NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação
Master Program in Statistics and Information Management

Project success prediction in the brazilian crowdfunding ecosystem:

A Case Study of Benfeitoria.com

Rodrigo Nogueira de Carvalho

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and 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

Success prediction in the brazilian crowdfunding ecosystem:

A Case Study of Benfeitoria.com platform

by

Rodrigo Nogueira de Carvalho

Project Work report presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

Advisor: Professor Fernando Bação, PhD

July, 2018

ABSTRACT

Crowdfunding in developing markets is growing steady every year and became an opportunity for small project owners to fund their ideas and products. In 2015, South America market generated at $ 85.74M in funding volume. With the growth and maturation of crowdfunding platforms in South America, a more data-driven approach is needed from the thousands of data generated from the relationship between donors and projects to understand the behavioral patterns of a successful project. Even though it is a recent topic the crowdfunding phenomenon has increased the interest of researchers and investors around the globe. Many studies are being developed in the areas of predictive modelling and data mining using the database of the main crowdfunding platforms of the world such as Kickstarter. However, there are few studies related with predictive models on crowdfunding platforms in developing markets, such as Brazil. This case study aims to analyze crowdfunding success factors for a brazilian crowdfunding platform (Benfeitoria) and develop a predictive model that identifies if a new project will be successfully backed after seven days in the platform.

KEYWORDS

Predict model; Data Mining; Model; Brazil, crowdfunding

INDEX

1. Introduction 5
1.1. Study objectives 6
1.2. Relevance and importance 7
2. Literature review 9
3. Methodology 13
3.1 Project development 14
4. Results and discussion 19
4.1. Practical implications 23
5. Conclusions and considerations 24
6. Bibliography 27
7. Appendix 32

1. Introduction

Crowdfunding in developing markets is growing steady every year and became an opportunity for project owners to fund their ideas and products The World Bank predicted the potential market for crowdfunding will reach $96 billion in the next twenty-five years.

In 2015, South America crowdfunding market generated $85.74M dollars in funding volume, according to financial analysis firm, Massolution. (Crowdsourcing.org, 2015).

The World Bank also present a study of the crowdfunding ecosystem in the region investigating why the region has not reached its potential due to some key factors, such as regulatory framework and investment for early-stage funding marketplace. (The World Bank, 2013)

The firsts crowdfunding websites appear in Brazil in 2011, more than 80 platforms where created at that time, targeting on different scope of segments. With the evolution of the market, started a natural selection of websites and in 2018 three platforms stand out as the mainstream players in the market. Catarse, Kickante, and Benfeitoria.

With the growth and maturation of crowdfunding platforms, a more data-driven approach is needed from the amount of data generated from the relationship between donors and projects, to understand the behavioral patterns underneath a successful project, improving predictive analysis based on objective measurements (Ahlemeyer-Stubbe & Coleman, 2014; Kantardzic, 2011)

For this case study, Benfeitoria dataset will be used to develop a predictive model. The company was founded in 2011 and target the segment of social and cultural crowdfunding projects. It was the first platform for-profit in the world that do not charge fixed commission of the projects, letting the users (project donors and project owners) choose how much they want to donate to the platform if the project is successful.

Benfeitoria team develops a CRM approach in the platform to increase the project chances of getting a donor. They offer personal consulting along the crowdfunding campaign for each project that is registered in the platform with educational material. The objective is to help project owners to plan and execute better their digital marketing strategy during a campaign (Hui, Greenberg, & Gerber, 2013).

The company has an unusual business model based on donation and trust agreement between the platform and project owners (Milinski, 2016). They need to act quickly and identify in the CRM strategy what are the projects with more chances of succeed to amplify the engagement and flag project owners with low chances to achieve their goals for improve their effort in marketing until the deadline. (I. J. Chen & Popovich, 2003).

For this study, we will test predictive models, such as decision trees, that presented positive results with similar challenge.

Through this predictive algorithm, we could analyze using a specific selection of variables, what are the probabilities of each new project in the platform to be successful at the end of a period of 7 days after the campaign started. On average, a crowdfunding campaign has 50 days.

1.1. Study objectives

The main goal of our study is to develop a predictive model to measure which projects are most likely to succeed in order to attend their CRM strategy of supporting "best projects" (projects that are more likely to generate income for the platform).

We want to analyze and understand what are the factors responsible for engaging the first donors, which variables impact more the success of a project and the behavior of donors using primary data (Benfeitoria dataset) and secondary data (studies and researches that analyze the behavior of the crowdfunding donor's in Brazil).

