (less than 24 hours). In these studies, precipitation played a crucial role in the final model (Han, S., & Coulibaly, P. (2019)), being an important explanatory variable also for long-lead (5 to 15 days) extreme flood forecasting, along with other factors such as temperature and wind (Zhuang, et al. (2016)).

2.2. MODELING

Supervised machine learning techniques aim to predict a random variable, being classified according to it: regressors when the random variable is continuous and classifiers when it's discrete (Strecht, P., et al. (2015)). Three classifiers were considered for this project: logistic regression, decision tree, and random forest.

2.2.1. Logistic Regression

When a logistic regression is used for a binary problem, the outcome is the probability of success of an event, as the logistic function underlying presents a sigmoid shape, varying between 0 and 1 (Quinn, G. P., & Keough, M. J. (2002)). As a non-linear model, the logistic regression doesn't require the independent variables to follow a specific distribution nor any specific form (Chan, C. L., et al. (2010)).

Regularization is used to avoid overfitting, being the L1 and L2 penalties the main techniques. The Ridge Regression consists of applying an L2 penalty to the logistic function so that the variance of the estimate can be reduced with the introduction of the new bias. The Lasso Regression applies an L1 technique, allowing for feature selection, which is one of its main advantages, and differences from the L2 penalty (Owen, A. B. (2007)). However, choosing the L1 penalty might result in the loss of some accuracy when there are high correlations between predictors (Tibshirani, R. (1996)). The elastic net applies both L1 and L2 penalty to the model, overcoming the main limitations found in the Lasso regression: selecting only one variable when there are high pairwise correlations and the ill performance when there are more predictors than observations (Zou, H., & Hastie, T. (2005)).

Logistic regression has been extensively used in smart city predictive analytics, to predict freeway crashes (Abdel-Aty, M., et al. (2004)), estimate the outcome and severity of road accidents (Al-Ghamdi, A. S. (2002); Jones, A. P., & Jørgensen, S. H. (2003)), evaluate the risk of roof falling (Palei, S. K., & Das, S. K. (2009)) and even of landslides (Ohlmacher, G. C., & Davis, J. C. (2003); Ayalew, L., & Yamagishi, H. (2005)).

2.2.2. Decision Tree

A decision tree classifies an unknown observation using one or more decision functions sequentially, starting from a root node, which contains all possible classes, down to the feature nodes. Although the intermediary levels might have more than one class, the terminal nodes have only one class (Swain, P. H., & Hauska, H. (1977)). Decision trees are strong algorithms, able to handle missing values, imbalanced classes, and redundant attributes at a low computational cost. In order to improve performance, hyperparameters can be tuned using a range of processes, such as grid search,

random search, particle swarm organization or estimation of distribution algorithms. However, when performing said processes, it is important to consider that the tuning that yields the model with the highest performance in a dataset might not perform so well on other datasets (Mantovani, R. G., et al. (2016)).

A decision tree was one of the models used to predict wildfires in Slovenia, being the best performer with bagging (Stojanova, D., et al. (2006)). It has also been used to analyze road accidents (Shanthi, S., & Ramani, R. G. (2012)) and even yielded an AUC of almost 90% when predicting landslides (Nefeslioglu, H. A., et al. (2010)), although it is more common to be a part of a larger algorithm, random forests.

2.2.3. Random Forest

Ensemble methods combine several models to build a better one, making use of the wisdom of the crowd principle. Random forests are one of those cases, using bagging of decision trees: models are trained separately and each one votes on the output for the test examples, which is reached by majority voting (in case of a classification problem) or by averaging (in case of regression). This model differs from decision trees on the splitting mechanism, as each bootstrap sample grows an unpruned tree from a different set of features – only a random subset is considered for splitting, increasing the variability of the trees and differing from pure bagging of decision trees (Breiman, L. (2001); Liaw, A., & Wiener, M. (2002)). Besides the hyperparameters used to tune decision trees mentioned above, random forests also allow to adjust the strength of randomization with the number of estimators used and the number of features considered at each subset, which in turn is dependent on the relevancy of features (Bernard, S., Heutte, L., & Adam, S. (2009)).

Besides the events mentioned in the above section, random forests have also been used to predict wildfires in Slovenia (Stojanova, D., et al. (2006)) and in Austria (Arpaci, A., et al. (2014)).

2.2.4. Performance metrics

In machine learning, models are created with the aim of improving a performance metric and better predicting or representing something in the world. Being so, the choice of criteria is crucial for the end result and must be done taking into consideration the problem at hand (Caruana, R., & Niculescu-Mizil, A. (2006)). Accuracy, precision, recall, F_1-score, receiver operating characteristics curve and area under the curve were considered.

Accuracy is the ratio between correct predictions and the total number of predictions. Although it has extensively been used to evaluate the performance of a model, it has been shown to not be appropriate in cases where the data is imbalanced or the costs of the errors are very different (Chawla, N. V., et al. (2002); Chawla, N. V. (2009); Hossin, M., & Sulaiman, M. N. (2015)).

A common tool to evaluate a classifier is a confusion matrix (Table 1), which contains the predicted results in the columns and the actual target values in the rows, such that the number of correctly identified events are the True Positive (TP), in case of positive examples, and the True Negative (TN), in case of negative examples. Misclassifications are called False Positive (FP) when a negative event is predicted as positive and False Negative (FN) when a positive event is predicted as negative (Chawla, N. V., et al. (2002)).

	Predict negative	**Predict positive**
Actual negative	True Negative (TN)	False Positive (FP)
Actual positive	False Negative (FN)	True Positive (TP)

Table 1 - Confusion matrix format

Recall is the ratio of positive events that were correctly predicted, being represented by TP/(TP+FN), and precision is the ratio of predicted as positive events that were actually positive, being represented by TP/(TP+FN). Buckland, M., & Gey, F. (1994) proved that there is a trade-off between precision and recall, being that both being 1 would be the desirable point. The F-score is a combination of these two metrics, being F_1 = 2 x Precision x Recall / (Precision + Recall) (Goutte, C., & Gaussier, E. (2005)). Although Han, H., Wang, W. Y., & Mao, B. H. (2005) and Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006) considered F_1 to be a good metric for imbalanced data, this was later disproved by Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013), which found it to be affected by the imbalance, but only when the negative examples were the majority.

The receiver operating characteristics curve, known as ROC curve, plots the False Positive Rate, which is the ratio of incorrect predictions of actual negative events and can be represented by FP/(FP+TN), against the True Positive Rate, which is the recall, for each threshold. At random, it will be a diagonal between (0,0) and (1,1) while the perfect classification is (0,1). One of the advantages of this metric is that it is not affected by changes in class distribution (Fawcett, T. (2006)). To facilitate the comparison of models, the area under the ROC curve is used, known as AUC. Although this metric represents many advantages, as being objective, it has been considered to apply different misclassification costs according to the classifier being used (Hand, D. J. (2009)). Even so, AUC and ROC are the most used performance metrics in rare event prediction (Haixiang, G., et al. (2017)).

2.3. DATA HANDLING

2.3.1. Missing values

When working with real data, a common issue faced is missing values, which can be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (Rubin, D. (1976)) When data is MCAR, no relationship can be found between the missing data and the observed data, so there is independency. A less strict categorization is MAR, when there is a relationship between the missing data and other variables, but not the missing variable in itself. When the relationship is with the outcome variable, then it is MNAR (Enders, C. K. (2011)), and the missing data mechanism cannot be ignored and must be studied – a common example is the fact the people with higher incomes are less likely to report them in surveys (Soley-Bori, M. (2013)).

Three main techniques can be used to handle the issue of missing data: deleting the observations with missing values, estimating the parameters with maximum likelihood procedures and imputation of the values (Gad, I., & Manjunatha, B. R. (2017)). Although case deletion has advantages, such as its easy implementation and little impact on statistical analysis, it might result in the removal of a considerable amount of complete data and would require for the data to be missing completely at random (Enders, C., & Bandalos, D. (2001)). Alternatively, it is possible to use maximum likelihood estimation techniques to obtain the model with the observed data, through direct maximum likelihood (maximization of the multivariate normal likelihood function) or expectation-maximization algorithm (the maximum likelihood estimates are obtained through iteration of the expectation and

maximization step). This more efficient method tends to be computationally heavy and more sensible to outliers (Allison, P. D. (2001); Soley-Bori, M. (2013); Gad, I., & Manjunatha, B. R. (2017)). Imputation of missing data can be conventional, when the values are substituted by a reasonable guess (mean, median, last observation carried forward …), or advanced, as multiple imputations, which uses the values available to predict the missing data, the most robust strategy, especially with MAR data. However, imputation poses a relevant problem: different outcomes can be achieved with the same method and same initial data (Allison, P. D. (2001); Soley-Bori, M. (2013); Nakai, M., & Ke, W. (2011)).

Within the realm of meteorology and data science, Gad, I., & Manjunatha, B. R. (2017) conducted a research to predict missing values in weather data using machine learning techniques, concluding that imputing the missing data with 0 or with a constant value might generate noise and outliers; also removing the attributes with the missing data would impact the performance of the model, as well as decrease the size of the dataset. Having tested the kernel ridge, linear regression, random forest, support vector machine and k-nearest neighbors imputation procedures, Gad, I., & Manjunatha, B. R. (2017) found that random forest was the best method for wind speed, and performing well with other meteorological features too.

2.3.2. Feature normalization

The different ranges of variables result in some similarity measures, as the Euclidean distance, attributing different weights to them, requiring feature scaling for all to have the same effect. Min-max scaling is one of the methods used, where all observations are scaled to between 0 and 1 by subtracting the maximum value and then diving by the original range (Aksoy, S., & Haralick, R. M. (2001)). To guarantee that all features have a mean of 0 and unit variance, standardization, or normalization with z-scores, can be used, subtracting the sample mean to each observation and then dividing by the sample standard deviation (Dubes, R. C., & Jain, A. K. (1988)). While the first approach maintains the distribution but scales its range, the second approach changes the observations to follow a normal distribution, which is a requirement for some algorithms and statistical testing. When using either approach, it is important to rescale the data points in the end for interpretation (Bruce, P., & Bruce, A. (2017)).

2.3.3. Sampling

Random sampling is commonly used in cases where using the entire population is expensive and computationally heavy (Olken, F. (1993)), being its different types classified according to: method to determine the sample size, whether it is done with replacement or not and random or sequentially, if the population size is known and whether each observation has uniform inclusion probability (Olken, F., & Rotem, D. (1986)).

Simple random sampling without replacement is the simplest way to do so, when *n* units are drawn randomly from the population not already drawn, with an equal chance of selection (Cochran, W. G. (2007)).

2.3.4. Feature selection

Reducing both computation and storage requirements and facilitating data visualization and understanding are some of the potential benefits of feature selection (Guyon, I., & Elisseeff, A.

(2003)). The literature on the topic state that a good set of features must be relevant and avoid redundancy, so the independent variables should be highly correlated with the target variable, while uncorrelated with each other (Hall, M. A. (1999); Toloşi, L., & Lengauer, T. (2011)).

2.3.5. Imbalanced data sets

Real world data sets have imbalanced datasets, where one class, usually the one to be predicted, is under-represented. Instead of training with the original distribution, opting to use a sampling strategy to balance the representation of the minority class has led to more accurate predictions than the unbalanced split (Özçift, A. (2011)).

2.3.5.1. Over-sampling

Over-sampling allows to balance classes by repeatedly sampling the minority class, which can be done through three main strategies: random over-sampling, SMOTE, and ADASYN. Random over-sampling with replacement consists of increasing the minority class by duplicating its observations the necessary folds, which might lead to over-fitting of the learner, as the decision regions become more specific (Ling, C. X., & Li, C. (1998)).

Synthetic Minority Over-sampling Technique (SMOTE) consists of creating new synthetic observations using k-nearest neighbors (Chawla, et al. (2002)), having the opposite effect on decision regions: they become broader as there are more observations to learn from. Some variations exist to this method, as only over-sampling along the borderline (Han, H., Wang, W. Y., & Mao, B. H. (2005)) or using k-means instead of the k-nearest neighbors to create the new observations (Douzas, G., Bacao, F., & Last, F. (2018)). Finally, Adaptive Synthetic Sampling (ADASYN) is built on top of SMOTE, using a weighted distribution for different minority observations, such that more synthetic data is generated for the observations that are harder to learn (He, et al. (2008)).

2.3.5.2. Under-sampling

Under-sampling balances the training set by using only a subset of the majority class, which also contributes to faster training and less computational requirements. However, these advantages come at an informational cost, as potentially relevant data is not considered in the model (Liu, X. Y., Wu, J., & Zhou, Z. H. (2008)). Similar to the over-sampling strategy, random under-sampling randomly selects observations from the majority class and removes them from the dataset until the desired ratio of imbalanced is achieved (Prusa, J., et al. (2015)).

The Near Miss algorithm aims at increasing the distance between the classes by removing observations from the majority class, leading to a more balanced dataset. It can achieve such purpose through three strategies: version 1 removes the majority class instances whose average distance to the minority class's closest observations is the smallest, therefore removing those that are closer to some minority instances; version 2 removes the majority class instances whose average distance to the farthest minority class observations is the smallest, therefore removing those that are closer to all the less represented class; version 3 removes the majority class instances that are closest to each minority class observation, based on the k-nearest neighbors approach (Mani, I., & Zhang, I. (2003)).

2.4. MODEL INTERPRETATION

As machine learning evolved, models became more complex and harder for the user to comprehend. This increased opacity of models led to an increase of resistance to adopt their suggested outcomes, meaning that models with worse performance were being chosen due to their transparency and easy interpretability (Lundberg, S. M., & Lee, S. I. (2017)).

Quantitative input influence measures were one of the solutions used to increase the algorithmic transparency in systems that process personal information by capturing the level of influence of each feature on the prediction (Datta, A., Sen, S., & Zick, Y. (2016)). Local Interpretable Model-agnostic Explanations (LIME) is another alternative, aiming at locally finding a representation of the classifier that is interpretable to humans (Ribeiro, M. T., Singh, S., & Guestrin, C. (2016)), with the great advantage of being useful for any type of model.

More recently, Shapley addictive explanations has been created as a unified framework used to understand predictions of complex models, with three important properties: local accuracy (for a specific observation, it will perform at least as well as the original model), missingness (a feature missing in the original model will have no impact in the model used for interpretability) and consistency (the attributed influence of a feature should not decrease when a model changes due to an increase in that some feature's contribution nor due to that feature remaining constant, regardless of the other variables). The SHAP values show how to get from the base value (what would have been predicted if none of the features were known) to the output of each observation by attributing to each feature its contribution to the change in the prediction. This approach not only is more efficient in computational terms, as it also yields more accurate results and humans understood them better than other methods, as LIME (Lundberg, S. M., & Lee, S. I. (2017)).

3. RESEARCH CONTEXT AND DATA METHODOLOGY

3.1. CONTEXT

Time plays a crucial role in any emergent event, with a rapid response being key to an effective and efficient resolution. Being so, RSB aims to reach any call within 5 minutes of dispatch, which is greatly affected by the number of teams available at any moment and the location where each is at the moment of dispatch.

Currently, decision makers use their intuition and knowledge of disturbances that affect traffic (like marathons and construction work) to relocate their vehicles, although this is rarely done due to the logistic implications. Additionally, when more extreme weather conditions are expected (as heavy rain and winds), more teams are put on call to be able to answer an expected increase of requests. Besides municipal firefighters, the city of Lisbon also has 6 volunteer corporations which, under specific cases, can be requested to increase their response capacity. However, that is dependent on each corporation's availability. Unfortunately, no tracking is made of location changes nor of the number of teams available at each point in time, being only known the baseline.

The desired output of this project is a probability of occurrences throughout the city for each moment, which will later be combined with mapping software to evaluate potential locations for vehicles according to predicted traffic. With this information, decision makers will be able to decide the resources needed to respond to the expected requests with more accuracy and within a shorter timeframe, allowing them to make data-driven decisions. Once in use, the model generated can be evaluated by comparing the predicted occurrences with the reality and assess whether the suggested locations improved response times, compared with leaving from the station.

3.2. DATA METHODOLOGY

RSB is responsible for answering a broad spectrum of events: fires, accidents, rescues, infrastructural issues, medical emergencies, and safety checks, among others. However, not all of these events are considered emergent, therefore not falling under the scope of this project. Additionally, medical emergencies require a different technical team and vehicle, being managed by another department. A complete list of the types of calls RSB responds to with a target column that identifies as 1 those that are relevant to this project is available in the appendix (Table 53).

An analysis of predictive analytics on the events under the scope of this report (fires, accidents, and infrastructural issues) state-of-the-art was conducted in the literature review. Topography, road-urban distribution, building characteristics, and climacteric data were the main features to predict fires, both ignitions and spread. Regarding road accidents, the important independent variables were mainly related to road characteristics, such as the number of lanes and speed limit, and on pedestrian and vehicle flow. For infrastructural issues, building characteristics played a relevant role, as well as climacteric data, especially wind, precipitation and temperature. It was not possible to find any relevant literature on prediction of rescues.

Following the findings, three datasets were used to conduct this project: occurrences from the RSB, climacteric data and the census from 2011, which describe the population and building characteristics of the city split into 3 662 subsections. It was not possible to gather updated data

regarding the roads nor populational density, neither pedestrian nor vehicular. Additionally, the most recent data regarding the buildings was from 2011, not fully reflecting the current state of the city.

A partnership between Nova Information Management School and Lisbon's Municipality allowed to have access to the datasets used: occurrences from RSB and meteorological data from the Portuguese Weather Institute (Instituto Português do Mar e da Atmosfera, from now on IPMA). Additionally, data from the Statistics National Institute (Instituto Nacional de Estatística, from now on INE) available online was used to characterize the city.

Prior to modeling, it was necessary to evaluate the data quality in terms of missing data, outliers and duplicates. Once it was clean, an exploratory analysis was conducted to better understand the distribution of occurrences throughout the city and to assess any patterns between them and the climate.

The three datasets were combined such that, for each subsection of the city, each observation was an hour of the day, with the respective climacteric data, characterization and a target variable corresponding to whether or not there was an occurrence in such location at that hour. Due to the high volume of data (every hour from January 1st of 2013 to December 17th of 2018 corresponds to over 52 000 observations per subsection), it was necessary to sample the data: 100 subsections were randomly chosen. This sample was then used to perform standardization, scaling to zero mean and unit variance, and feature selection: a correlation analysis between the variables was made and, iteratively, the feature with smaller absolute correlation with the target among the highest correlated pair in absolute values was eliminated, being only kept features with absolute correlations under 0.8.

The occurrence of fires, accidents, rescues or infrastructural issues is a rare event, which results in the final dataset being heavily imbalanced: few positive occurrences to many negative ones. In these cases, it is expected for the model to perform poorly than it would if it were balanced (Özçift, A. (2011)). Therefore, the modeling approach can be divided into three strategies: first it was attempted to model using the imbalanced data, then it was attempted to over-sample the minority class and, finally, it was attempted to under-sample the majority class. Although many techniques exist, it was only possible to use two for each strategy due to computational limitations: random over-sampling, SMOTE, random under-sampling, and Near Miss.

Many algorithms can be used to tackle a binary classification supervised problem such as the presented one. Supported by the literature, logistic regression, decision trees, and random forests were chosen, as they tend to perform well under these problems and were computationally feasible for the available resources.

When using the imbalanced data, it was only necessary to split the dataset into a train, where each algorithm was trained using cross validation to find the best hyperparameters, and a test set, where it would be tested. Due to the high level of imbalance present in rare events, it is expected that all observations are predicted as negative in these cases. On the other hand, when trying to balance the dataset, for each sampling strategy (over and under-sampling), the dataset was split into three: a train, a validation, and a test set, being that the sampling technique was only applied to the first one. For each ratio of imbalance, the model was trained on the training set using cross validation to find the best hyperparameters, which was then run on the validation set to get the performance metrics.

The best ratio of imbalance for that specific model and sampling technique would be found according to those performance metrics and then the model would be evaluated according to its performance on new data – the test set.

As presented in literature, the fact that the data is imbalanced poses a big concern on the choice of evaluation metric, as criteria as accuracy assign equal cost to both false negatives (FN) and false positives (FP), therefore leading to considering a model that predicts all observations as negative as the best – an undesirable outcome. Area under de curve was found to be the least affected by the imbalance of data (Haixiang, G., et al. (2017)), being the chosen criteria to evaluate the performance of the models.

Additionally, it was important to also evaluate the performance of each model on different subsections than those used to train and evaluate it, in order to understand each model's capacity to generalize to the entire city. Following the reasoning applied previously, a new sample of 100 different subsections was randomly chosen and prepared according to the same process. Each model of each technique was then used to test the generalization capacity.

The best performant model was chosen based on its AUC on the test set and ability to generalize to the new sample. The time-independency assumption was then validated by training the winning model on the data from 2013 to 2017 and then test it on 2018, applying the most appropriate sampling strategy. Finally, in order to guarantee that it would be an appropriate model to apply to the entire city, 10 new samples of 100 different randomly selected subsections were created and the model was tested for each of them.

Having found and validated the best performant model, SHAP values were calculated to understand the role and importance of each feature in the final prediction, allowing for a needed better interpretation of the model. This was an important step to increase the chances of adoptability of the suggested locations by the firefighters, as it made the model more transparent.

All the analysis and modeling were made using Python on Jupyter Notebook, resorting to Spark in Databricks when the computational needs required it. The maps were made with ArcGis.

4. DATA AND EXPLORATORY ANALYSIS

The following section will present each dataset used in the model and its handling in terms of scope and missing values. The exploratory analysis conducted presents the distribution of occurrences across the city for the years at study (2013-2018), succeeded by a closer analysis of the last year's data relation with climacteric data and time variability. Before building the final dataset, the data is standardized, and the redundant features are eliminated by correlation analysis. A random sample of 100 subsections is then created for the modeling process.

4.1. RSB OCCURRENCES

RSB provided a historical dataset with all the occurrences their firefighters responded to from August 19, 2011, to December 17, 2018. These can span from fires and accidents to medical emergencies and visits for equipment check. As the scope of this project is to improve the response time of RSB to emergent events, all non-emergent events weren't considered. Being so, all fires, accidents, urgent rescues and infrastructural issues were considered as positive, while the remaining events were considered as negatives – a list of all occurrence types and respective target variable in the appendix (Table 53). It was chosen to not use the data from 2011, as it was very incomplete (information was only available for one third of the year), as well as from 2012, since 87% of the observations did not include the type of call, therefore not being possible to classify the target variable. There were no duplicates in the remaining 55 871 observations.

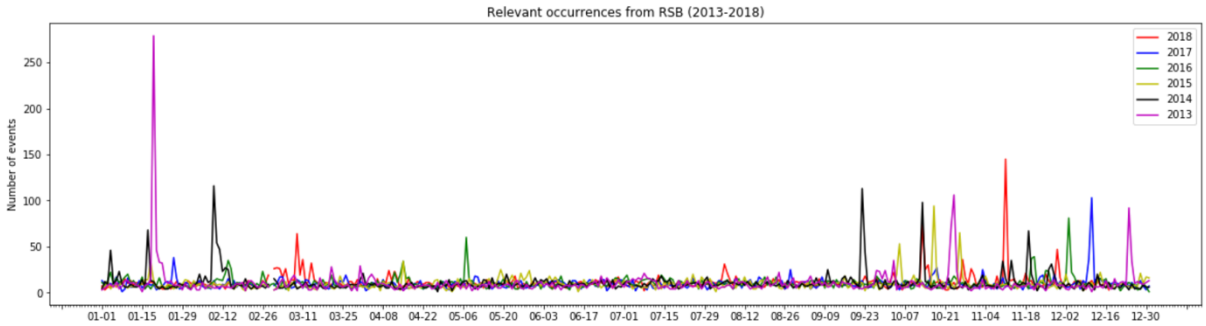

Figure 1 - RSB relevant occurrences from 2013 to 2018

Figure 1 above represents the distribution of relevant occurrences throughout each year, being that each color represents a specific year: 2018 in red, 2017 in blue, 2016 in green, 2015 in yellow, 2014 in black and 2013 in magenta. It is noticeable that there tend to be peaks at the end of the year, throughout the last quarter and even more in the month of November. In addition, it is interesting to note that there is an unusual peak in the middle of January 2013.

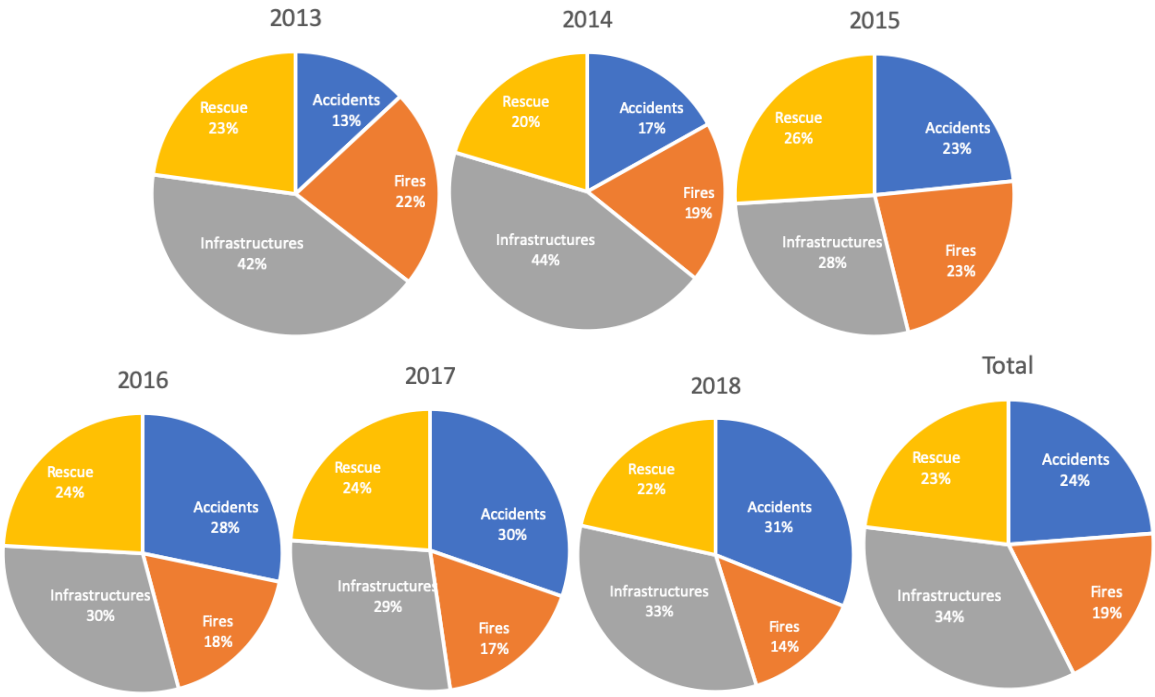

Figure 2 - Distribution of relevant occurrences from RSB (2013 - 2018)

Throughout time, rescue activities (includes human and animal search and rescue on both land and aquatic environments, as well as opening doors with a suspect of a person trapped inside) remained stable at around 20 to 25% of the total occurrences RSB responded to. Infrastructural issues, which include trees falling, floods, structures falling, landslides and loose electrical cables, predominated, reaching 44% in 2014. Unfortunately, the amount of accidents has been proportionally increasing, reaching almost one third of the total relevant events in 2018. Although the creation of RSB was a response to fires in the city of Lisbon, this is actually the less recurrent event type, proportionally decreasing each year.

4.2. Census Data

Every 10 years, INE collects information regarding the population and the buildings, allowing to better understand the country and contributing to its planification, in terms of schools, hospitals, safety, and transportation, for example (Instituto Nacional de Estatística. (n.d.)). Lastly collected in 2011, census is mostly self-reported data, as each house answers a survey with multiple questions accompanied by an instruction manual. Even though it is mandatory by law to answer truthfully to the surveys (Lei nº22/2008, article 26º (Instituto Nacional da Casa da Moeda. (n.d.))), census data is bound to contain some error which is important to take into account (Gonzalez, M. E., et al. (1975)).

The data is collected per house, but is only available in aggregate, so as to guarantee the anonymity required by law. For the specific case of Lisbon, the city is first divided into parishes, then sections and finally subsections, reaching a total of 3 662 portions.

Following the literature, three types of information were considered: the area of the subsection, the building characteristics and the population that lives in it. The names and descriptions of the features are available in the appendix (Table 52).

4.3. METEOROLOGICAL DATA

IPMA provided three datasets from the stations within the city's limits: Geofísico, Gago Coutinho and Tapada da Ajuda. The six features provided were: average air temperature (Celsius degrees), relative air humidity (percentage), average wind direction (degree), average wind speed (meters per second), total precipitation (millimeters) and total sun radiation (kilojoule per square meter). The dataset has an hourly granularity, from 2013 to 2018, corresponding to a total of 52 584 expected observations per variable. However, there are many missing values, as accounted in Table 2 below.

Feature	Geofísico	Gago Coutinho	Tapada da Ajuda	Missing in all stations
Temperature	6 440 (12,25%)	263 (0,5%)	495 (1%)	1
Humidity	13 328 (25,34%)	241 (0,45%)	495 (1%)	1
Wind direction	45 868 (87,22%)	296 (0,56%)	52 584 (100%)	271 (0,5%)
Wind speed	45 866 (87,22%)	277 (0,5%)	11 215 (21,3%)	9
Precipitation	6 811 (12,95%)	251 (0,5%)	715 (1,36%)	2
Sun radiation	46 246 (87,9%)	252 (0,5%)	556 (10,6%)	5

Table 2 - Missing data for each station from IPMA in absolute values (% of the total data)

The amount of missing data made it unfeasible to discard neither the observations nor the features that were incomplete, so it was imputed with the data from the closest station: Tapada da Ajuda and Geofísico were the closest ones at 3 067m, followed by Gago Coutinho and Geofísico at 5 577m. Finally, Tapada da Ajuda and Gago Coutinho were the farthest at 7 917m. Having imputed those values, 289 were still missing. As it is uncommon to have abrupt changes in meteorological data, the observation of the previous or the next hour was used to impute, up to a limit of 3 imputations, resulting in only wind direction presenting missing data: 209 observations. Considering the literature (Gad, I., & Manjunatha, B. R. (2017)), a random forest was used to predict those values based on the remaining weather conditions at the time.

4.4. EXPLORATORY ANALYSIS

4.4.1. Geographic distribution of occurrences

Each occurrence in the dataset is associated with a pair of coordinates, corresponding to a specific location, which was then linked to a subsection using polygons, allowing to understand the distribution of occurrences throughout the city.

A hotspot analysis (How Hot Spot Analysis (Getis-Ord Gi*) works, n.d.) consists of computing a z-score, such that a hotspot must have a high value and be surrounded by other high values. The higher the value, the more intense is the hotspot. Similarly, cold spots are associated with low negative values.

Figure 3 presents a hotspot analysis made to the relevant occurrences from 2013 to 2018, showing hotspots in the area of the airport, Monsanto, Penha de França, Avenidas Novas, Campo Grande, and Benfica and cold spots in the areas of Belém, Baixa and Olivais, for example.

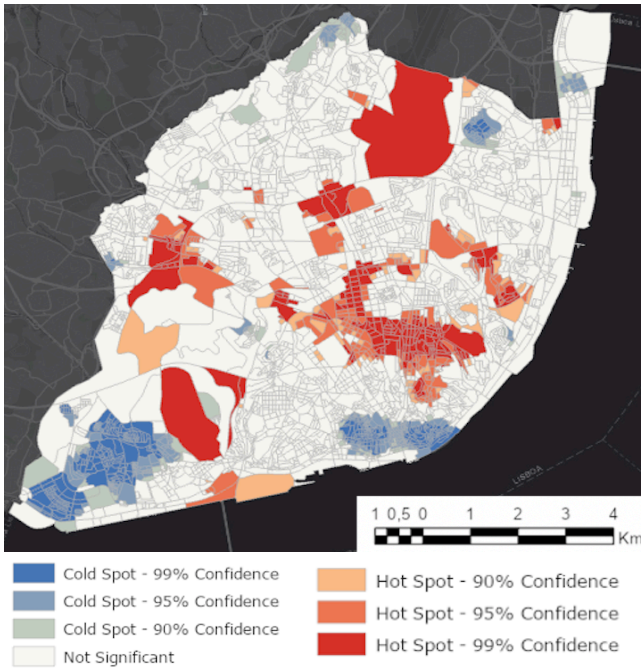

Figure 3 - Hotspot analysis (2013-2018)

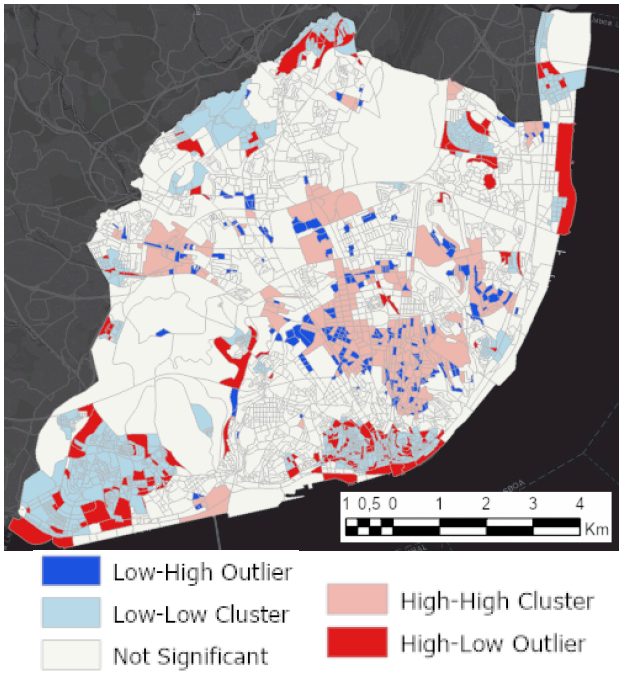

Figure 4 - Cluster-outlier analysis (2013-2018)

A cluster-outlier analysis (How Cluster and Outlier Analysis (Anselin Local Moran's I) works, n.d.) allows to identify two types of clusters: of high values, which consist in high values among high values, known as high-high cluster and cluster of low values, which consist in low values among low values, known as low-low clusters; and two types of outliers: high values among low values, known as high-low outliers, and low values among high values, known as low-high outliers.

Comparing Figure 3 with Figure 4, one can see that there are some comparatively high values in the area of Belém, marked in red as high-low outliers among a light blue low-low cluster, and that there are some comparatively low values in the area of Avenidas Novas – Penha de França, marked in blue as low-high outliers among a light red high-high cluster. These outliers among a cluster mean that even though an area has more relevant occurrences, there are some subsections with fewer events than the neighboring ones.

4.4.2. Year of 2018

An exploratory analysis of the occurrences in the year of 2018 was made to understand the overall volatility across the year, as well as of specific events: fires, accidents, rescues, and infrastructural issues. Additionally, this data was crossed with weather data to detect visual patterns between them and correlations.

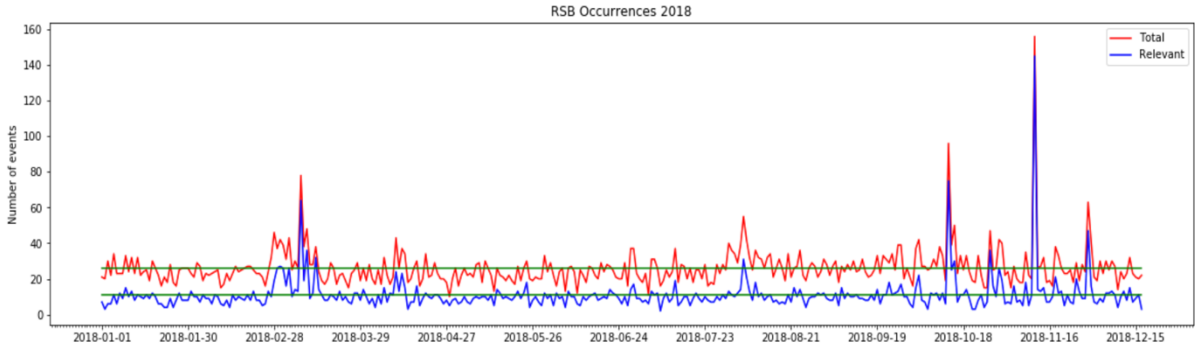

Figure 5 - RSB occurrences from 2018 (total and relevant)

From Figure 5 above, it is possible to detect a few peaks, in particular, November 11 (with 145 occurrences), October 13 (at 75), March 9 (at 64) and November 29 (at 47). On average, there are 26.22 events per day, being that 11.19 of those are considered as relevant by the chief. Further analysis showed that the peak of November 11 was due to a peak in floods (79 calls for private spaces and 49 for public ones), the peak in October 13 was mostly related to trees (20 calls) and buildings (27 calls) falling, while the peak in March was once again related to private spaces being flooded (32 calls). Being so, it seems that peaks in events, seem to be caused by weather condition changes, making it relevant to analyze both information together.

Figure 6 - RSB occurrences (total and relevant) and weather from 2018

When plotting the same occurrences against weather factors for the same period, as in Figure 6, it is noticeable that the peaks in events tend to follow peaks in precipitation (two bottom plots of Figure 6). However, a more careful analysis is necessary prior to conclude such relationship.

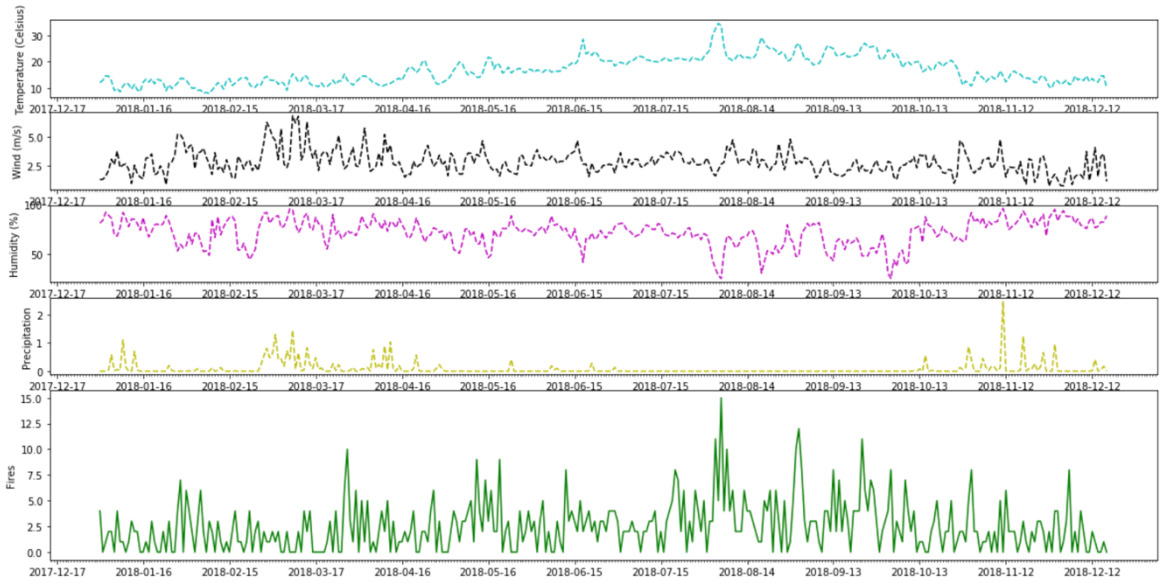

Figure 7 - RSB fire responses and weather from 2018

The peak of fires was at 5 of August (15 occurrences), usually ranging between 0 and 5 per day. Visually it is hard to detect any pattern between fires and weather conditions in Figure 7, although there is a 0.35 correlation with temperature and of -0.37 with humidity.