In pursuance of project's main goal, specific objectives have been defined and enumerated:

1. Research about crowdfunding behavior pattern for project donors and project owners through data exploration and machine learning methods, by reviewing the literature, to select the most adequate methods and techniques to apply;

2. Evaluate the accuracy of the created models

1.2. Relevance and importance

Academic literature analyzing the success factors of crowdfunding projects in developing markets, such as Brazil, is very limited. In 2014 the Catarse platform presented a study of 3336 individuals from their database to analyze the current situation of crowdfunding in the country using quantitative and qualitative data to present a profile of who are the donors and project owners in Brazil, their socialcultural profile, education level, geographic location and most common type of project categories. ("Retrato financiamento coletivo Brasil - Catarse e Chorus," n.d.)

The survey concentrates on two main areas: Profiling the donors and project owner in Brazil and understand the behavior and motivation for supporting a project.

The results showed that most people who participate in a crowdfunding campaign are men (59%) and the predominant age group is between 25 and 30 years (31%), with a higher educational level. Donors also presented three characteristics to support a project: Transparency during the campaign, project quality and reward offered by the project.

Some authors present initial studies to understand some characteristics of successful and failures crowdfunding campaigns based on the largest crowdfunding platform in the world, Kickstarter.(Greenberg & Gerber,

2014; Kraus, Richter, Brem, Cheng, & Chang, 2016; Song & van Boeschoten, 2015).

According to Mollick (Mollick, 2014), the variables related to the quality of the project tend to have a higher impact on the result of a campaign. One of the challenges in creating a predictive model is to identify the pattern of donor profile for each different type of crowdfunding campaigns. Projects more related with art and culture end up depending a lot on the external social capital (friends and family) while projects on technology, games, and comics tend to be funded by recurrent donors (An, Quercia, & Crowcroft, 2014).

The information regarding donors and project owners behavior are used to profiling and developing a CRM marketing strategy to produce a competitive advantage, generate more profit and increase reputation among donors of the platform (Gordon S. Linoff & Michael J. A. Berry, 2004; Milinski, 2016).

The majority of the predict models that we could find for our research select the variables under two main categories identified in Mollick work (Mollick, 2014). The first group are related with Preparedness (existence of video, spelling check, number of updates in the campaign) and Social Capital (number of the founder's Facebook friends, number of backers in the first week) which are strongly related to the success of a project.

Early contributions have an important aspect in a crowdfunding campaign to reduce uncertainties about the quality and feasibility of the project and the trustworthy barriers that will be tested in the project (Colombo, Franzoni, & Rossi-Lamastra, 2015). Through a predictive model it will be possible better define which projects are more likely to generate extra income for the platform, through the behavior of the first donors, and those who need to make some changes in their marketing campaign to achieve their goals.

From this information it is possible to concentrate efforts on CRM strategy approach to support "best projects" (projects that are more likely to generate more income for the platform) and flag projects that require special attention for the project owners.

2. Literature review

Crowdfunding has its origins in the concepts of microfinance and crowdsourcing but has become a unique category of raising money for projects through the exponential growth of financial transactions made by users on the internet. We can define as efforts of individuals or groups to fund their ventures by drawing on small contributions from a relatively large number of individuals without standard financial intermediaries.

Crowdfunding is different from the traditional financing of new ventures in important aspects. Relatively small contributions of many individuals over a fixed period and transparency allowing potential donors to see the level of support of a campaign before deciding if they will fund or not a project. According with some researchers this type of social information can influence directly in the final outcome of a campaign. (Kuppuswamy & Bayus, 2018)

In the study *Inferring the Impacts of Social Media on Crowdfunding* (Lu, Xie, Kong, & Yu, 2014) crowdfunding is presented as a mixture of entrepreneurship with social network participation, in which the donors play an unexpected role: investors.

Different from conventional investors in venture capital firms, donors who have support a project are more likely to promote the campaign among their online social networks and this behavior directly influences digital marketing campaigns, putting social media platforms at the center of crowdfunding campaigns.

The most common forms of crowdfunding are: reward-based crowdfunding, donation-based crowdfunding, equity crowdfunding and lending-based crowdfunding.

In rewards-based crowdfunding, backers typically contribute with small amounts of money in exchange for a reward. This reward is often, but not always, the item being produced, such as a watch, an album or a film. In donation-based crowdfunding, donors generally donate small amounts. Donation-based crowdfunding is typically used to raise money for a non-profit or a cause.