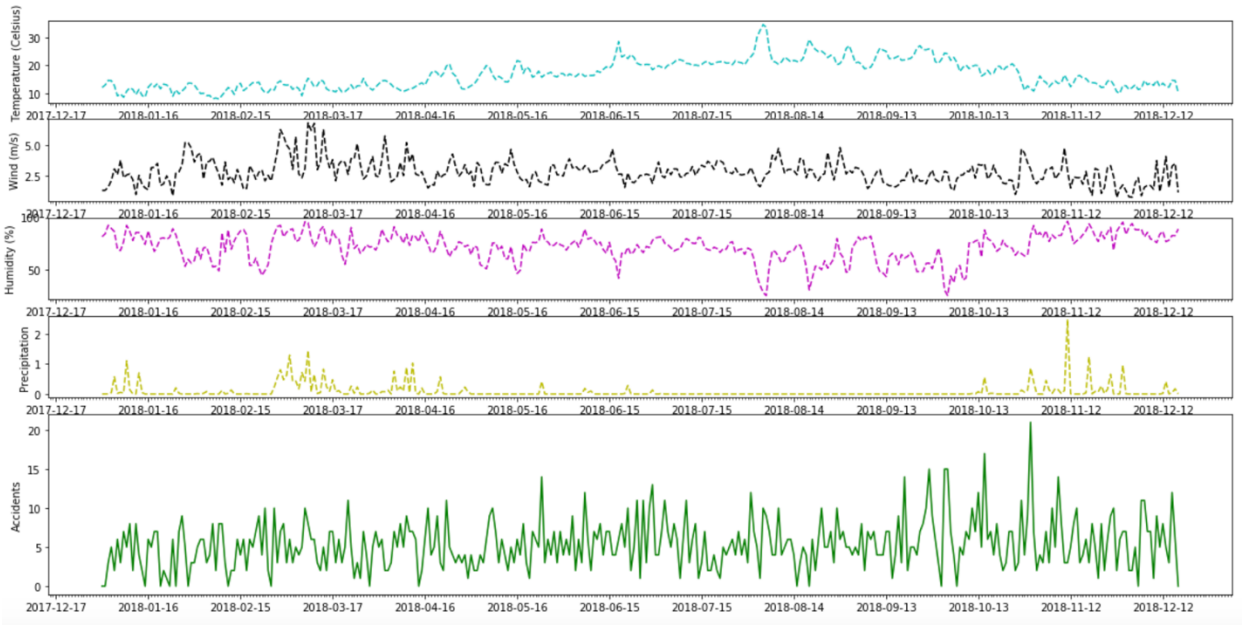

Figure 8 - RSB accident responses and weather from 2018

October 30 was the day with more accidents (21 occurrences), followed by 17 two weeks earlier, as plotted in Figure 8. On average, there were less than 6 accidents per day, being most common in the last months of the year, especially October. No pertinent correlation was found between accidents and meteorological factors.

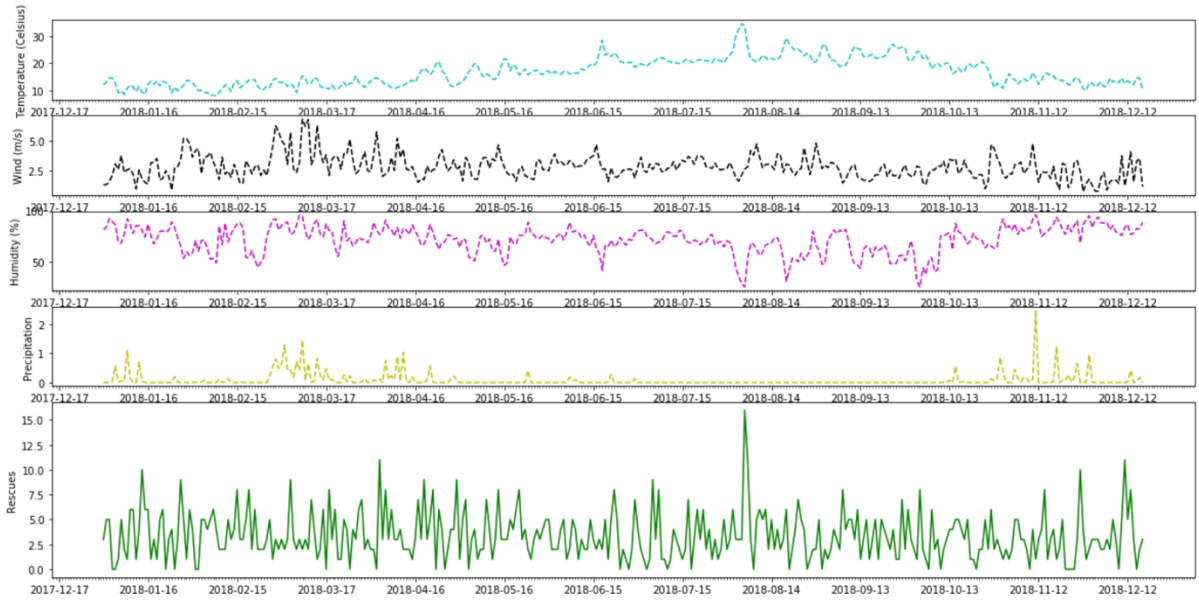

Figure 9 - RSB rescue responses and weather from 2018

August 5 was the day when most rescue calls surged (at 16) compared to an average of 4. Rescues don't seem to correlate with meteorological factors and tend to be quite uniform across the entire year, as visible in Figure 9.

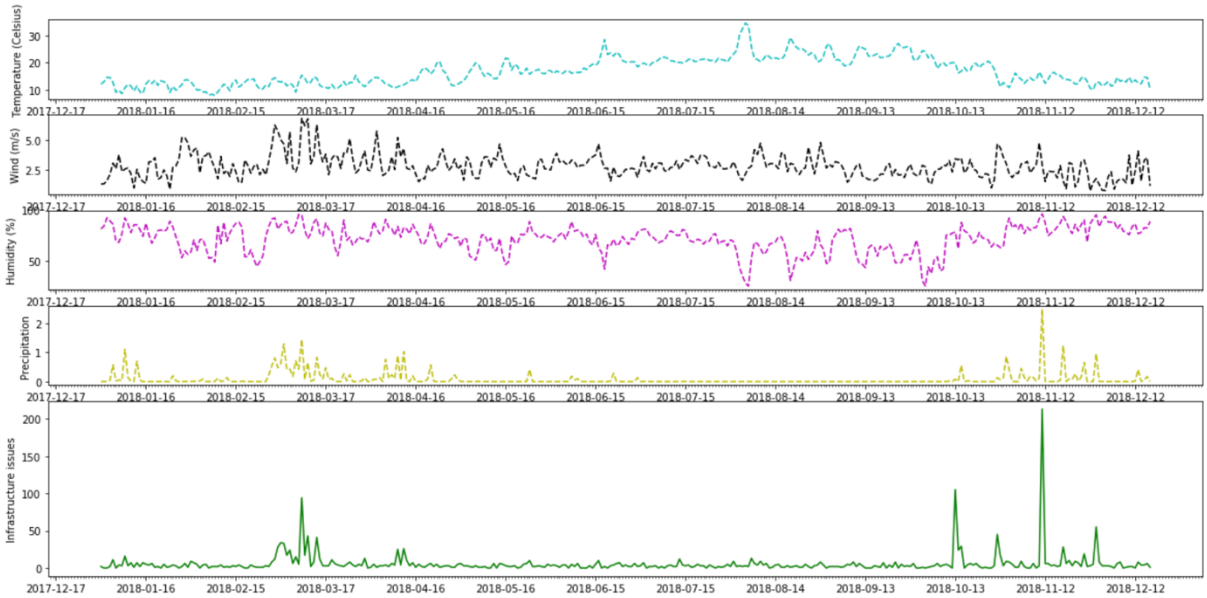

Figure 10 - RSB responses to infrastructural issues and weather from 2018

It is under infrastructural issues that more volatility is found, with a great peak at November 11 (213 calls) and two smaller ones on October 13 and March 9 (105 and 94 calls, respectively) – Figure 10. This category encompasses the most diverse type of events: 41% are landslides and structural collapses, 28% are floods, 18% are trees falling and 12% refer to damage with electrical cables. The visual similarity between these events and precipitation throughout 2018 is supported by a correlation of 0.73 (raises to 0.743 when only floods are considered). Literature found that the majority of trees falls occurred when winds were above 7 meters per second in the previous 6 hours

(Lopes, et al. (2009)), which is aligned with the positive correlation found between trees falling and wind.

4.5. DATA PRE-PROCESSING

The dataset has 55 871 observations, of which 22 287 are positive events. Being the aim to predict occurrences, the data was organized per hour, which meant that each observation corresponded to an hour of the year. The end goal of this project is to reduce the response time of RSB vehicles to occurrences, so the location of the predicted events is crucial: a too granular model would make the data very sparse, while a macro one would yield less valuable results.

A dataset was built for each subsection, the highest level of granularity of the census data available, with one observation per hour. As expected, these constitute very unbalanced datasets, with only 110 positive observations in a total of 52 248 for the subsection with most occurrences. A random sample was built with 100 randomly selected subsections, resulting in a total of 5 224 813 observations and 92 features, both meteorological and demographic. The sample was then standardized by removing the mean and scaling to unit variance.

Following the literature (Hall, M. A. (1999)), the correlation of the features was computed – matrix in Figure 39 in the Appendix. The table below shows the pairs of variables with the highest correlations, as well as which one from the pair which was eliminated – the one with the lowest absolute correlation with the target variable. As this was an iterative process, as a variable is eliminated, the correlations it would have with other features are no longer relevant – Table 3.

Variable 1	**Variable 2**	**Correlation**	**Eliminated**
N_INDIVIDUOS_RESIDENT_20A64	N_INDIVIDUOS_RESIDENT_25A64	0,9987	1
N_INDIVIDUOS_RESIDENT_M_20A64	N_INDIVIDUOS_RESIDENT_M_25A64	0,9986	1
N_INDIVIDUOS_RESIDENT_H_20A64	N_INDIVIDUOS_RESIDENT_H_25A64	0,9983	1
N_IND_RESID_EMPREGADOS	N_IND_RESID_EMPREG_SECT_TERC	0,9977	2
N_INDIVIDUOS_RESIDENT	N_INDIVIDUOS_RESIDENT_M	0,9959	1
N_INDIVIDUOS_PRESENT	N_INDIVIDUOS_PRESENT_M	0,9958	1
N_INDIVIDUOS_RESIDENT_25A64	N_INDIVIDUOS_RESIDENT_M_25A64	0,9951	1
N_INDIVIDUOS_RESIDENT_65	N_INDIVIDUOS_RESIDENT_M_65	0,9941	2
N_INDIVIDUOS_RESIDENT_14A19	N_INDIVIDUOS_RESIDENT_15A19	0,9939	2
N_INDIVIDUOS_RESIDENT_65	N_IND_RESID_PENS_REFORM	0,9925	1
N_INDIVIDUOS_PRESENT_M	N_INDIVIDUOS_RESIDENT_M	0,9920	1
N_INDIVIDUOS_RESIDENT_M_14A19	N_INDIVIDUOS_RESIDENT_M_15A19	0,9907	2
N_INDIVIDUOS_PRESENT_H	N_INDIVIDUOS_RESIDENT_H	0,9891	1
N_INDIVIDUOS_RESIDENT_H_14A19	N_INDIVIDUOS_RESIDENT_H_15A19	0,9887	1
N_INDIVIDUOS_RESIDENT_H	N_INDIVIDUOS_RESIDENT_H_25A64	0,9879	2
N_INDIVIDUOS_RESIDENT_M_25A64	N_IND_RESID_EMPREGADOS	0,9856	2
N_INDIVIDUOS_RESIDENT_H	N_INDIVIDUOS_RESIDENT_M_25A64	0,9842	1

N_INDIVIDUOS_RESIDENT_M	N_INDIVIDUOS_RESIDENT_M_25A64	0,9832	2
N_INDIVIDUOS_RESIDENT_20A24	N_INDIVIDUOS_RESIDENT_H_20A24	0,9744	2
N_INDIVIDUOS_RESIDENT_20A24	N_INDIVIDUOS_RESIDENT_M_20A24	0,9735	1
N_INDIVIDUOS_RESIDENT_M	N_IND_RESID_SEM_ACT_ECON	0,9730	1
N_INDIVIDUOS_RESIDENT_H_65	N_IND_RESID_PENS_REFORM	0,9721	1
N_INDIVIDUOS_RESIDENT_10A13	N_IND_RESIDENT_FENSINO_2BAS	0,9713	1
N_EDIFICIOS_1OU2_PISOS	N_EDIFICIOS_CLASSICOS_1OU2	0,9711	1
N_INDIVIDUOS_RESIDENT_0A4	N_INDIVIDUOS_RESIDENT_H_0A4	0,9682	2
N_IND_RESID_PENS_REFORM	N_IND_RESID_SEM_ACT_ECON	0,9681	2
N_INDIVIDUOS_RESIDENT_0A4	N_INDIVIDUOS_RESIDENT_M_0A4	0,9670	1
N_INDIVIDUOS_RESIDENT_5A9	N_INDIVIDUOS_RESIDENT_H_5A9	0,9614	1
N_INDIVIDUOS_RESIDENT_14A19	N_IND_RESID_ESTUD_MUN_RESID	0,9609	1
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESIDENT_ENSINCOMP_2BAS	0,9568	2
N_EDIFICIOS_CLASSICOS_1OU2	N_EDIFICIOS_CLASSICOS_EMBANDA	0,9464	1
N_INDIVIDUOS_RESIDENT_H_10A13	N_IND_RESIDENT_FENSINO_2BAS	0,9406	1
N_INDIVIDUOS_RESIDENT_H_5A9	N_IND_RESIDENT_FENSINO_1BAS	0,9390	2
N_IND_RESIDENT_FENSINO_SEC	N_IND_RESID_ESTUD_MUN_RESID	0,9363	2
N_INDIVIDUOS_RESIDENT_M_10A13	N_IND_RESIDENT_FENSINO_2BAS	0,9290	2
N_IND_RESIDENT_ENSINCOMP_SUP	N_IND_RESIDENT_FENSINO_SUP	0,9285	1
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESIDENT_ENSINCOMP_3BAS	0,9177	2
N_IND_RESIDENT_ENSINCOMP_SEC	N_IND_RESIDENT_FENSINO_SUP	0,9080	2
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESID_DESEMP_PROC_EMPRG	0,9078	2
N_INDIVIDUOS_RESIDENT_M_14A19	N_IND_RESIDENT_FENSINO_SEC	0,9038	2
N_INDIVIDUOS_RESIDENT_H_15A19	N_IND_RESIDENT_FENSINO_3BAS	0,9001	2
N_IND_RESIDENT_ENSINCOMP_SEC	N_IND_RESID_PENS_REFORM	0,8961	1
N_INDIVIDUOS_RESIDENT_H_15A19	N_INDIVIDUOS_RESIDENT_M_20A24	0,8827	1
N_INDIV_RESIDENT_N_LER_ESCRV	N_IND_RESIDENT_ENSINCOMP_1BAS	0,8742	1
N_INDIVIDUOS_RESIDENT_M_10A13	N_INDIVIDUOS_RESIDENT_M_20A24	0,8690	1
N_IND_RESID_TRAB_MUN_RESID	N_IND_RESID_PENS_REFORM	0,8639	1
N_ALOJAMENTOS_VAGOS	N_EDIFICIOS_CLASSICOS_3OUMAIS	0,8602	2
N_INDIVIDUOS_RESIDENT_M_0A4	N_INDIVIDUOS_RESIDENT_M_5A9	0,8477	2
N_INDIVIDUOS_RESIDENT_H_5A9	N_INDIVIDUOS_RESIDENT_M_0A4	0,8419	2
N_INDIVIDUOS_RESIDENT_M_20A24	N_IND_RESID_EMPREG_SECT_SEQ	0,8382	1
N_ALOJAMENTOS_VAGOS	N_EDIFICIOS_CLASSICOS	0,8313	2
N_INDIVIDUOS_RESIDENT_M_14A19	N_IND_RESID_EMPREG_SECT_SEQ	0,8164	1
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESID_EMPREG_SECT_SEQ	0,8138	1

Table 3 - Feature selection process through correlation

This process led to the elimination of 53 variables, therefore remaining 39 independent features. The correlation matrix of those features is visible in Figure 11, below. The dark red diagonal represents the correlation of 1 that each feature has with itself, there being no other relevant correlations – the highest is 0,79 and the lowest is -0,62.

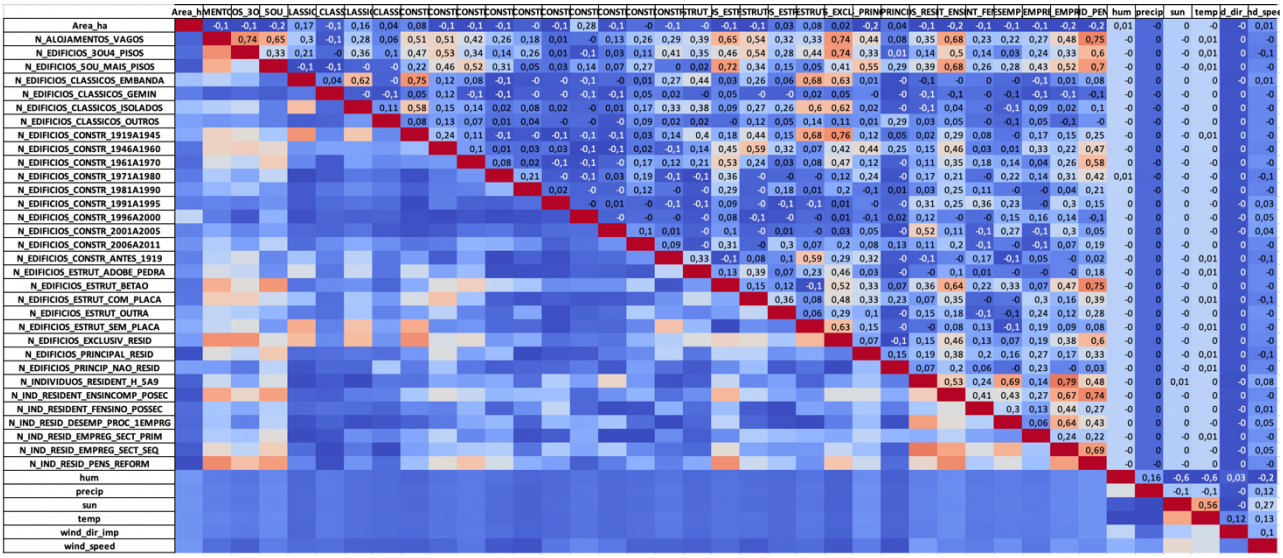

Figure 11 - Correlation matrix after feature selection

As a result of this process, all of the meteorological features were kept (humidity, precipitation, sun, temperature and direction and speed of the wind) as well as 33 from the census, of which 25 related to buildings and their characteristics (unoccupied houses; in terms of floors per classical building: buildings with 3 or 4 floors, buildings with 5 or more floors; in terms of closeness of the buildings in the block: isolated, in pairs, with 3 or more in a row, other; in terms of year of construction: prior to 1919, 1919 to 1945, 1946 to 1960, 1961 to 1970, 1971 to 1980, 1981 to 1990, 1991 to 1995, 1995 to 2000, 2001 to 2005, 2006 to 2011; in terms of construction materials: adobe and stone, concrete and others; whether the building has reinforced steel in the floors; in terms of occupation: only residential, mostly residential and finally mostly non-residential) and 7 to the habitants (male residents aged from 5 to 9, in terms of education: residents currently on a pos high school course and residents who have completed a pos high school course; regarding employment status: unemployed looking for the first job, employed in the primary sector, employed in the secondary sector and finally retired), plus the size of the subsection.