Equity-crowdfunding, investors give larger amounts of money. When investors give the money, they don't get a reward, but instead, a small piece of equity in the company itself. As a result, equity crowdfunding is typically used to raise money to fund the launch or growth of a company, not just initiate a creative project or cause. Often, these companies go on to raise money from angel investors or venture capitalists.

Lending crowdfunding is relatively new. Instead of owning a stake in a business, investors' money is matched to a loan for a person or business and repaid over a defined term with interest.

This evolution in the methods and process of raising money for causes, cultural projects and business ventures has created a new paradigm and challenges for finding supporters. Crowdfunding platforms have gained more popularity in recent years. However, many projects fail to achieve their goals and understanding donor behavior is essential to make fundraising successful.

One of the first studies to analyze the phenomenon of crowdfunding platforms came from the researcher Ethan Mollick, in the paper "Dynamics of crowdfunding: An exploratory study". The author investigated the exponential growth of crowdfunding campaigns and was one of the pioneers in researching the dynamics that contribute to the success of a campaign

by analyzing a sample of 48.526 projects extracted from the Kickstarter platform. The result of this study, besides mapping the variables most correlated with the success of a campaign, also paved the way for new studies on donor behavior and the relationship of trust between breeders and supporters.

During the review of the literature two key aspects emerge as a foundation for a successful crowdfunding campaign that were divided between factors related to campaign preparations and factors related during the campaign period such as: the proximity of donors to the campaign, the size of the project creator's social network and the digital marketing strategy to generate word-of-mouth for the project to attract more early donors and feedback for the project owners (Dellarocas, 2003).

This social network influence was presented on Mollick research (Mollick, 2014) with a positive correlation with the numbers of Facebook friends and the amount of money raised for the campaign.

In the paper "Internal social capital and the attraction of early contributions in crowdfunding" (Colombo et al., 2015), the researchers explore the concept of social capital to attract contributions in the very early days of a campaign under conditions of maximum uncertainty and how social connections increase the probability of success.

The authors identify two types of social capital: external and internal. The first one is related with family and friends and the second is related with your peers (other project owners) and possible donors inside the crowdfunding platform. “These interactions entail behavior that is specific to crowdfunding communities and includes giving attention, money, feedback, and visibility to the projects of other members of the community”.

There is a strong relation between collaboration, reciprocity, trust and reputation involving donations for charity and social projects according with Milinski (Milinski, 2016). The author analyzed through social experiments

that reputation is an important asset for a campaign and can be used as a universal currency between different social groups to identify if a specific person is a trustworthy social partner. In the experiments, the person's reputation had a significant effect on the donor's decision either to give money or not.

Researching for predictive models of crowdfunding platform we found similar studies using different predict methods, such as random forest, text mining, logistic regression model, and SVM. We discovered two papers with an interesting approach for these predictions problems. The first paper "Crowdfunding Support Tools: Predicting Success & Failure" (Greenberg, Pardo, Hariharan, & Gerber, 2013) explored the efficacy of using machine learning classifiers to determine whether projects will be successful before they launch in order to create a tool for novice crowdfunders using variables related with Preparedness of a crowdfunding campaign (Mollick, 2014) such as Twitter and Facebook followers, the existence of a campaign video, duration, monetary value of the campaign and number of rewards.

Using only these variables they obtained 67% accuracy, a satisfactory result considering that the available data do not include the number of donors. The second paper "Money Talks: A Predictive Model on Crowdfunding Success Using Project Description" (Du et al., 2015) explores 154,561 project descriptions across 15 funding categories to evaluate if the project will succeed. The study obtained 73% of accuracy.

These early studies show the potential of predictive models for crowdfunding platforms is no longer a trend but a necessity for platforms that need to increase their success rates and expand their markets while offering more personalized services to help campaigns reach your goals.

3. Methodology

For this project we are going to take a data-driven approach, using data mining techniques. According to the authors (Gordon S. Linoff & Michael J. A. Berry, 2004), data mining is an iterative learning process that creates results over time. "Data Mining is an important component of analytic customer relationship management. The goal of analytic customer relationship management is to recreate, to the extent possible, the intimate, learning relationship that a well-run small business enjoys with its customers".

We are going to use SEMMA methodology, developed by SAS Institute which is an anachronism for Sample, Explore, Modify, Model, and Assess ("Data Mining and the Case for Sampling Solving Business Problems Using SAS ® Enterprise Miner ™ Software," n.d.). The objective according to the authors is to "make easier for business analysts to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy". Our objective with this method is to understand the correlations between variables, identify possible outliers, analyzing the patterns in the dataset to help generate the most suitable predict model for our problem.