5. MODELING

The present results refer to a random sample of 100 subsections (2,7% of the dataset) of the city of Lisbon, between 2013 and 2018. The 5 224 813 observations were split into a training and testing set, following a 70-30 ratio.

5.1. RANDOM SAMPLING

The logistic regression was trained using a cross-validation technique, with 4 folds, considering different parameters for the elastic net proportions (0 would be pure Ridge penalty and 1 pure Lasso penalty) and for the weight of the regularization. Using the area under the ROC curve as an evaluation metric, a logistic regression with Ridge penalty and a small regularization (0,001) was found to be the best model – performance metrics in Table 4 – although it predicted all the observations as 0, which means for no event to be occurring.

The decision tree was first trained without any hyperparameter tuning. As the algorithm only created one node, it wasn't necessary to do any pruning. As expected in such imbalanced situations, all observations were predicted as 0, performing as the logistic regression (Table 4).

A random forest was trained using cross-validation, which yielded 5 decision trees, each with a maximum depth of 15 and, at most, 30 bins as the best performant random forest. This classifier wasn't able to predict any positive events (Table 4).

	Area under the ROC	**Recall**	**F_1-score**	**Accuracy**
Logistic Regression	0.5	0.5	0.5	0.9996
Decision Tree	0.5	0.5	0.5	0.9996
Random Forest	0.5	0.5	0.5	0.9996

Table 4 - Performance metrics with random sampling

The metrics recall and F_1-score were first calculated for each label and then the unweighted average was computed, therefore not accounting for class imbalance. If accounting only for the positive label, it could result in a model always predicting to be an occurrence, which is not a desirable outcome. On the other hand, the fact that the classes are so imbalanced means that a weighted average doesn't reflect the metrics of the relevant class, which is the under-represented one.

Being that under these three algorithms the prediction was always not to be an occurrence, it was not relevant to plot the ROC curve nor show the confusion matrix.

5.2. OVER-SAMPLING

Facing a sample with only 1 344 positive observations (target=1) for 3 656 447 negative ones (target=0) in the train set, there is an extreme imbalance in the dataset – over 1:2 720. Under such circumstances it is common for models to predict only the majority class, still resulting in high accuracy. A study by Batista, G. E., Prati, R. C., & Monard, M. C. (2004) found that over-sampling methods performed better than under-sampling when using the AUC as a metric and even that random over-sampling was very competitive when compared to more complex techniques, so that was the strategy applied first.

To allow this analysis, the sample was split into three sets: training, validation and test set, being that only the first was over-sampled with random over-sampling and then using SMOTE.

5.2.1. Random over-sampling with replacement

The logistic regression was trained for the best hyperparameters – found using 5-fold cross validation for each penalty (Ridge, Elastic Net, and Lasso) and for the weight of the regularization – with different ratios of oversampling. It is interesting to note that regardless of the size of the over-sampling, the best model was consistently a Ridge logistic regression with a penalty of 0,001. A table of the AUC results on the validation set according to the level of over-sampling and respective graph can be seen below (Table 5 and Figure 12).

Ratio	AUC
0.1	0,52053
0.2	0,52548
0.3	0,54102
0.4	0,56093
0.5	0,58212
0.6	0,60007
0.7	0,61002
0.8	0,62740
0.9	**0,63396**
1	0,62692

Table 5 - AUC of logistic regression for each ratio of random over-sampling

Figure 12 - AUC of logistic regression for different ratios of random over-sampling

The main performance metrics from the test set of the logistic regression with the optimal hyperparameters and a ratio of imbalance of 0.9 can be found in Table 11 – the improvement in the logistic regression by using random over-sampling allowed to correctly predict 308 relevant events out of 574, although at the cost of wrongly predicting 431 788 events as relevant out of 1 566 448 negative events (confusion matrix in Table 6). The ROC curve was plotted in Figure 13, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 134 660	431 788
True positive	266	308

Table 6 - Confusion matrix of the logistic regression with random over-sampling

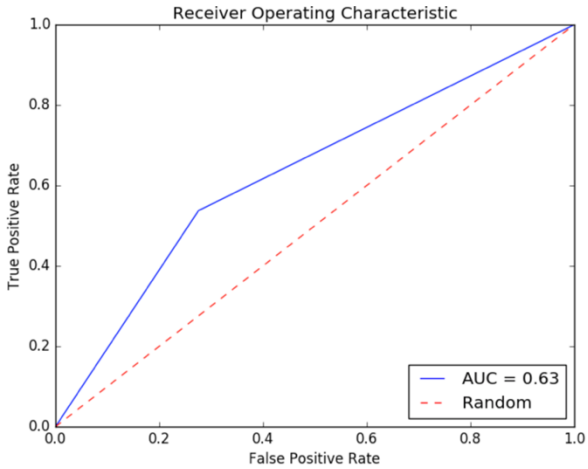

Figure 13 - ROC curve for logistic regression with random over-sampling

The same strategy was used to train the decision tree, which also performed consistently better with the identical hyperparameter for all ratios of imbalance: a depth of 30. A table of the AUC results on the validation set according to the level of over-sampling and respective graph can be seen below (Table 7 and Figure 14).

Ratio	AUC
0.1	0,52907
0.2	0,53025
0.3	0,53332
0.4	0,53206
0.5	0,52968
0.6	0,53066
0.7	0,53400
0.8	**0,53803**
0.9	0,53754
1	0,53728

Table 7 - AUC of decision tree for each ratio of random over-sampling

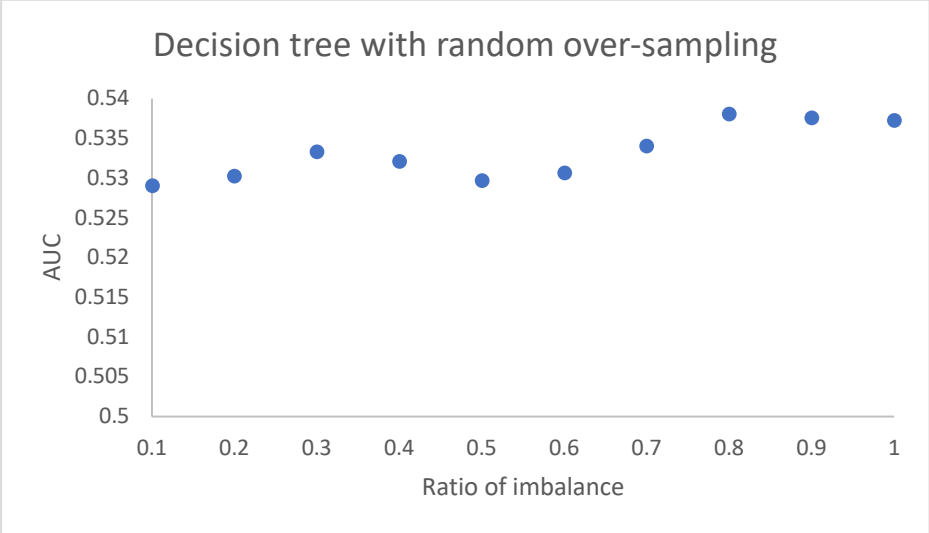

Figure 14 - AUC of decision tree for different ratios of random over-sampling

The main performance metrics from the test set of the decision tree with the optimal hyperparameters and a ratio of imbalance of 0.8 can be found in table 11 – the improvement in the decision tree by using random over-sampling allowed to correctly predict 50 relevant events out of 574, although at the cost of wrongly predicting 8 548 events as relevant out of 1 566 448 negative events (confusion matrix in Table 8). The ROC curve was plotted in Figure 15, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 557 900	8 548
True positive	524	50

Table 8 - Confusion matrix of the decision tree with random over-sampling

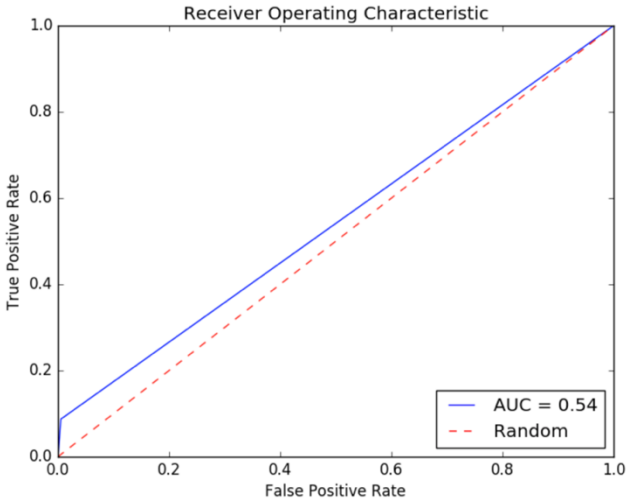

Figure 15 - ROC curve for decision tree with random over-sampling

Contrarily to the models above, there was some variability regarding the best hyperparameters for the random forest, with either 5 or 30 being the depth for the constant 200 estimators with 6 features. A table of the AUC results on the validation set according to the level of over-sampling and respective graph can be seen below (Table 9 and Figure 16), being that the first two had a depth of 5 while the remaining had of 30.

Ratio	AUC
0.1	0,51311
0.2	0,51311
0.3	0,53196
0.4	0,55311
0.5	0,56702
0.6	0,59633
0.7	0,62809
0.8	0,65108
0.9	0,64735
1	**0,65286**

Table 9 - AUC of random forest for each ratio of random over-sampling

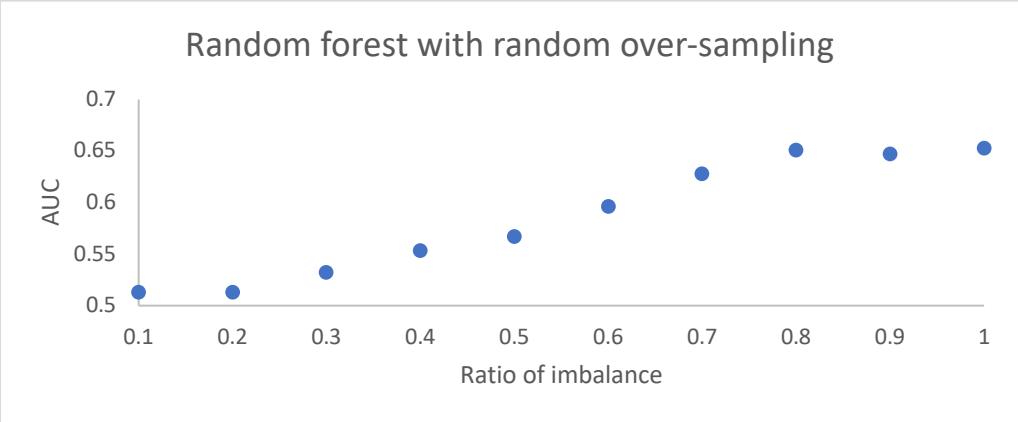

Figure 16 - AUC of random forest for different ratios of random over-sampling

The main performance metrics from the test set of the random forest with the optimal hyperparameters (200 estimators with a depth of 5) and a ratio of imbalance of 1 can be found in Table 11 – the improvement in the random forest by using random over-sampling allowed to correctly predict 404 relevant events out of 574, although at the cost of wrongly predicting 551 873 events as relevant out of 1 566 448 negative events (confusion matrix in Table 10). The ROC curve was plotted in Figure 17, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 014 575	551 873
True positive	170	404

Table 10 - Confusion matrix of the random forest with random over-sampling

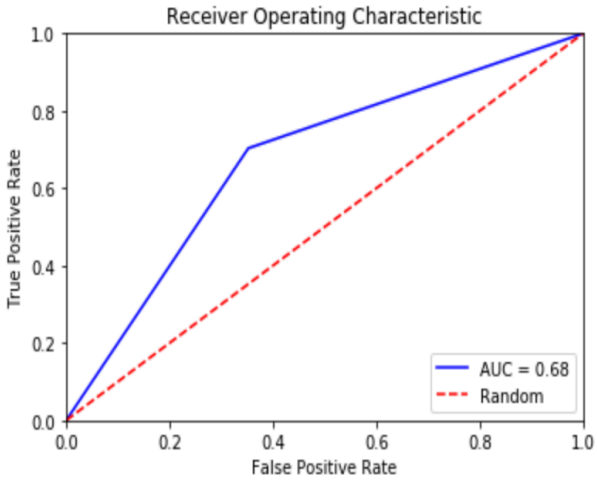

Figure 17 - ROC curve for random forest with random over-sampling

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	0.9	0.6309	0.6309	0.4208	0.7243
Decision Tree	0.8	0.5409	0.5409	0.5045	0.9942
Random Forest	1	0.6755	0.6755	0.3939	0.6477

Table 11 - Performance metrics with random over-sampling on the test set

When using random over-sampling, the best performant model according to the AUC metric was the random forest. As the increase in performance comes at the cost of wrongly predicting many events as relevant, it translates in a big decrease in the accuracy and being the lowest F_1-score.

Additionally, the models were also used on a new sample with 100 new subsections to test for the adaptability to the entire city – the main performance metrics can be found in Table 12.

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	0.9	0.5926	0.5926	0.4418	0.7904
Decision Tree	0.8	0.5122	0.5122	0.4988	0.9922
Random Forest	1	0.6587	0.6587	0.4558	0.8358

Table 12 - Performance metrics with random over-sampling on a new data set

The confusion matrix of each model on the new sample can be seen in Tables 13, 14, and 15.

	Predict negative	Predict positive
True negative	4 129 416	1 094 388
True positive	606	395

Table 13 - Confusion matrix of the logistic regression with random over-sampling on the new sample

	Predict negative	Predict positive
True negative	5 183 772	40 032
True positive	969	32

Table 14 - Confusion matrix of the decision tree with random over-sampling on the new sample

	Predict negative	Predict positive
True negative	4 366 474	857 330
True positive	519	482

Table 15 - Confusion matrix of the random forest with random over-sampling on the new sample

When applying random over-sampling, the random forest was the best performant model. It is interesting to note that this algorithm is also the one that best adjusted to the new sample of 100 different subsections, losing less AUC and even over-performing the logistic regression in terms of F_1-score and accuracy.

5.2.2. SMOTE

The logistic regression was trained for the best hyperparameters – found using 5-fold cross validation for each penalty (Ridge, Elastic Net and Lasso) and the weight of the regularization – with different ratios of oversampling. Once again, it is interesting to note that regardless of the size of the oversampling, the best model was consistently a Ridge logistic regression with a penalty of 0,001. A table of the AUC results on the validation set according to the level of over-sampling and respective graph can be seen below (Table 16 and Figure 18).

Ratio	AUC
0.1	0,517130285
0.2	0,526094327
0.3	0,545908843
0.4	0,558976585
0.5	0,576686715
0.6	0,594633772
0.7	0,61045669
0.8	0,624509271
0.9	**0,626362398**
1	0,625116261

Table 16 - AUC of logistic regression for each ratio of SMOTE

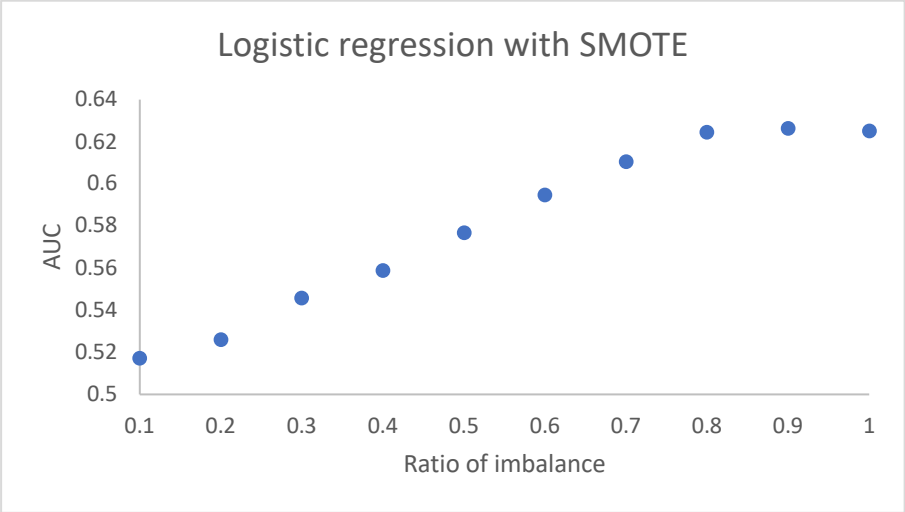

Figure 18 - AUC of logistic regression for different ratios of SMOTE

The main performance metrics from the test set of the logistic regression with the optimal hyperparameters and a ratio of imbalance of 0.9 can be found in Table 22 – the improvement in the logistic regression by using SMOTE allowed to correctly predict 338 relevant events out of 574, although at the cost of wrongly predicting 475 469 events as relevant out of 1 566 448 negative events (confusion matrix in Table 17). The ROC curve was plotted in Figure 19, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 090 979	475 469
True positive	236	338

Table 17 - Confusion matrix of the logistic regression with SMOTE

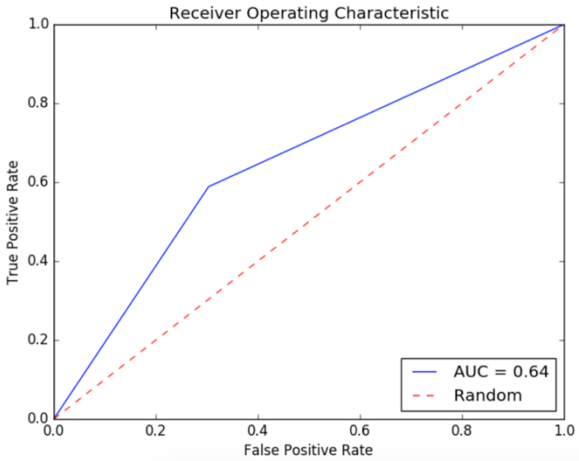

Figure 19 - ROC curve for logistic regression with SMOTE

The same strategy was used to train the decision tree, which also performed consistently better with the identical hyperparameter for all ratios of imbalance: a depth of 30. A table of the AUC results on the validation set according to the level of over-sampling and respective graph can be seen below (Table 18 and Figure 20).