The methodology implemented is described below

1. Selection: In this phase, the dataset provided by Benfeitoria will be uploaded to SAS Enterprise Miner.

2. Pre-processing: Cleaning and scrubbing data to improve the effectiveness of the predict model. Moreover, the noise and possible outliers are removed. Strategies for dealing with missing data fields are also addressed.

3. Transformation: Transforming data into appropriate forms to perform a prediction.

4. Modeling/Segmentation: testing the most suitable prediction algorithms.

5. Assessment: Comparing, analyzing and selecting the best predictive model.

For the predictive model, we will use decision tree model to analyze the dataset. Other robust algorithm, such as neural networks, can give us more precision in our results, but we opted for this approach because our objective is to understand and identify which variables most influence the outcome of a crowdfunding campaign at the end of seven days and could be easily implemented for the startup through SQL language.

3.1 Project development

For this research we had access to Benfeitoria's database. In total we extracted 1411 observations, at the time we have collected the data, and a subset of 12 variables of 1411 covering all the history of campaigns executed in the platform. In total we had 1011 successful and 400 unsuccessful campaigns. We also define a timeframe of 7 days from the beginning of the campaign for all variables. SAS Enterprise Miner is the selected software to perform the required analysis. Through this tool, it is our aim to disclose the predictive model. The first step was to import the dataset into SAS, analyze the given variables and define their role and type.

Initially, the dataset was explored, and some elements were identified, which needed further exploration. The following nodes were used: StatExplore, Graph Explore, and Multiplot to have a statistical description and visualization of the uploaded dataset. (see Table 1 - Benfeitoria dataset)

Variables	Type	Description
Category	Nominal	Type of categories
P_name	Nominal	Project name
P_description	Nominal	Project description
P_development	Interval	days until the project is ready to start
M_raised	Interval	Amount of money raised
N_ supporters	Interval	Number of supporters
N_tags	Interval	Number of tags for project
N_ updates	Interval	Number of project updates in the platform
N_rewards	Interval	Number of rewards
Duration	Interval	How many days before the deadline
PER_raised	Interval	Percentage of amount of money raised
M_goal	Interval	money objective to be raised
Campaign_result	Binary/Target	1-successful | 0- not successful

Table 1 - Benfeitoria dataset

Analyzing the dataset and the StatExplore outputs enables to extract descriptive statistics per class and interval variables, identify total number of missing values and outliers. Below are shown some preliminary statistics per class and interval variables:

- There are 100 missing values in the variable M_raised which will be replaced by the number zero since no financial value was collected in that time frame;
- The statistical information indicates the existence of outliers in our dataset.

New variables were created for our predictive model. The following table explains the transformed variables used for prediction modeling and the formulas used to create them.

Variable	Description	Formula
Avg_M_supporters	Average donated by supporters	M_raised/N_supporters
Avg_reward	Average of reward	M_raised/N_rewards

Table 2 transformed variables

To identify the existing outliers within the variable's distribution, visualization methods (scatter plot, boxplot, and histogram), statistical data and business logic were considered. At the end of this procedure, 41 observations were excluded.

I used statistical information and graphical distribution of each variable with and without outlier to exemplify these decisions. To illustrate these changes, I present the variable Duration after the removal of the outliers where we can see a more normalized distribution in the image below. The other images can be found in the Appendix (see page 32).

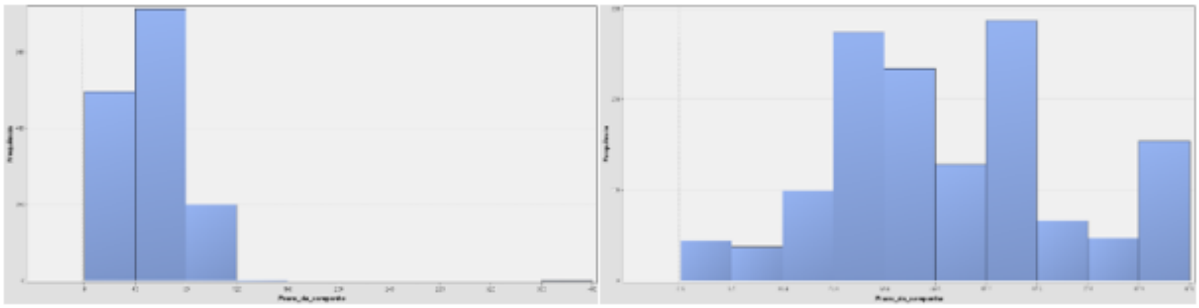

Figure 1 – Duration variable before and after outliers

To provide an understanding of the dynamics of crowdfunding, we'll profile some key variables of our dataset that are related in a successful campaign.