Ratio	AUC
0.1	0,54191331
0.2	0,54974172
0.3	0,54432357
0.4	0,54511817
0.5	0,54532176
0.6	**0,55684134**
0.7	0,54826982
0.8	0,54948302
0.9	0,54240057
1	0,55362546

Table 18 - AUC of decision tree for each ratio of SMOTE

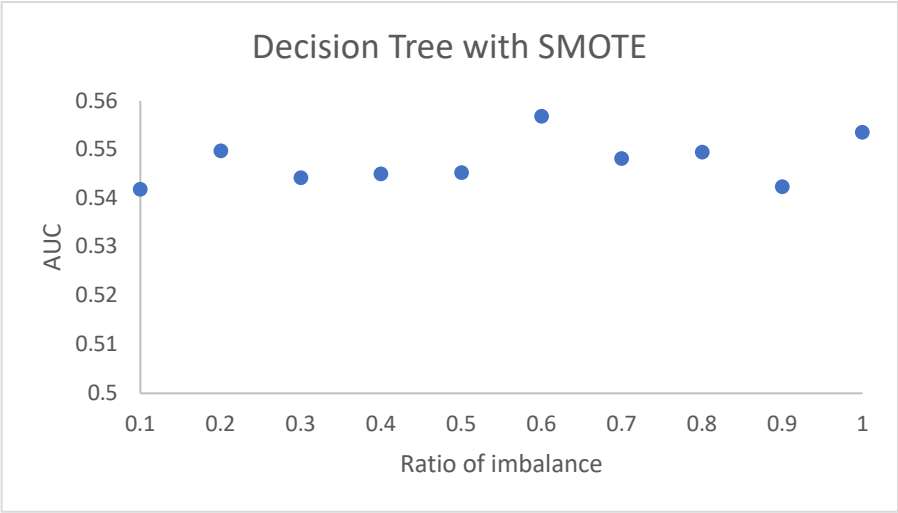

Figure 20 - AUC of decision tree for different ratios of SMOTE

The main performance metrics from the test set of the decision tree with the optimal hyperparameters and a ratio of imbalance of 0.6 can be found in Table 22 – the improvement in the decision tree by using SMOTE allowed to correctly predict 47 relevant events out of 574, although at the cost of wrongly predicting 34 304 events as relevant out of 1 566 448 negative events (confusion matrix in Table 19). The ROC curve was plotted in Figure 21, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 532 144	34 304
True positive	527	47

Table 19 - Confusion matrix of the decision tree with SMOTE

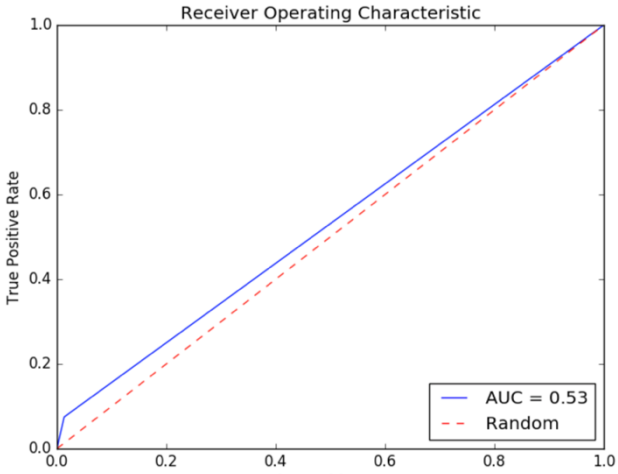

Figure 21 - ROC curve for decision tree with SMOTE

The best hyperparameters for the random forest with SMOTE were always 200 trees with 6 features. However, the best depth for a ratio of 0.1 was 30, for a ratio of 0.2 was 15 and for all the remaining was 5. A table of the AUC results on the validation set according to the level of over-sampling and respective graph can be seen below (Table 20 and Figure 22).

Ratio	AUC
0.1	0,51430
0.2	0,51907
0.3	0,52803
0.4	0,55185
0.5	0,56851
0.6	0,60060
0.7	0,62264
0.8	0,64105
0.9	0,65291
1	**0,65318**

Table 20 - AUC of random forest for each ratio of SMOTE

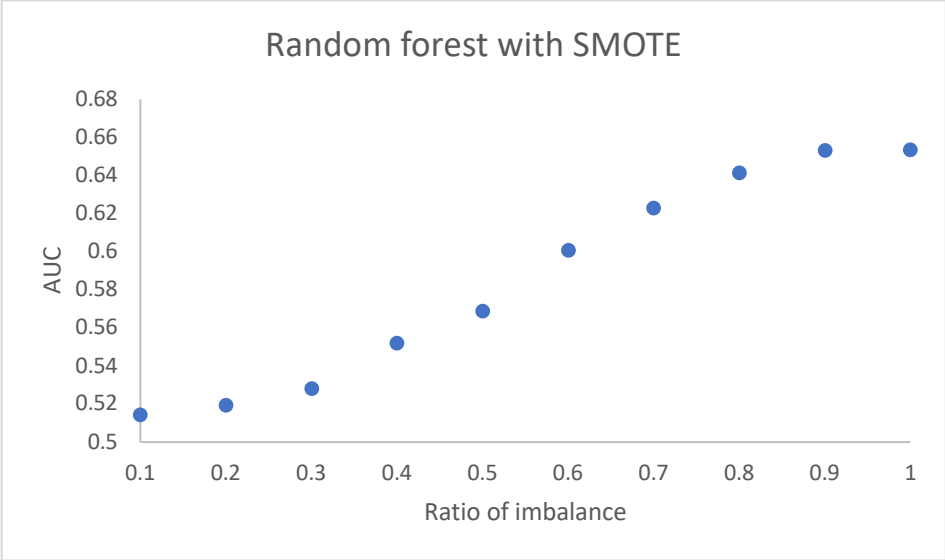

Figure 22 - AUC of random forest for different ratios of SMOTE

The main performance metrics from the test set of the random forest with the optimal hyperparameters and a ratio of imbalance of 1 can be found in table 22 – the improvement in the random forest by using SMOTE allowed to correctly predict 355 relevant events out of 574, although at the cost of wrongly predicting 482 412 events as relevant out of 1 566 448 negative events (confusion matrix in Table 21). The ROC curve was plotted in Figure 23, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 084 036	482 412
True positive	219	355

Table 21 - Confusion matrix of the random forest with SMOTE

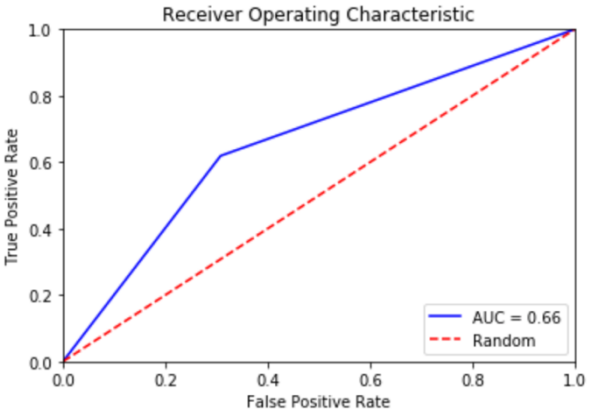

Figure 23 - ROC curve for random forest with SMOTE

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	0.9	0.6427	0.6427	0.4112	0.6964
Decision Tree	0.6	0.5300	0.5300	0.4957	0.9778
Random Forest	1	0.6556	0.6556	0.4098	0.6920

Table 22 - Performance metrics with SMOTE on the test set

When using SMOTE, the best performant model according to the AUC metric was the random forest. As the increase in performance comes at the cost of wrongly predicting many events as relevant, it translates in a big decrease in the accuracy and being the lowest F_1-score.

Additionally, the models were also used on a new sample with 100 new subsections to test for the adaptability to the entire city – the main performance metrics can be found in Table 23.

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	0.9	0.5899	0.5899	0.4354	0.7702
Decision Tree	0.6	0.4453	0.4453	0.3236	0.4781
Random Forest	1	0.6554	0.6554	0.4520	0.8232

Table 23 - Performance metrics with SMOTE on a new data set

The confusion matrix of each model on the new sample can be seen in Tables 24, 25 and 26.

	Predict negative	Predict positive
True negative	4 023 916	1 199 888
True positive	591	410

Table 24 - Confusion matrix of the logistic regression with SMOTE on the new sample

	Predict negative	Predict positive
True negative	2 497 359	2 726 445
True positive	588	413

Table 25 - Confusion matrix of the decision tree with SMOTE on the new sample

	Predict negative	Predict positive
True negative	4 300 321	923 483
True positive	513	488

Table 26 - Confusion matrix of the random forest with SMOTE on the new sample

It is interesting to note that the logistic regression performed considerably worse on a new dataset than the random forest, which had similar results both on the test set and on the new sample, with 100 different subsections.

Overall, the best model with over-sampling was the random forest with random over-sampling at an AUC of 0,6755 on the test set. This conclusion goes in line with Batista, G. E., Prati, R. C., & Monard, M. C. (2004) findings regarding the high performance of random over-sampling.

5.3. UNDER-SAMPLING

Alternatively to creating new observations of the minority class, models were trained using a smaller subset of the observations of the majority class. Even though under-sampling is less computational heavy than over-sampling as it requires using less data, it was only possible to use the technique of random under-sampling and the version 1 of Near Miss, as the other techniques require more computational capacity than available.

5.3.1. Random under-sampling

The logistic regression was trained for the best hyperparameters – found using 5-fold cross validation for each penalty (Ridge, Elastic Net, and Lasso) and for the weight of the regularization – with different ratios of oversampling. Contrary to what happened when using random over-sampling, there was some variability in the best penalty, being that Ridge was often the best hyperparameter, and with varying weights of regularization. A table of the AUC results on the validation set according to the level of under-sampling and respective graph can be seen below (Table 27 and Figure 24).

Ratio	AUC
0.1	0,50955
0.2	0,53598
0.3	0,55298
0.4	0,56519
0.5	0,58340
0.6	0,60705
0.7	0,62753
0.8	0,63409
0.9	**0,64326**
1	0,63873

Table 27 - AUC of logistic regression for each ratio of random under-sampling

Figure 24 - AUC of logistic regression for different ratios of random under-sampling

The main performance metrics from the test set of the logistic regression with the optimal hyperparameters (Lasso penalty with a regularization of 10) and a ratio of imbalance of 0.9 can be found in Table 33 – the improvement in the logistic regression by using random under-sampling allowed to correctly predict 330 relevant events out of 574, although at the cost of wrongly predicting 426 016 events as relevant out of 1 566 448 negative events (confusion matrix in Table 28). The ROC curve was plotted in Figure 25, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 140 432	426 016
True positive	244	330

Table 28 - Confusion matrix of the logistic regression with random under-sampling

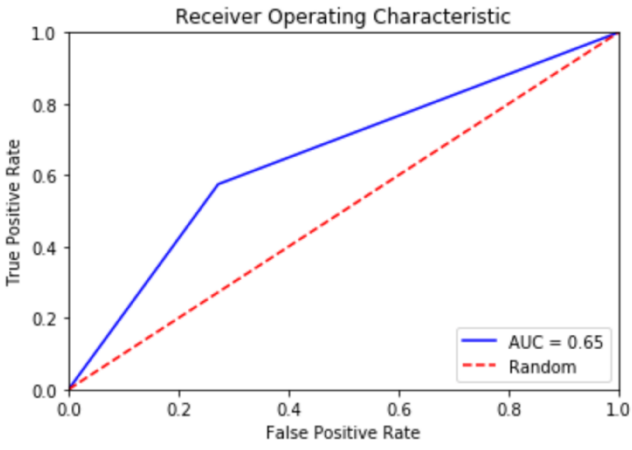

Figure 25 - ROC curve for logistic regression with random under-sampling

The same strategy was used to train the decision tree, which performed consistently better with the identical hyperparameter for all ratios of imbalance: a depth of 5. A table of the AUC results on the validation set according to the level of under-sampling and respective graph can be seen below (Table 29 and Figure 26).

Ratio	AUC
0.1	0,52420
0.2	0,53080
0.3	0,55132
0.4	0,57281
0.5	0,58953
0.6	0,60246
0.7	0,60785
0.8	0,63761
0.9	0,63366
1	**0,64895**

Table 29 - AUC of decision tree for each ratio of random under-sampling

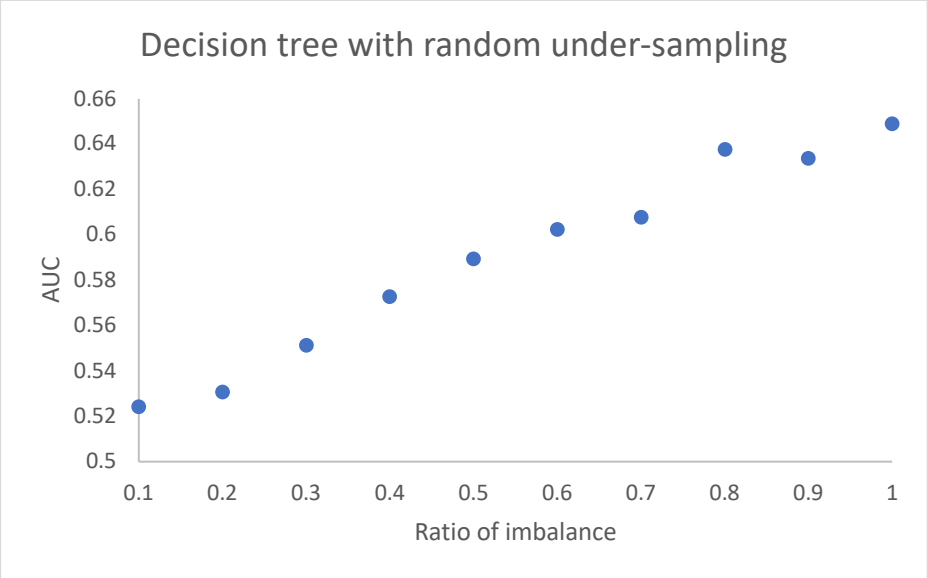

Figure 26 - AUC of decision tree for different ratios of random under-sampling

The main performance metrics from the test set of the decision tree with the optimal hyperparameters and a ratio of imbalance of 1 can be found in table 33 – the improvement in the decision tree by using random under-sampling allowed to correctly predict 357 relevant events out of 574, although at the cost of wrongly predicting 471 303 events as relevant out of 1 566 448 negative events (confusion matrix in Table 30). The ROC curve was plotted in Figure 27, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 095 145	471 303
True positive	217	357

Table 30 - Confusion matrix of the decision tree with random under-sampling

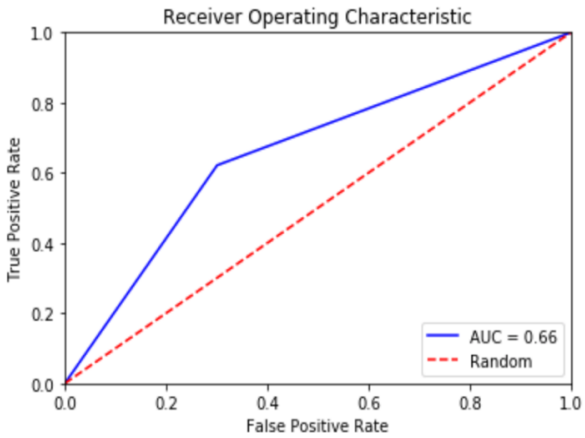

Figure 27 - ROC curve for decision tree with random under-sampling

The random forest model was the one with more variability in terms of hyperparameters: as the ratio of imbalance grew, the optimal depth of the trees decreased and the number of estimators varied between 50 and 300, with either 5 or 6 features to split. A table of the AUC results on the validation set according to the level of under-sampling and respective graph can be seen below (Table 31 and Figure 28).

Ratio	AUC
0.1	0,52725
0.2	0,55697
0.3	0,58355
0.4	0,58537
0.5	0,61510
0.6	0,63692
0.7	0,63650
0.8	0,64100
0.9	**0,66203**
1	0,65859

Table 31 - AUC of random forest for each ratio of random under-sampling

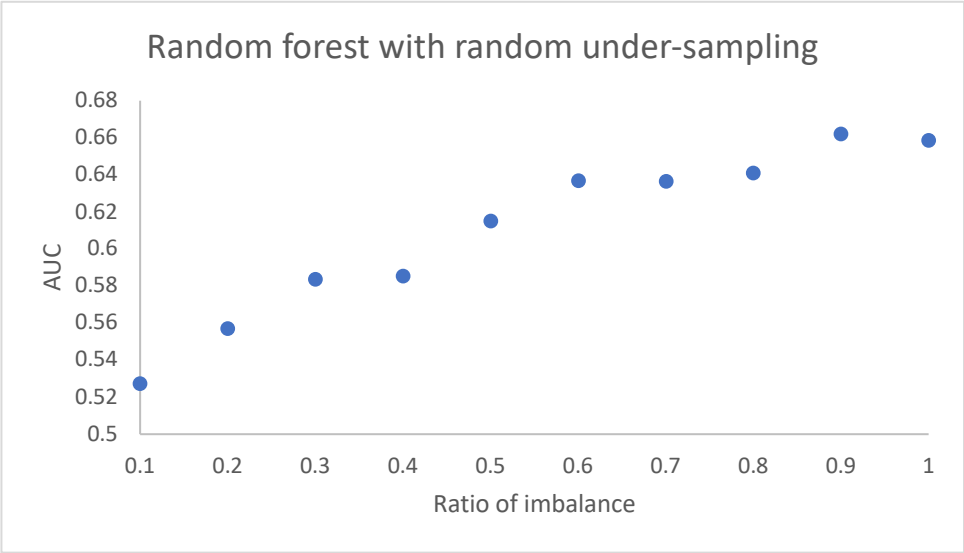

Figure 28 - AUC of random forest for different ratios of random under-sampling

The main performance metrics from the test set of the random forest with the optimal hyperparameters (100 estimators with a depth of 5) and a ratio of imbalance of 0.9 can be found in Table 33 – the improvement in the random forest by using random under-sampling allowed to correctly predict 373 relevant events out of 574, although at the cost of wrongly predicting 464 760 events as relevant out of 1 566 448 negative events (confusion matrix in Table 32). The ROC curve was plotted in Figure 29, showing the improvement from random.

	Predict negative	Predict positive
True negative	1 101 688	464 760
True positive	201	373

Table 32 - Confusion matrix of the random forest with random under-sampling

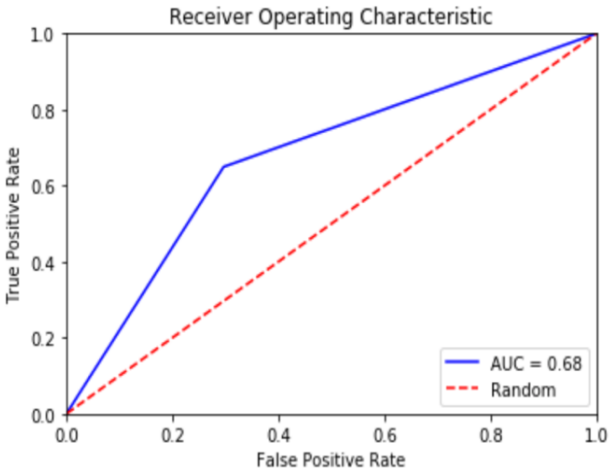

Figure 29 - ROC curve for random forest with random under-sampling

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	0.9	0.6514	0.6514	0.4221	0.7280
Decision Tree	1	0.6603	0.6603	0.4123	0.6603
Random Forest	0.9	0.6764	0.6764	0.4138	0.7033

Table 33 - Performance metrics with random under-sampling on the test set

When using random under-sampling, the best performant model according to the AUC metric was, once again, the random forest. The usage of this simple technique leads to great improvements from the random level and had better results overall than the over-sampling ones.

Additionally, the models were also used on a new sample with 100 new subsections to test for the adaptability to the entire city – the main performance metrics can be found in Table 34.

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	0.9	0.5777	0.5777	0.4208	0.7258
Decision Tree	1	0.6488	0.6488	0.4571	0.8340
Random Forest	0.9	0.6538	0.6538	0.4638	0.8630

Table 34 - Performance metrics with random under-sampling on a new data set

The confusion matrix of each model on the new sample can be seen in Tables 35, 36 and 37.

	Predict negative	Predict positive
True negative	3 791 587	1 432 217
True positive	571	430

Table 35 - Confusion matrix of the logistic regression with random under-sampling on the new sample

	Predict negative	Predict positive
True negative	4 388 244	835 560
True positive	543	458

Table 36 - Confusion matrix of the decision tree with random under-sampling on the new sample

	Predict negative	Predict positive
True negative	4 508 480	715 324
True positive	556	445

Table 37 - Confusion matrix of the random forest with random under-sampling on the new sample

When applying random under-sampling, the random forest was the best performant model, adjusting well to the 100 subsections of new data. The logistic regression, on the other hand, was the algorithm that performed worst, with a decrease of almost 8% comparing to around 2% in the decision tree and random forest.

5.3.2. Near Miss

The logistic regression was trained for the best hyperparameters – found using 5-fold cross validation for each penalty (Ridge, Elastic Net, and Lasso) and the weight of the regularization – with different ratios of under-sampling. Once again, it is interesting to note that regardless of the size of the under-sampling, the best model was consistently a Ridge logistic regression with a penalty of 1000. A table of the AUC results on the validation set according to the level of under-sampling and respective graph can be seen below (Table 38 and Figure 30).