M_Goal: Benfeitoria crowdfunding model is targeted to "all or nothing" campaigns, which means, that pledge money is only collected if the goal is reached. One of the objectives of this type of model is to create an urgency for campaigns, forcing both creators and supporters to strengthen engagement around the project. One of the challenges in choosing a campaign goal is related to the campaign's perception that it is financially feasible to achieve. Unrealistic goals may decline investors from backing the project and setting an appropriate funding goal mediates this risk. (S.-Y. Chen et al., 2015). Currently there are also other flexible goal models that can be tailored for different projects. In our study we will only address all or nothing campaigns. 66.5% of the successful campaigns aimed to raise between R$6196 - R$18567 per project. (1370€ – 4100€)

PER_raised: The percentage of a project's goal that is raised for the campaign. Projects can raise more than their goal. The authors Forbes and Scaefer defined some guidelines for successful crowdfunding: *"A lower funding goal means the percentage funded increases more per pledge. This*

increases the chance of success by attracting more potential backers. Finally, a higher chance of success means the other advantages of crowdfunding such as marketing, validation of product and product feedback can be more effectively exploited" (Forbes & Schaefer, 2017). In our dataset 50,5% of successful campaigns raised between 13% - 30% in the first week.

N_ supporters: The number of supporters the project had in the first 7 days is an important indicator for measuring the success of a project. Early supporters are essential in spread the campaign when there is a high degree of uncertainty serving as a direct indication of the importance of the project to the platform and its network of supporters (Colombo et al., 2015). 67% of successful projects had between 9 and 33 donors in the first week.

Duration: Total days that each project will accept funds. In the paper "Crowdfunding creative ideas: The dynamics of project backers in Kickstarter" (Kuppuswamy & Bayus, 2018) the researcher's study shows that most donors enter at the beginning of the campaign and in the last days. 25% of the projects with less than 28 days were not successful.

Updates: the number of project updates helps to create a greater link with potential donors as well as answering questions and spreading the campaign. In some studies, the frequent number of updates was identified as a positive influencer to increase the rate of successful projects.(Xu et al., 2014). Only 4% of projects update their info.

N_rewards: Number of rewards related with a campaign. The choice of each reward should be made to encourage campaign donors. The proper selection of each reward must consider the target public of the campaign. 76% of successful projects had 7-10 rewards per campaign.

N_Tags: number of tags/categories used to identify project in the plataform. The same project can have more than 1 tag.

Splitting dataset

The first stage of this process consists of partitioning the data to train the different models that will be tested later in the project. The Training Set is the data that is used to train the algorithm. Validation set is used to assess how well the model created in training set is currently performing. Our dataset has 1411 observations, a standard split of 70% training set, and 30% validation set.

After partitioning the data, the next step to choose the best possible variables to be used to train the different models. Our objective is to find the best combination of variables to be used as inputs in a series of different models, which will ultimately return the greatest prediction accuracy.

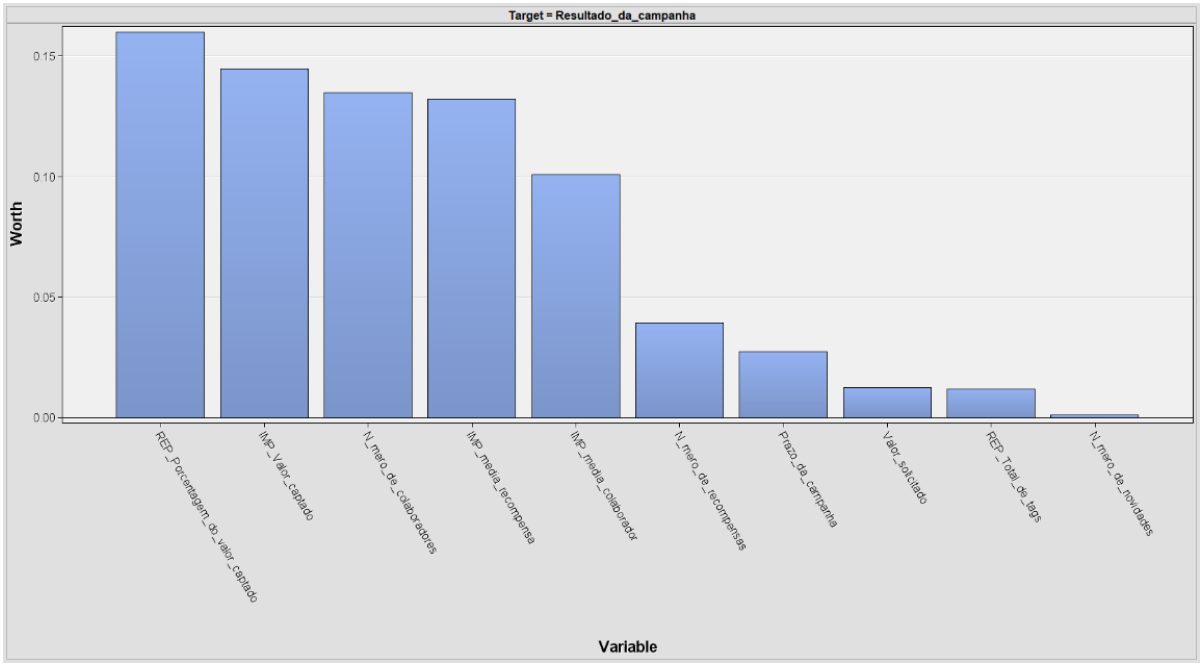

Figure 2 - Variable worth

The results from the Stat Explore explaining the variable worth for predicting our target variable were considered when selecting variables in the prediction models. In our case, we will keep all the variables because Decision Trees is a very robust algorithm for outliers and not affect by multicollinearity issues.

4. Results and discussion

The modelling stage of the project consists on running a set of different predictive models using the above list of variables as inputs. For this dataset was use the Decision Tree classification model, one of the most robust and easy to interpret prediction algorithms.

Decision trees are powerful and popular tools for classification and prediction. In this project, we used 4 decision tree nodes using different combinations of maximum branch and depth. When using the ROC curve to identify the best model, the rule is: the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model.

Our baseline for the prediction model was a priori probability of success in Benfeitoria Projects, which was 72% in this dataset.

The results we achieved through this set of variables are positive. All four models tested (see Table 3 - Results from Decision Tree validation models) showed results much higher than the baseline. After the decision tree pruning, the best model was able to predict the success of a crowdfunding project with 86.6% accuracy, an improvement of 14.6% over the baseline

Model	Decision Tree Branch/Depth	Validation: ROC Index	Validation: Misclassification rate
1	Decision tree 4/4	0.868	0.16748
2	Decision tree 4/3	0.851	0.16990
3	Decision tree 3/4	0.808	0.17961
4	Decision tree 2/4	0.78	0.17233

Table 3 - Results from Decision Tree validation models

As it can be seen from the table above, Decision Tree with 4 branch/4 depth performs better than other Decision Trees Models, considering the Validation ROC Index Value (0.868). From the results presented using different criteria, specifically using the presented results from ROC Index, Valid Misclassification Rate (see Table 4 - Classification Table for the best

model – validation), as a conclusion, it can be estimated that the Decision Tree 1 is the best prediction model[1].

n= 412	Predict 0	Predict 1	
Actual 0	TN =75	FP = 41	116
Actual 1	FN =28	TP = 268	296
	103	309	

Table 4 - Classification Table for the best model – validation

To interpret the results, we will use Decision Tree model 1, to highlight the most important variables of the model. The variable *Per_raised* (percentage of the money raised by the project) is responsible for the first node and split in four branches. We also identify nodes with letters to facilitate their explanation during interpretation (see Figure 3 – Decision tree best model).