Ratio	AUC
0.1	0,45483
0.2	0,46599
0.3	0,47451
0.4	0,47794
0.5	0,47876
0.6	0,48108
0.7	0,48696
0.8	0,48928
0.9	0,49094
1	**0,49225**

Table 38 - AUC of logistic regression for each ratio of Near Miss

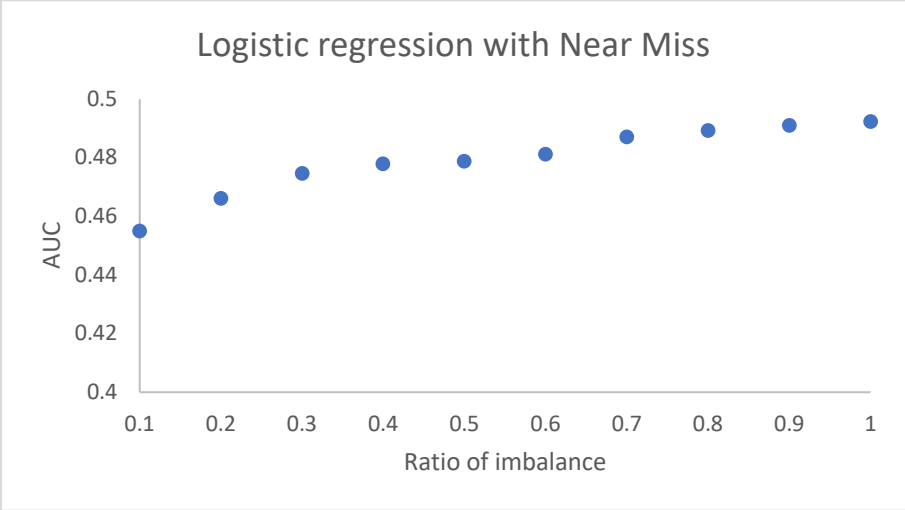

Figure 30 - AUC of logistic regression for different ratios of Near Miss

The main performance metrics from the test set of the logistic regression with the optimal hyperparameters and a ratio of imbalance of 1 can be found in Table 44 – the improvement in the logistic regression by using random under-sampling allowed to correctly predict 534 relevant events out of 574, although at the cost of wrongly predicting 1 472 848 events as relevant out of 1 566 448 negative events (confusion matrix in Table 39). The ROC curve was plotted in Figure 31, showing the improvement from random.

	Predict negative	Predict positive
True negative	93 600	1 472 848
True positive	40	534

Table 39 - Confusion matrix of the logistic regression with Near Miss

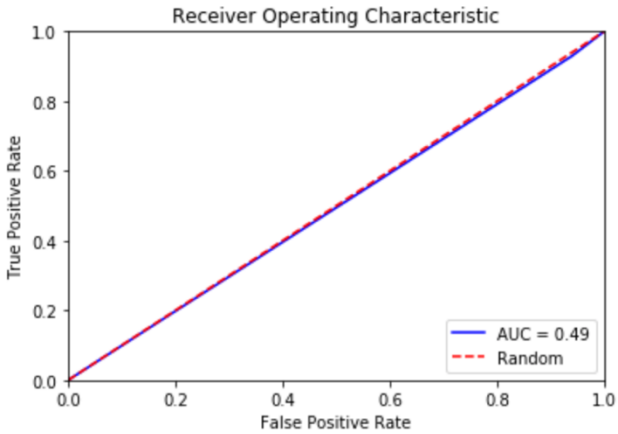

Figure 31 - ROC curve for logistic regression with Near Miss

The same strategy was used to train the decision tree, which also performed consistently better with a depth of 10 for all ratios of imbalance, except 0.1 where 20 was better. A table of the AUC results on the validation set according to the level of under-sampling and respective graph can be seen below (Table 40 and figure 32).

Ratio	AUC
0.1	0,48338
0.2	0,47254
0.3	0,47695
0.4	0,48552
0.5	0,49153
0.6	0,49484
0.7	0,49851
0.8	0,49715
0.9	0,49845
1	**0,50184**

Table 40 - AUC of decision tree for each ratio of Near Miss

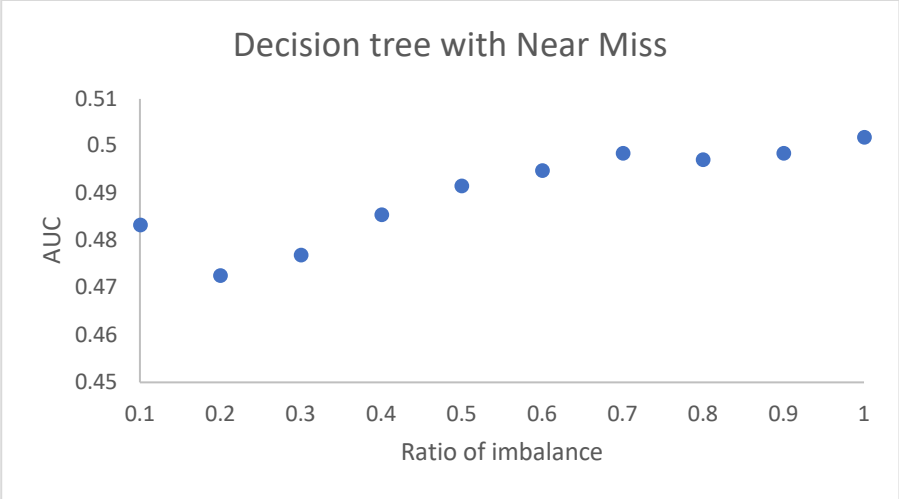

Figure 32 - AUC of decision tree for different ratios of Near Miss

The main performance metrics from the test set of the decision tree with the optimal hyperparameters and a ratio of imbalance of 1 can be found in Table 44 – the improvement in the decision tree by using Near Miss allowed to correctly predict 559 relevant events out of 574, although at the cost of wrongly predicting 1 518 409 events as relevant out of 1 566 448 negative events (confusion matrix in Table 41). The ROC curve was plotted in Figure 33, showing the improvement from random.

	Predict negative	Predict positive
True negative	48 039	1 518 409
True positive	15	559

Table 41 - Confusion matrix of the decision tree with Near Miss

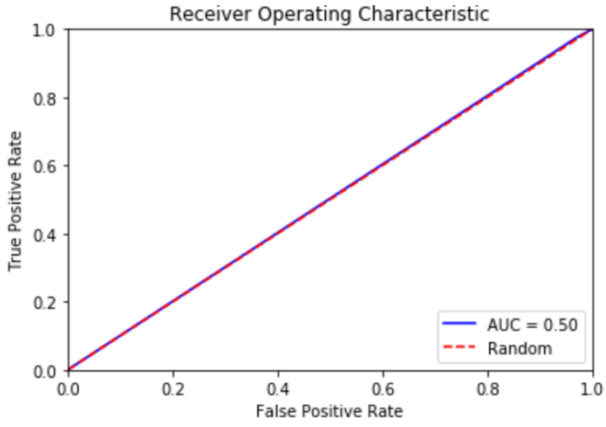

Figure 33 - ROC curve for decision tree with Near Miss

The best hyperparameters for the random forest with Near Miss were a depth of either 15 or 30 with 5 or 6 features considered at each split decision. It was on the number of estimators that more volatility was found, varying between 50 and 200 trees. A table of the AUC results on the validation set according to the level of under-sampling and respective graph can be seen below (Table 42 and Figure 34).

Ratio	AUC
0.1	0,49260
0.2	0,48245
0.3	0,48838
0.4	0,49954
0.5	0,49507
0.6	0,49704
0.7	0,49972
0.8	0,50025
0.9	0,50228
1	**0,50282**

Table 42 - AUC of random forest for each ratio of Near Miss

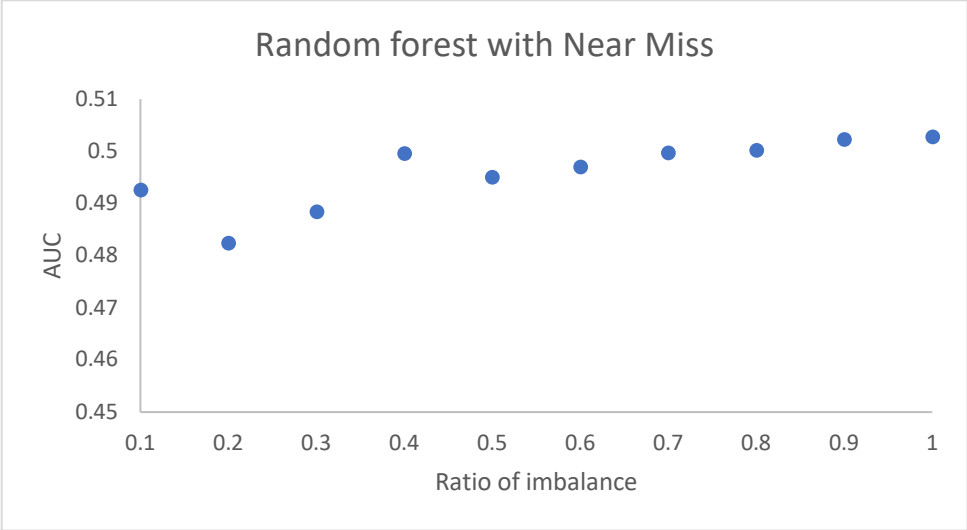

Figure 34 - AUC of random forest for different ratios of Near Miss

The main performance metrics from the test set of the random forest with the optimal hyperparameters and a ratio of imbalance of 1 (50 trees with a maximum depth of 15 and 5 features) can be found in Table 44 – the improvement in the random forest by using Near Miss allowed to correctly predict 561 relevant events out of 574, although at the cost of wrongly predicting 1 529 583 events as relevant out of 1 566 448 negative events (confusion matrix in Table 43). The ROC curve was plotted in Figure 35, showing the improvement from random.

	Predict negative	Predict positive
True negative	36 865	1 529 583
True positive	13	561

Table 43 - Confusion matrix of the random forest with Near Miss

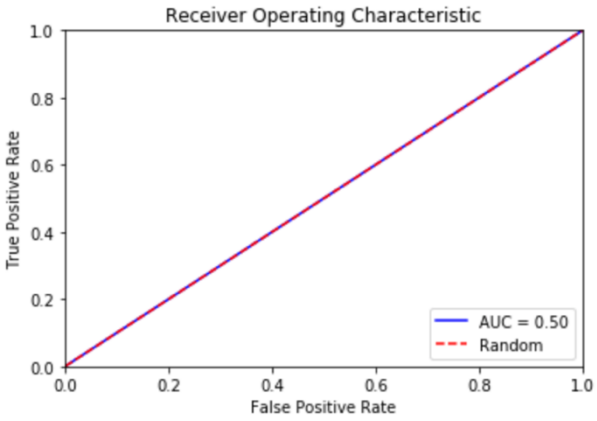

Figure 35 - ROC curve for random forest with Near Miss

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	1	0.4948	0.4948	0.0568	0.0601
Decision Tree	1	0.5025	0.5025	0.0302	0.0311
Random Forest	1	0.5000	0.5000	0.0234	0.0239

Table 44 - Performance metrics with Near Miss on the test set

When using Near Miss, the best performant model according to the AUC metric was the decision tree, although it was quite close to both the other algorithms. This technique performed poorly, even below random guessing for the logistic regression, not being suitable for this data set.

Additionally, the models were also used on a new sample with 100 new subsections to test for the adaptability to the entire city – the main performance metrics can be found in Table 45.

	Ratio	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	1	0.5387	0.5387	0.1077	0.1206
Decision Tree	1	0.5073	0.5073	0.0289	0.0297
Random Forest	1	0.5078	0.5078	0.0264	0.0271

Table 45 - Performance metrics with Near Miss on a new data set

The confusion matrix of each model on the new sample can be seen in tables 46, 47 and 48.

	Predict negative	Predict positive
True negative	629 137	4 594 667
True positive	43	958

Table 46 - Confusion matrix of the logistic regression with Near Miss on the new sample

	Predict negative	Predict positive
True negative	154 385	5 069 419
True positive	15	986

Table 47 - Confusion matrix of the decision tree with Near Miss on the new sample

	Predict negative	Predict positive
True negative	140 556	5 083 248
True positive	11	990

Table 48 - Confusion matrix of the random forest with Near Miss on the new sample

It is interesting to note that when the models trained with Near Miss are applied to a new data set (not under-sampled), they actually perform better than with never seen data from the same subsection. The opposite would be expected, as it has been seen for the other techniques used to balance the data. However, this doesn't mean that there was an error using Near Miss, as models adjust differently to new data.

6. RESULTS AND CONCLUSIONS

As seen in the section above, Near Miss under-sampling was the sampling technique that performed worst, even worse than simple random sampling for the logistic regression. On the other hand, random over and random under-sampling were the best performant strategies, always achieving the best results with the random forest, supporting the principle of wisdom of the crowd. Table 49 aggregates all the results previously presented on the test set.

	Sampling technique	Area under the ROC	Recall	F_1-score	Accuracy
Logistic Regression	Random	0.5	0.5	0.5	0.996
Decision Tree	Random	0.5	0.5	0.5	0.996
Random Forest	Random	0.5	0.5	0.5	0.996
Logistic Regression	Random over	0.6309	0.6309	0.4208	0.7243
Decision Tree	Random over	0.5409	0.5409	0.5045	0.9942
Random Forest	Random over	0.6755	0.6755	0.3939	0.6477
Logistic Regression	SMOTE	0.6427	0.6427	0.4112	0.6964
Decision Tree	SMOTE	0.5300	0.5300	0.4957	0.9778
Random Forest	SMOTE	0.6556	0.6556	0.4098	0.6920
Logistic Regression	Random under	0.6514	0.6514	0.4221	0.7280
Decision Tree	Random under	0.6603	0.6603	0.4123	0.6603
Random Forest	Random under	0.6764	0.6764	0.4138	0.7033
Logistic Regression	Near Miss	0.4948	0.4948	0.0568	0.0601
Decision Tree	Near Miss	0.5025	0.5025	0.0302	0.0311
Random Forest	Near Miss	0.5000	0.5000	0.0234	0.0239

Table 49 - Performance metrics for all models and sampling techniques on the test set

Overall, the random forest algorithm tends to perform better than the logistic regression and the decision tree for this dataset, except when using Near Miss, although the difference is small. The random forest with random under-sampling performed slightly better than the random forest with random over-sampling, with an AUC of 0.6765 and 0.6755, respectively. That strategy also yielded better results in all the other performance criteria considered – recall, f_1-score, and accuracy. However, when testing for the new sample of 10 subsections, the random over-sampling strategy performed slightly better in terms of AUC (0.6587 against 0.6538). In terms of other metrics, random under-sampling achieved better results.

Bearing in mind the slightly better performance of the random forest with random under-sampling on the test set and advantages in terms of time and computational requirements, the model that is considered to best predict the occurrences is a random forest with random under-sampling.

6.1. TEMPORAL ASSUMPTION

The predictive model was done under the strong assumption that the occurrences are not time dependent, therefore allowing to randomly select the training and testing samples. In order to evaluate the accuracy of such assumption, a random forest with the best hyperparameters found previously (100 trees, 6 features and a maximum depth of 5) was trained on the data from 2013 to 2017, which was under-sampled using random under-sampling to a ratio of imbalance of 0.9. The

metrics for its performance on the data of 2018 can be found in Table 50, compared with the random train-test split.

	Area under the ROC	Recall	F_1-score	Accuracy
Cross-temporal validation	0.6794	0.6794	0.4261	0.7388
Random train-test split	0.6764	0.6764	0.4138	0.7033

Table 50 - Performance metrics of random forest with random under-sampling, with and without cross-temporal validation

When using cross-temporal validation, the model performed slightly better than when randomly splitting the sample into a train and test set, allowing for the use of the applied.

6.2. Geographical Generalization

Due to computational limitations, it was not possible to test the model on the entire dataset, using all the subsections. In order to evaluate the ability of the best performant model to adjust to new subsections, 10 new samples of 100 different subsections were created (data the model has never seen). The score of the random forest with random under-sampling for each of these samples is available in Table 51 below.

	Area under the ROC	Recall	F_1-score	Accuracy
Sample 1	0.6031	0.6031	0.4698	0.8844
Sample 2	0.6141	0.6141	0.4680	0.8776
Sample 3	0.5982	0.5982	0.4654	0.8688
Sample 4	0.6212	0.6212	0.4624	0.8581
Sample 5	0.6295	0.6295	0.4715	0.8900
Sample 6	0.6044	0.6044	0.4670	0.8744
Sample 7	0.6010	0.6010	0.4717	0.8914
Sample 8	0.6536	0.6536	0.4729	0.8952
Sample 9	0.6314	0.6314	0.4577	0.8426
Sample 10	0.6243	0.6243	0.4698	0.8846

Table 51 - Performance metrics of the random forest with random under-sampling on 10 new samples

Although the model lost some of its performance when predicting new subsections, it adjusted well to new data. On average, the random forest had an AUC of 0.62.

6.3. Model Interpretation

SHAP values, available in the git repository https://github.com/slundberg/shap, allow to interpret black-box models by attributing a value corresponding to the change each feature impacted the base estimator for the final prediction.

Figure 36 provides a bird-eye view of the impact of each feature for all the training data, ordered by importance. Each point represents an observation, with its horizontal location corresponding to the contribution for the final prediction (the left side means it predicted 0 – no occurrence –, the right side means it predicted 1 – a relevant occurrence) and the color represents whether the feature has a high or low value for that instance. Being so, one can understand from Figure 36 that the size of the

subsection is the most important feature and that smaller subsections tend to be related with fewer occurrences. Precipitation, on the other hand, is related to occurrences when it takes higher values, being the third most important feature. It is interesting to note that humidity plays such a different role than precipitation, being that lower values of humidity are related to positive predictions.

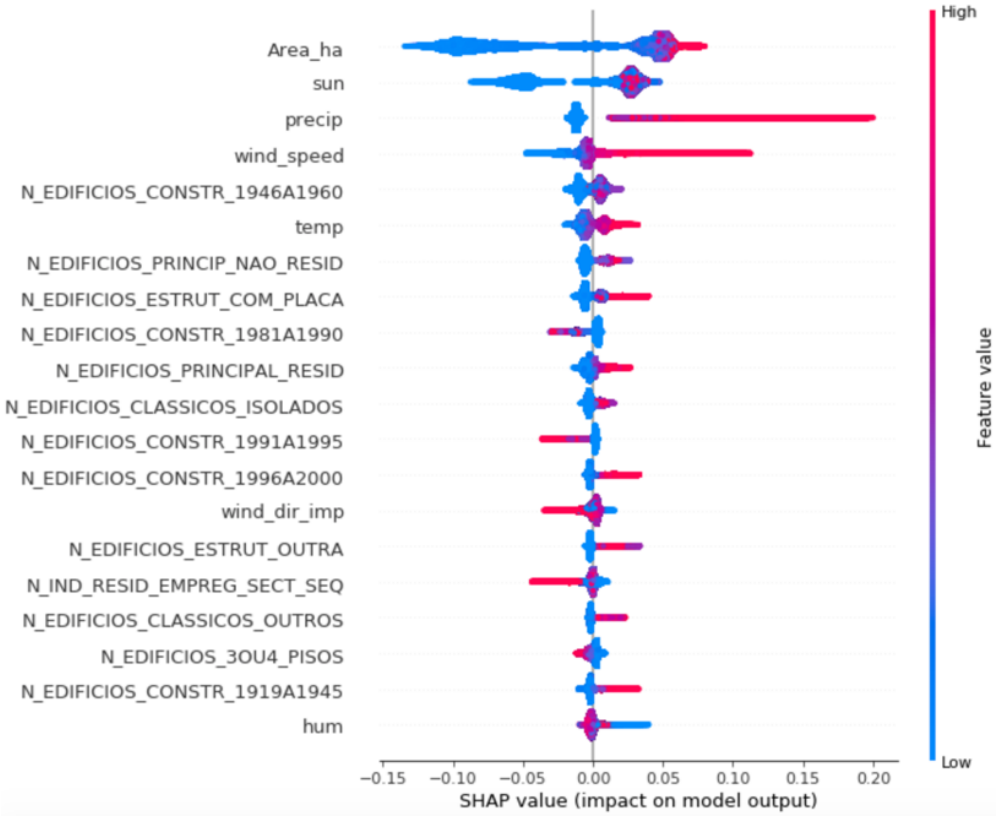

Figure 36 - SHAP values of the best performant model (random forest with random under-sampling)

Figure 37 and Figure 38 show the SHAP values for two observations, depicting the change that each feature impacted in the prediction of the output value. Starting from the base value of 0.4727, the size of the subsection and the sun were the main independent variables leading to an output of 0.31 (Figure 37). Using the information from Figure 36, one can know that this subsection is a small one, as the area has a negative influence on the output value, and that it refers to an hour with low precipitation – higher values of precipitation are related with positive predictions, being necessary a low value to have the opposite impact. Figure 38, on the other hand, has to be a larger subsection at a moment of higher precipitation, which is the variable with a greater impact in pushing the base value to an output of 0.76.