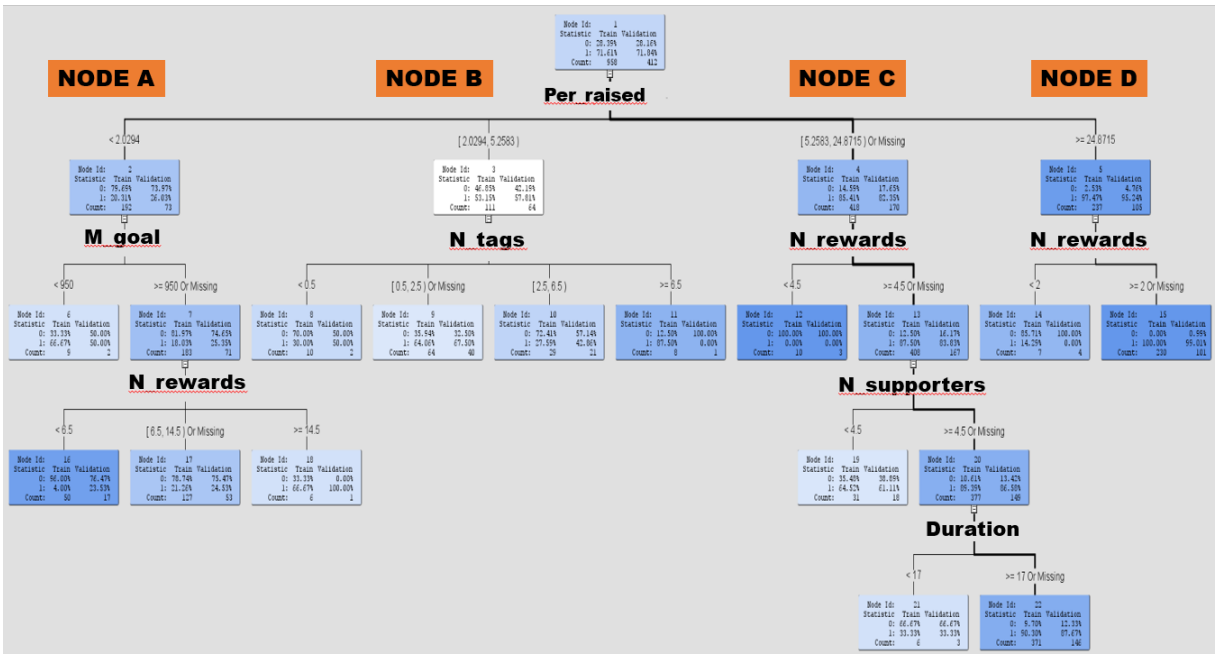

Figure 3 – Decision tree best model

- NODE A - Projects that raised less than [*2.0294%]* of total amount in the first week have 73.97% of chances of failure. This node will go through to more two depth level.
- NODE B - From [*2.0294% – 5.2583%]* of the amount raised the

[1] The decision tree of the best model is in the appendix

chances of being successful go to 57.81%; This node will go through one depth level.

- NODE C - *[5.2583% – 24.8715% or Missing]* - chances of success increase to 82.35%. This node will go through to more three depth level.

- NODE D - Above [*24.8715%]* raised we got 95.24% of probability of having successful project. This node will go through one depth level.

Analyzing the NODE A we had a second and third degree of depth. The second criteria to divide projects was the variable *M_GOAL* (money objective to be raised).

- If a project that raised less than 2% and has a small goal**,** lower than R$950, [<950], we obtained inconclusive result with 50% of chance of failure or success, with only two observations. This result must be careful interpreted because of risk of bias;

- If the goal was equal or higher than R$950 [>=950 or Missing] chances of failure were 74.65%;

The leaf [>=950 or Missing variables] split for the third degree of depth using the variable *N_rewards* (number of rewards).

- Projects with less than [<6.5] rewards have 76.47% of failure;

- Projects with [*(6.5 – 14.5) or Missing]* rewards has 24.53% of success;

- Projects with [>=14.5] rewards have 66.67% of success in train dataset with only 6 observations. Validation dataset presents only one observation with 100% of success. This result must be careful interpret because of risk of bias, since we had only 1 observation in this leaf;

NODE B - From [*2.0294% – 5.2583%]* of the amount raised (variable PER_raised), the second node split in four leafs using the variable *N_Tags* (number of tags).

- A leaf with less than [<0.5] presents 70% of failure in the train dataset. Validation dataset presents an inconclusive result (50%) due to lack of data. Only two observations were present in this leaf. Results must be careful interpret because of risk of bias;
- [(*0.5 – 2.5) or Missing]* present 67.50% probability of success in validation;
- [2.5 - 6.5] present 57.14% of failure in validation dataset;
- [>=6.5] present 87.50% of failure in the train dataset with eight observations. Validation dataset presents an inconclusive result. Only one observation was present. Results must be careful interpret because of risk of bias;

NODE C - From [(5*.2583 – 24.8715) or Missing]* of the amount raised (variable *PER_raised*), the second node split in 2 branches using the variable *N_rewards* (number of rewards).