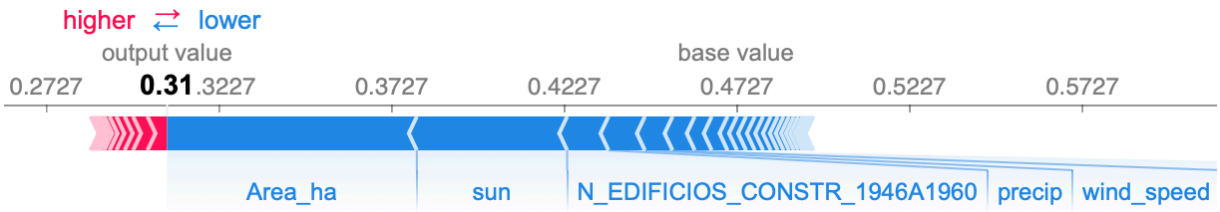

Figure 37 - SHAP values for an observation predicted as 0

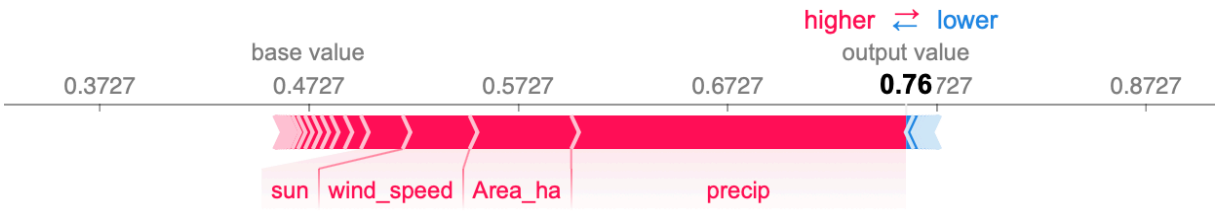

Figure 38 - SHAP values for an observation predicted as 1

6.4. Practical Implications

The objective of this project was to predict occurrences to which the firefighters respond in the city of Lisbon, aiming to contribute to a better allocation of vehicles and therefore a faster response time.

The predictive model created allows to predict the subsections where it is more likely for there to be an occurrence, for each hour according to the expected weather. Allying this information with traffic forecasts, a suggestion of locations that allow to reach every predicted call within the desired timeframe of 5 minutes can be created. In addition, it is possible to understand which are the days (and areas) that require additional teams and take the necessary action: ask the volunteer corporations if they have availability to increase their response or increase the response of the RSB.

SHAP values allowed for a deeper understanding of the impact of each variable in the final model, which is crucial information to improve the city. The fact that areas with more buildings built from 1996 to 2000 and from 1919 to 1945 are more susceptible to occurrences can have practical implications in urban planning and city requirements for those homeowners (Figure 36), such as guaranteeing better access for those areas (larger roads and closer control on vehicles stopped in the road, for example) and ensuring both a good flow of water (preventing floods) and the functioning of fire hydrants (fire response). On the other hand, areas with more buildings built from 1991 to 1995 seem to be less susceptible to occurrences, therefore not requiring additional attention from the city in this regard (Figure 36), although keeping the current level of safety and response.

6.5. Limitations and Future Work

During the realization of this project, the main limitation found was computational power. Working with big data, the high volume required a lot of computational power to process it and train models. Two strategies were used to overcome this obstacle: sampling 100 subsections out of the entire city, which meant working with only around 5 224 000 observations and using Spark in Databricks for the more computationally heavy portion of data training – the over-sampling. However, it was still not possible to use all the techniques and algorithms (over 24 hours wasn't enough to over-sample the data to a ratio of imbalance of 0.1 with ASADYN, for example) and each model would take a long time to train, even using parallel computing in Spark.

Meteorological data was consistently considered as an important predictor of occurrences in the literature. For the purpose of this project, it was only possible to have access to the three weather stations that IPMA has in the city of Lisbon, which translates to little variability. In order to improve the prediction, it is important to have access to this feature with a greater granularity – a future model can make use of the climacteric sensors that have been implemented in Lisbon in 2019.

The fact that census data is only collected every 10 years means that the most recent dataset no longer accurately depicts the city of Lisbon. As new information will only be collected in 2021, the models in this paper used data from 2011. In order to better characterize the city, a future model could use the most recent data and other features that it was not possible to have access at the time, but have been found to be important predictors, as road characteristics (condition, number of lanes, number of exits, …), traffic flow (both pedestrian and automobile), trees (condition, size and species) and land use.

Projects as the underlying this thesis are very relevant to improve the city's response to emergent events. Although the scope of this one is the firefighters, it could be replicated to other first response services, as medical emergency (especially ambulances) and police, and in other locations.

7. BIBLIOGRAPHY

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transportation Research Record, 1897(1), 88-95.

Aksoy, S., & Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. Pattern recognition letters, 22(5), 563-582.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention, 34(6), 729-741.

Allison, P. D. (2001). Missing data (Vol. 136). Sage publications.

Arpaci, A., Malowerschnig, B., Sass, O., & Vacik, H. (2014). Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. Applied Geography, 53, 258-270.

Ayalew, L., & Yamagishi, H. (2005). The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology, 65(1-2), 15-31.

Ayodeji, O. (2011). An examination of the causes and effects of building collapse in Nigeria. Journal of Design and Built environment, 9(1).

Aziz, K., Rahman, A., Fang, G., & Shrestha, S. (2014). Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. Stochastic environmental research and risk assessment, 28(3), 541-554.

Bandara, D., Mayorga, M. E., & McLay, L. A. (2014). Priority dispatching strategies for EMS systems. Journal of the Operational Research Society, 65(4), 572–587.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

Bernard, S., Heutte, L., & Adam, S. (2009). Influence of hyperparameters on random forest accuracy. In International Workshop on Multiple Classifier Systems (pp. 171-180). Springer, Berlin, Heidelberg.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Bruce, P., & Bruce, A. (2017). Practical statistics for data scientists: 50 essential concepts. " O'Reilly Media, Inc.".

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. Journal of the American society for information science, 45(1), 12-19.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

Celik, O. C., & Ellingwood, B. R. (2010). Seismic fragilities for non-ductile reinforced concrete frames – Role of aleatoric and epistemic uncertainties. Structural Safety, 32(1), 1–12.

Chan, C. L., Chen, Y. J., Chen, K. P., & Chiu, S. J. (2010). Elderly inpatient fall risk factors: A study of decision tree and logistic regression. In The 40th International Conference on Computers & Indutrial Engineering (pp. 1-6). IEEE.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Church, R., & Velle, C. R. (1974). The maximal covering location problem. Papers in Regional Science, 32(1), 101–118.

Cochran, W. G. (2007). Sampling techniques. John Wiley & Sons.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (pp. 598-617). IEEE.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122–135

De Vasconcelos, M. P., Silva, S., Tome, M., Alvim, M., & Pereira, J. C. (2001). Spatial prediction of fire ignition probabilities: comparing logistic regression and neural networks. Photogrammetric engineering and remote sensing, 67(1), 73-81.

Denham, M., Wendt, K., Bianchini, G., Cortés, A., & Margalef, T. (2012). Dynamic Data-Driven Genetic Algorithm for forest fire spread prediction. Journal of Computational Science, 3(5), 398–404.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465, 1-20.

Dubes, R. C., & Jain, A. K. (1988). Algorithms for clustering data.

Dwoskin, E. (2014, January 24). How New York's Fire Department Uses Data Mining. Retrieved July 12, 2019, from https://blogs.wsj.com/digits/2014/01/24/how-new-yorks-fire-department-uses-data-mining/.

Enders, C., & Bandalos, D. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 8(3), 430–457.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. Psychological methods, 16(1), 1.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Feero, S., Hedges, J. R., Simmons, E., & Irwin, L. (1995). Does out-of-hospital EMS time affect trauma survival? The American Journal of Emergency Medicine, 13(2), 133–135.

Gad, I., & Manjunatha, B. R. (2017). Performance evaluation of predictive models for missing data imputation in weather data. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1327-1334). IEEE.

Gårder, P. E. (2004). The impact of speed and other variables on pedestrian safety in Maine. Accident Analysis & Prevention, 36(4), 533–542.

Gonzalez, M. E., Ogus, J. L., Shapiro, G., & Tepping, B. J. (1975). Standards for discussion and presentation of errors in survey and census data. Journal of the American Statistical Association, 70(351b), 5-23.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval (pp. 345-359). Springer, Berlin, Heidelberg.

Greibe, P. (2003). Accident prediction models for urban roads. Accident Analysis & Prevention, 35(2), 273–285.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Hall, M. A. (1999). Correlation-based feature selection for machine learning.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.

Han, S., & Coulibaly, P. (2019). Probabilistic Flood Forecasting Using Hydrologic Uncertainty Processor with Ensemble Weather Forecasts. Journal of Hydrometeorology, 20(7), 1379–1398.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning, 77(1), 103-123.

Harmsen, A. M. K., Giannakopoulos, G. F., Moerbeek, P. R., Jansma, E. P., Bonjer, H. J., & Bloemers, F. W. (2015). The influence of prehospital time on trauma patients outcome: A systematic review. Injury, 46(4), 602–609.

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, 73, 220-239.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (pp. 1322-1328). IEEE.

Hogan, K., & ReVelle, C. (1986). Concepts and Applications of Backup Coverage. Management Science, 32(11), 1434–1444.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.

How Cluster and Outlier Analysis (Anselin Local Moran's I) works. (n.d.). Retrieved October 28, 2019, from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm.

How Hot Spot Analysis (Getis-Ord Gi*) works. (n.d.). Retrieved October 28, 2019, from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm.

Instituto Nacional da Casa da Moeda. (n.d.). Lei 22/2008, 2008-05-13. Retrieved October 21, 2019, from https://dre.pt/pesquisa/-/search/249237/details/maximized.

Instituto Nacional de Estatística. (n.d.). Censos 2011: Perguntas Frequentes. Retrieved October 21, 2019, from https://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=ine_faqs_censos2011.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE.

Jones, A. P., & Jørgensen, S. H. (2003). The use of multilevel models for the prediction of road accident outcomes. Accident Analysis & Prevention, 35(1), 59-69.

Kolesar, P., & Blum, E. H. (1973). Square Root Laws for Fire Engine Response Distances. Management Science, 19(12), 1368–1378.

Kolesar, P., & Walker, W. E. (1974). An Algorithm for the Dynamic Relocation of Fire Companies. Operations Research, 22(2), 249–274.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

Lau, C. K., Lai, K. K., Lee, Y. P., & Du, J. (2015). Fire risk assessment with scoring system, using the support vector machine approach. Fire Safety Journal, 78, 188–195.

Leden, L. (2002). Pedestrian risk decrease with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario. Accident Analysis & Prevention, 34(4), 457–464.

Lerner, E. B. (2003). Is Total Out-of-hospital Time a Significant Predictor of Trauma Patient Mortality? Academic Emergency Medicine, 10(9), 949–954.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. In Kdd (Vol. 98, pp. 73-79).

Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.

Lopes, A., Oliveira, S., Fragoso, M., Andrade, J. A., & Pedro, P. (2009). Wind Risk Assessment in Urban Environments: The Case of Falling Trees During Windstorm Events in Lisbon. In Bioclimatology and Natural Hazards (pp. 55–74). Springer Netherlands.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Madasseri Payyappalli, V. (2019). Data-Driven Fire Risk Management: Spatio-Temporal Prediction and Resource Allocation Models (Doctoral dissertation, State University of New York at Buffalo).

Mani, I., & Zhang, I. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets (Vol. 126).

Mantovani, R. G., Horváth, T., Cerri, R., Vanschoren, J., & de Carvalho, A. C. (2016). Hyper-parameter tuning of a decision tree induction algorithm. In 2016 5th Brazilian Conference on Intelligent Systems (BRACIS) (pp. 37-42). IEEE.

McCue, C. (2006). Data Mining and Predictive Analytics in Public Safety and Security. IT Professional, 8(4), 12–18.

Mountain, L., & Fawaz, B. (1996). Estimating accidents at junctions using routinely available input data. Traffic engineering & control. Traffic Engineering & Control 37 (11), 624–628.

Nakai, M., & Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. International Journal of Mathematical Analysis, 5(1), 1-13.

Nefeslioglu, H. A., Sezer, E., Gokceoglu, C., Bozkir, A. S., & Duman, T. Y. (2010). Assessment of landslide susceptibility by decision trees in the metropolitan area of Istanbul, Turkey. Mathematical Problems in Engineering, 2010.

Nordin, N. A. M., Zaharudin, Z. A., Maasar, M. A., & Nordin, N. A. (2012). Finding shortest path of the ambulance routing: Interface of A* algorithm using C# programming. In 2012 IEEE Symposium on Humanities, Science and Engineering Research. IEEE.

Ohlmacher, G. C., & Davis, J. C. (2003). Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Engineering geology, 69(3-4), 331-343.

Olken, F. (1993). Random sampling from databases (Doctoral dissertation, University of California, Berkeley).

Olken, F., & Rotem, D. (1986). Simple random sampling from relational databases.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. Contemporary Mathematics, 443(7), 59-72.

Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Computers in biology and medicine, 41(5), 265-271.

Palei, S. K., & Das, S. K. (2009). Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. Safety Science, 47(1), 88-96.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015). Using random undersampling to alleviate class imbalance on tweet sentiment data. In 2015 IEEE international conference on information reuse and integration (pp. 197-202). IEEE.

Quinn, G. P., & Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge university press.

Revelle, C., & Hogan, K. (1989). The maximum reliability location problem and α-reliablep-center problem: Derivatives of the probabilistic location set covering problem. Annals of Operations Research, 18(1), 155–173.

ReVelle, C., & Marianov, V. (1991). A probabilistic FLEET model with individual vehicle reliability requirements. European Journal of Operational Research, 53(1), 93–105.

ReVelle, C. (1991). Siting Ambulances and Fire Companies: New Tools for Planners. Journal of the American Planning Association, 57(4), 471–484.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.

Rubin, D. (1976). Inference and Missing Data. Biometrika, 63(3), 581-592.

Sampalis, J. S., Lavoie, A., Williams, J. I., Mulder, D. S., & Kalina, M. (1993). Impact of on-site care, prehospital time, and level of in-hospital care on survival in severely injured patients. The Journal of Trauma: Injury, Infection, and Critical Care, 34(2), 252–261.

Shanthi, S., & Ramani, R. G. (2012). Feature relevance analysis and classification of road traffic accident data through data mining techniques. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 24-26).

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. Boston University, 23.

Stojanova, D., Panov, P., Kobler, A., Dzeroski, S., & Taskova, K. (2006). Learning to predict forest fires with different data mining techniques. In Conference on data mining and data warehouses (SiKDD 2006), Ljubljana, Slovenia (pp. 255-258).

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. International Educational Data Mining Society.

Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics, 15(3), 142-147.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288.

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27(14), 1986-1994.

van Buuren, M., Kommer, G. J., van der Mei, R., & Bhulai, S. (2015). A simulation model for emergency medical services call centers. In 2015 Winter Simulation Conference (WSC). IEEE.

Vieira Gomes, S. (2013). The influence of the infrastructure characteristics in urban road accidents occurrence. Accident Analysis & Prevention, 60, 289–297.

Viglino, D., Vesin, A., Ruckly, S., Morelli, X., Slama, R., Debaty, G., … Timsit, J.-F. (2017). Daily volume of cases in emergency call centers: construction and validation of a predictive model. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 25(1).

Watts, Jr., J. M., & Kaplan, M. E. (2001). Fire Technology, 37(2), 167–180.

Watts, Jr., J. M. (2003). Fire-Risk Indexing: A Systemic Approach to Building-Code “Equivalency” for Historic Buildings. APT Bulletin, 34(4), 23.

Wang, X., Fan, T., Chen, M., Deng, B., Wu, B., & Tremont, P. (2015). Safety modeling of urban arterials in Shanghai, China. Accident Analysis & Prevention, 83, 57–66.

White, J. A., & Case, K. E. (1974). On Covering Problems and the Central Facilities Location Problem. Geographical Analysis, 6(3), 281–294.

Xin, J., & Huang, C. (2013). Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management. Fire Safety Journal, 62, 72–78.

Zhuang, Y., Yu, K., Wang, D., & Ding, W. (2016). An evaluation of big data analytics in feature selection for long-lead extreme floods forecasting. In 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC). IEEE.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320.

8. APPENDIX

Variable name	Variable description
Area_ha	Area of the subsection
N_EDIFICIOS_CLASSICOS	Nº of classical buildings
N_EDIFICIOS_CLASSICOS_1OU2	Nº of classical buildings with 1 or 2 apartments
N_EDIFICIOS_CLASSICOS_ISOLADOS	Nº of isolated classical buildings
N_EDIFICIOS_CLASSICOS_GEMIN	Nº of classical buildings built in pairs (side by side)
N_EDIFICIOS_CLASSICOS_EMBANDA	Nº of classical buildings with more than 3 joined in a row
N_EDIFICIOS_CLASSICOS_3OUMAIS	Nº of classical buildings with 3 or more apartments
N_EDIFICIOS_CLASSICOS_OUTROS	Nº of other classical buildings
N_EDIFICIOS_EXCLUSIV_RESID	Nº of buildings exclusively residential
N_EDIFICIOS_PRINCIPAL_RESID	Nº of buildings mostly residential
N_EDIFICIOS_PRINCIP_NAO_RESID	Nº of buildings mostly non-residential
N_EDIFICIOS_1OU2_PISOS	Nº of buildings with 1 or 2 floors
N_EDIFICIOS_3OU4_PISOS	Nº of buildings with 3 or 4 floors
N_EDIFICIOS_5OU_MAIS_PISOS	Nº of buildings with 5 or more floors
N_EDIFICIOS_CONSTR_ANTES_1919	Nº of buildings built prior to 1919
N_EDIFICIOS_CONSTR_1919A1945	Nº of buildings built between 1919 and 1945
N_EDIFICIOS_CONSTR_1946A1960	Nº of buildings built between 1946 and 1960
N_EDIFICIOS_CONSTR_1961A1970	Nº of buildings built between 1961 and 1970
N_EDIFICIOS_CONSTR_1971A1980	Nº of buildings built between 1971 and 1980
N_EDIFICIOS_CONSTR_1981A1990	Nº of buildings built between 1981 and 1990
N_EDIFICIOS_CONSTR_1991A1995	Nº of buildings built between 1991 and 1995
N_EDIFICIOS_CONSTR_1996A2000	Nº of buildings built between 1996 and 2000
N_EDIFICIOS_CONSTR_2001A2005	Nº of buildings built between 2001 and 2005
N_EDIFICIOS_CONSTR_2006A2011	Nº of buildings built between 2006 and 2011
N_EDIFICIOS_ESTRUT_BETAO	Nº of buildings with reinforced steel structure
N_EDIFICIOS_ESTRUT_COM_PLACA	Nº of buildings of masonry walls with steel structure
N_EDIFICIOS_ESTRUT_SEM_PLACA	Nº of buildings of masonry walls without steel structure
N_EDIFICIOS_ESTRUT_ADOBE_PEDRA	Nº of buildings of adobe and stone
N_EDIFICIOS_ESTRUT_OUTRA	Nº of buildings with other structure
N_ALOJAMENTOS_VAGOS	Nº of empty houses
N_INDIVIDUOS_PRESENT	Nº of people in the house (or arriving within 12h)
N_INDIVIDUOS_PRESENT_H	Nº of male people in the house (or arriving within 12h)
N_INDIVIDUOS_PRESENT_M	Nº of female people in the house (or arriving within 12h)
N_INDIVIDUOS_RESIDENT	Nº of people that live in the house
N_INDIVIDUOS_RESIDENT_H	Nº of male people that live in the house
N_INDIVIDUOS_RESIDENT_M	Nº of female people that live in the house
N_INDIVIDUOS_RESIDENT_0A4	Nº of people that live in the house aged between 0 and 4
N_INDIVIDUOS_RESIDENT_5A9	Nº of people that live in the house aged between 5 and 9
N_INDIVIDUOS_RESIDENT_10A13	Nº of people that live in the house aged between 10 and 13
N_INDIVIDUOS_RESIDENT_14A19	Nº of people that live in the house aged between 14 and 19

N_INDIVIDUOS_RESIDENT_15A19	Nº of people that live in the house aged between 15 and 19
N_INDIVIDUOS_RESIDENT_20A24	Nº of people that live in the house aged between 20 and 24
N_INDIVIDUOS_RESIDENT_20A64	Nº of people that live in the house aged between 20 and 64
N_INDIVIDUOS_RESIDENT_25A64	Nº of people that live in the house aged between 25 and 64
N_INDIVIDUOS_RESIDENT_65	Nº of people that live in the house older than 64
N_INDIVIDUOS_RESIDENT_H_0A4	Nº of men that live in the house aged between 0 and 4
N_INDIVIDUOS_RESIDENT_H_5A9	Nº of men that live in the house aged between 5 and 9
N_INDIVIDUOS_RESIDENT_H_10A13	Nº of men that live in the house aged between 10 and 13
N_INDIVIDUOS_RESIDENT_H_14A19	Nº of men that live in the house aged between 14 and 19
N_INDIVIDUOS_RESIDENT_H_15A19	Nº of men that live in the house aged between 15 and 19
N_INDIVIDUOS_RESIDENT_H_20A24	Nº of men that live in the house aged between 20 and 24
N_INDIVIDUOS_RESIDENT_H_20A64	Nº of men that live in the house aged between 20 and 64
N_INDIVIDUOS_RESIDENT_H_25A64	Nº of men that live in the house aged between 25 and 64
N_INDIVIDUOS_RESIDENT_H_65	Nº of men that live in the house older than 64
N_INDIVIDUOS_RESIDENT_M_0A4	Nº of women that live in the house aged between 0 and 4
N_INDIVIDUOS_RESIDENT_M_5A9	Nº of women that live in the house aged between 5 and 9
N_INDIVIDUOS_RESIDENT_M_10A13	Nº of women that live in the house aged between 10 and 13
N_INDIVIDUOS_RESIDENT_M_14A19	Nº of women that live in the house aged between 14 and 19
N_INDIVIDUOS_RESIDENT_M_15A19	Nº of women that live in the house aged between 15 and 19
N_INDIVIDUOS_RESIDENT_M_20A24	Nº of women that live in the house aged between 20 and 24
N_INDIVIDUOS_RESIDENT_M_20A64	Nº of women that live in the house aged between 20 and 64
N_INDIVIDUOS_RESIDENT_M_25A64	Nº of women that live in the house aged between 25 and 64
N_INDIVIDUOS_RESIDENT_M_65	Nº of women that live in the house older than 64
N_INDIV_RESIDENT_N_LER_ESCRV	Nº of residents that can't read nor write
N_IND_RESIDENT_FENSINO_1BAS	Nº of residents going to primary school
N_IND_RESIDENT_FENSINO_2BAS	Nº of residents going to 5^{th} or 6^{th} grade
N_IND_RESIDENT_FENSINO_3BAS	Nº of residents going to 7^{th}, 8^{th} or 9^{th} grade
N_IND_RESIDENT_FENSINO_SEC	Nº of residents going to high school
N_IND_RESIDENT_FENSINO_POSSEC	Nº of residents going to a pos high school course
N_IND_RESIDENT_FENSINO_SUP	Nº of residents going to university
N_IND_RESIDENT_ENSINCOMP_1BAS	Nº of residents with primary school
N_IND_RESIDENT_ENSINCOMP_2BAS	Nº of residents with 6^{th} grade
N_IND_RESIDENT_ENSINCOMP_3BAS	Nº of residents with 9^{th} grade
N_IND_RESIDENT_ENSINCOMP_SEC	Nº of residents with high school
N_IND_RESIDENT_ENSINCOMP_POSEC	Nº of residents with a pos high school course
N_IND_RESIDENT_ENSINCOMP_SUP	Nº of residents with a university degree
N_IND_RESID_DESEMP_PROC_1EMPRG	Nº of residents unemployed looking for their first job
N_IND_RESID_DESEMP_PROC_EMPRG	Nº of residents unemployed looking for a new job
N_IND_RESID_EMPREGADOS	Nº of residents with a job
N_IND_RESID_PENS_REFORM	Nº of retired residents
N_IND_RESID_SEM_ACT_ECON	Nº of residents without economic activity
N_IND_RESID_EMPREG_SECT_PRIM	Nº of residents working in the primary sector
N_IND_RESID_EMPREG_SECT_SEQ	Nº of residents working in the secondary sector

N_IND_RESID_EMPREG_SECT_TERC	Nº of residents working in the tertiary sector
N_IND_RESID_ESTUD_MUN_RESID	Nº of residents studying in the municipality where they live
N_IND_RESID_TRAB_MUN_RESID	Nº of residents working in the municipality where they live

Table 52 - Census data variables

Ocorrência	Target	2013	2014	2015	2016	2017	2018	Total
1100 - Incêndio- Povoamento Florestal	1	8	1	1	1	8	1	20
1200 - Incêndio- Agricula	1			1	1			2
1300 - Incêndio- Inculto	1	193	132	144	116	82	42	709
1401 - Incêndio- Edifício (Infra-estrutura/Instalação) - Habitação	1	118	125	98	130	117	122	710
1402 - Incêndio- Edifício (Infra-estrutura/Instalação) - Estacionamento	1	3	3		2	5	4	17
1403 - Incêndio- Edifício (Infra-estrutura/Instalação) - Serviços	1	1	3	5	4	3	7	23
1404 - Incêndio- Edifício (Infra-estrutura/Instalação) - Escolar	1	1			2	1	2	6
1405 - Incêndio- Edifício (Infra-estrutura/Instalação) - Hospitalar/Lar	1		2	3	2	2		9
1406 - Incêndio- Edifício (Infra-estrutura/Instalação) - Espectáculo/Lazer/Culto Religioso	1		1	1	1	1		4
1407 - Incêndio- Edifício (Infra-estrutura/Instalação) - Hoteleira e similar	1	14	12	10	12	14	9	71
1408 - Incêndio- Edifício (Infra-estrutura/Instalação) - Comercial/Lojas/Feiras/Gare de Transporte	1	4	6	1	1		2	14
1410 - Incêndio- Edifício (Infra-estrutura/Instalação) - Militar/Forças Segurança	1				1			1
1411 - Incêndio- Edifício (Infra-estrutura/Instalação) - Indústria/Oficina/Armazém	1	4	5	4	3	2	4	22
1420 - Incêndio- Edifício (Infra-estrutura/Instalação) - Edifício Devoluto/Degradado	1	14	7	18	18	7	6	70
1500 - Incêndio- Equipamentos (sem afectação do ambiente)	1	2	6	2		1	1	12
1501 - Incêndio- Equipamentos (sem afectação do ambiente) - Contentores de lixo	1	143	138	132	94	96	105	708
1701 - Incêndio- Transportes - Rodoviário	1	81	78	60	84	71	89	463
1702 - Incêndio- Transportes - Aéreo	1			1			1	2
1703 - Incêndio- Transportes - Ferroviário	1						1	1
1704 - Incêndio- Transportes - Aquático	1				1		2	3
1800 - Incêndio- Detritos	1	258	211	315	187	192	156	1319
2101 - Acidentes - Rodoviários - Atropelamento	1	9	33	58	98	106	116	420
2102 - Acidentes - Rodoviários - Com viaturas	1	166	358	465	653	647	753	3042
2103 - Acidentes - Rodoviários - C/ Encarcerados	1	80	71	74	81	72	78	456
2301 - Acidentes - Ferroviário - Atropelamento	1	5	3	1	3	8	4	24
2303 - Acidentes - Ferroviário - Choque	1		2					2
2304 - Acidentes - Ferroviário - Descarrilamento	1		1				1	2
2305 - Acidentes - Ferroviário - Com Encarcerados	1	1	1	1				3
2400 - Acidentes - Aquático	1			1				1

2401 - Acidentes - Aquático - Queda ao Rio	1	5	7	4	2	6	3	27
2500 - Acidentes - Equipamentos	1	4	1	4	1	2	2	14
2501 - Acidentes - Equipamentos - Elevadores	1	216	184	214	223	207	265	1309
3100 - Infra-estruturas e Vias de Comunicação - Queda de Árvore	1	371	195	190	127	195	234	1312
3202 - Infra-estruturas e Vias de Comunicação - Corte de abastecimento - Electricidade	1	1						1
3300 - Infra-estruturas e Vias de Comunicação - Desabamento	1	7	16	2	10	1	4	40
3301 - Infra-estruturas e Vias de Comunicação - Desabamento - Queda de Revestimento	1	361	425	250	287	282	272	1877
3400 - Infra-estruturas e Vias de Comunicação - Deslizamento	1		1		1	1	2	5
3500 - Infra-estruturas e Vias de Comunicação - Inundação	1	433	644	211	337	12	9	1646
3501 - Infra-estruturas e Vias de Comunicação - Inundação Espaço Privado	1					98	232	330
3502 - Infra-estruturas e Vias de Comunicação - Inundação Espaço Publico	1					48	131	179
3600 - Infra-estruturas e Vias de Comunicação - Desentupimento/Tamponamento	1	7	7	6	10	11	3	44
3700 - Infra-estruturas e Vias de Comunicação - Danos/Queda Cabos Eléctricos	1	16	23	30	28	39	38	174
3701 - Infra-estruturas e Vias de Comunicação - Danos/Queda Cabos Eléctricos - Curto-circuito	1	109	128	121	117	103	121	699
3800 - Infra-estruturas e Vias de Comunicação - Queda de Estruturas	1	250	270	170	210	195	260	1355
4100 - Pré-Hospitalar - Intoxicação	0	4	3	2	4	5	1	19
4200 - Pré-Hospitalar - Doença Súbita	0	764	752	905	540	340	387	3688
4300 - Pré-Hospitalar - Traumatismo/Queda	0	162	145	208	89	116	110	830
4400 - Pré-Hospitalar - Queimado	0				2	1		3
4500 - Pré-Hospitalar - Parto	0	5	4	3	4	3	3	22
4600 - Pré-Hospitalar - Afogamento	0				1			1
5101 - Conflitos Legais - Explosivos - Ameaça	0		1	1				2
5102 - Conflitos Legais - Explosivos - Explosão	0			3				3
5200 - Conflitos Legais - Agressão/Violação	0	5	2	13	2	7	4	33
5301 - Conflitos Legais - Suicídio/Homicídio - Tentativa	0	8	10	3	8	9	9	47
5302 - Conflitos Legais - Suicídio/Homicídio - Consumado	0	1		2				3
5600 - Conflitos Legais - Apoio à Autoridade	0	28	35	33	44	27	27	194
6102 - Tecnológicos Industriais - Acidentes Matérias Perigosas - Químicos	0	4	1		3	5	2	15
6301 - Tecnológicos Industriais - Fuga de Gás - Canalização/Conduta	0	112	113	107	76	85	85	578
6302 - Tecnológicos Industriais - Fuga de Gás - Garrafa	0	16	10	10	19	10	9	74
6303 - Tecnológicos Industriais - Fuga de Gás - Depósito/Reservatório	0		1		2	1	2	6
6401 - Tecnológicos Industriais - Situações Suspeitas - Verificar Fumos	0	111	134	116	135	160	170	826
6402 - Tecnológicos Industriais - Situações Suspeitas - Verificar Cheiros	0	87	131	131	121	164	149	783

6403 - Tecnológicos Industriais - Situações Suspeitas - Verificar SADI/Alarmes	0	60	57	58	56	66	56	353
7101 - Serviços - Prevenções - Patrulhamento/Vigilância	0	4	12	7	14	94	125	256
7102 - Serviços - Prevenções - Espectáculo	0	13	33	16	11	20	41	134
7103 - Serviços - Prevenções - Desporto	0	20	26	18	10	6	16	96
7104 - Serviços - Prevenções - Queimadas	0	1	1	1	1			4
7105 - Serviços - Prevenções - Transportes	0		1					1
7106 - Serviços - Prevenções - Pré-Posicionamento Meios	0	22	39	15	34	37	51	198
7200 - Serviços - Limpeza de Via/Conservação	0	246	200	224	228	186	145	1229
7201 - Serviços - Limpeza de Via/Conservação - Sinalizar Buraco	0	119	175	102	146	83	114	739
7202 - Serviços - Limpeza de Via/Conservação - Óleo no Pavimento	0	375	396	352	335	332	276	2066
7301 - Serviços - Abastecimento de Água - População	0	2			4	1		7
7302 - Serviços - Abastecimento de Água - Entidade Pública	0	8	18	15	16	13	11	81
7303 - Serviços - Abastecimento de Água - Entidade Privada	0	6	7	2	3	2	2	22
7401 - Serviços - Abertura de Porta - Com Socorro	1	843	788	904	890	805	834	5064
7402 - Serviços - Abertura de Porta - Sem Socorro	0	1318	1344	1332	1069	790	662	6515
7500 - Serviços - Fecho de água	0	1433	1581	1631	1525	1462	1398	9030
7600 - Serviços - Reboque/Desempanagem	0	8	5	2	3	4	8	30
7701 - Serviços - Transporte Doentes - Geral	0			1	4	3	1	9
7702 - Serviços - Transporte Doentes - Inter-Hospital	0		1					1
7703 - Serviços - Transporte Doentes - Auxílio p/ Transporte de Doentes	0	145	116	104	133	187	181	866
7800 - Serviços - Resgate/Recolha de Animais	0	104	133	169	191	172	137	906
8201 - Actividades - Busca/Resgate (Pessoas e Animais) - Terrestre	1	8	6	5	13	19	7	58
8202 - Actividades - Busca/Resgate (Pessoas e Animais) - Aquático	1	2	3	2	3	2	5	17
8301 - Actividades - Operações Nacionais - Socorro	0				1			1
8302 - Actividades - Operações Nacionais - Assistencia	0					1		1
8500 - Actividades - Exercício/Simulacro	0	20	14	18	11	18	35	116
8603 - Actividades - Deslocações - Serviço Geral	0			5	15	23	43	86
8700 - Actividades - Assistência à População/Apoio Social	0	412	414	455	463	506	529	2779
8701 - Actividades - Acompanhamento Hospitalar	0					42	74	116
8702 - Actividades - Acompanhamento Consultas	0					17	32	49
8703 - Actividades - Acompanhamento Retornos	0					14	17	31
8710 - Actividades - Apoio Social Monitorização	0					17	13	30
8711 - Actividades - Apoio Social Avaliação	0					130	123	253
8711 - Actividades - Apoio Social Visitas Pontuais	0					3		3
8712 - Actividades - Apoio Social Sinalização	0					26	21	47
8720 - Actividades - Tele assistência Adesão	0					91	80	171
8721 - Actividades - Tele assistência Manutenção	0					76	81	157

Avarias							
8722 - Actividades - Tele assistência Recolha Equipamentos	0				17	19	36
8723 - Actividades - Tele Assist. Documentação Expediente	0				7	18	25
8730 - Actividades - Reuniões Eventos	0					7	7
9600 - Eventos de Protecção Civil - Visita Técnica	0				3	3	6
Grand Total		**9361**	**9812**	**9573**	**9077**	**8843**	**9205**	**55871**

Table 53 - List of occurrence types from RSB, with respective target

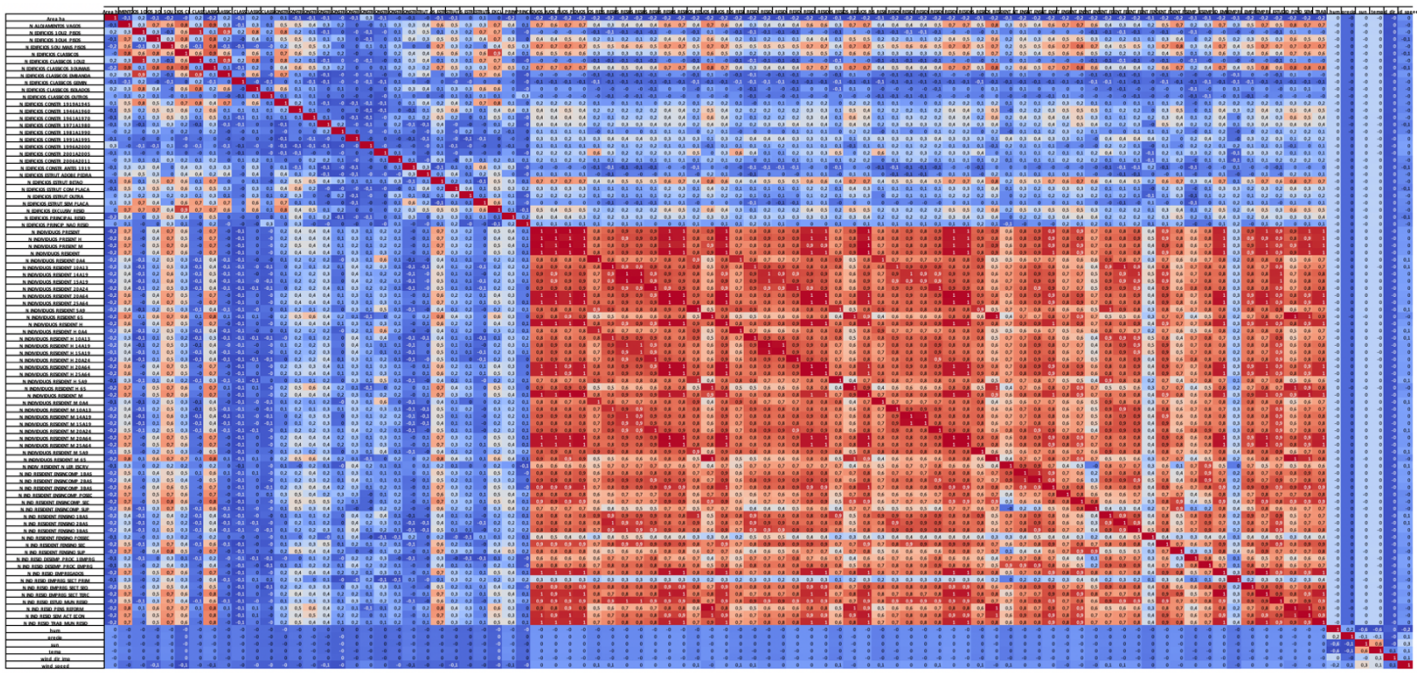

Figure 39 - Correlation matrix with all features (prior to feature selection)