- Projects with number of rewards lower than [<4.5] presents 100% chances of failure in the train dataset and validation. Results must be careful interpret because of risk of bias due to the lack of observations;
- [>=4.5 or Missing] has 83.83% chances of success;

The leaf Number of rewards [>=4.5 or Missing] split for the third degree of depth using the variable *N_supporters* (*Number of supporters)*

- Project with number of supporter lower than [<4.5] has 61.11% of success;
- Project with more than [>=4.5 or Missing] has 86.58% of succeed

and split by the fourth degree of depth in using variable Duration (How many days before the deadline)

This fourth node divide in two leafs:

- Projects with less than 17 days [<17] has 66.67% of failure. Results must be careful interpret because of risk of bias due to the lack of observations.
- Project with equal or above 17 days [>=17 or missing] has 87.67% of success.

NODE D - From [>=*24.8715]* of the amount raised (variable *PER_raised*), the second criteria to divide projects was the variable N_rewards (number of reward).

- Project with less than two [<2] has 100% of failure in validation. Results must be careful interpret because of risk of bias due to the lack of observations.
- Project with two or more rewards [>=2 or Missing] has 99.01%. This result apparently seems to be right because due the high number of observations in this leaf.

From the result of our predictive model it was possible to identify the variables that most impact our model. The percentage that each project collects in the first week has a strong impact on the result of the campaign. Projects with small goals are more likely to attract more donors and exceed the initial goal. Another variable that stood out in our model is related to the reward number. Projects with more rewards can reach different types of donors.

4.1. Practical implications

During the study we were able to observe the impact of variables related to the social capital of the project creators. In the future, the platform can use

more variables related to measuring social network sharing, donor engagement, and communication reach of each project in social networks to help improve the predictive models.

Project owners and donors can keep track of a crowdfunding campaign on social media and obtain useful insights. The platform can analyze social media data to predict the outcomes and help project owners marketing strategies, while project donors can also make use of this information to identify promising projects that are more likely to be successful.

By analyzing the results of our predictive model, it is possible to establish a CRM strategy for the platform as for the project owners, based on some assumptions identified in the study that directly impact the positive or negative result of each crowdfunding campaign.

For the platform, identifying projects with a high degree of success is essential for the business model. Structuring an automated CRM strategy to alert projects about their likelihood of success or failure allows us to prevent most common mistakes of projects and develops educational materials to empower creators.

Developing a marketing plan for a successful campaign should focus its efforts in the first few days to generate visibility and engagement from the first donors. The project creators network is critical to generate awareness, reach early donors and building trust.

I also recommend using the predictive model to create an insights tool for the platform audience to generate more qualified leads and educate the public about the needs of engaging their target audience during the first week before even starting the crowdfunding campaign.

5. Conclusions and considerations

We propose an exploratory analysis of the possibilities of developing predictive models on Brazilian crowdfunding platforms to increase the

efficiency of the campaigns using a data driven approach. Although the market has become more mature in the last 7 years in Brazil and the crowdfunding models have become popular, the use of machine learning in this segment is in an early stage.

This study is based on data collected from a Benfeitoria's crowdfunding platform. Our approach is useful for describing and predict on this platform. From this method, we proposed a decision tree technique to predict the results (success or failure) of a crowdfunding campaign after 7 days on the platform focus only on reward and donation models for social and cultural projects. We collect a limited number of observations and variables and for future studies should try to extract features related with social media (variables) by using scrapping data techniques. Using a larger database our predictor variables might be different.

The adoption of other machine learning algorithms, such as text mining can also lead to interesting future research to predict success campaigns analyzing their description and title (Yuan, Lau, & Xu, 2016). Qualitative studies can be applied to examine the expectations and perceptions from both donors and project owners. (Kraus et al., 2016)

One of the challenges found in this project was directly related to the amount of data. Since we are working with a startup we receive a lot of unstructured data that we could not use and where we tried to apply the GIGO (Garbage In, Garbage out) concept to minimize the bias of the data.

The importance of project owners' social capital in initial donor mobilization is extremely necessary to engage early donors. Although it is not possible to directly measure in this study the direct impact on the campaigns, we were able to analyze their influence by the profile of the projects (low financial goals and focused on donations for social and cultural projects).

Unlike a company that seeks to test and validate new product ideas, where interest in the product may outweigh the direct relationship of the project

owner, a donation campaign relies exclusively on the previous reputation of the project creator to mobilize his network to collaborate with the project.

The importance of this initial network established by the project owner to achieve its outcome is linked to the emotional aspects of the initial donors (usually family and friends) with the project. They act as social validators to motivate other donors to collaborate, contribute, and disseminate information by expanding the trust deposited by early donors for secondary connections, as demonstrated in the paper *Internal social capital and the attraction of early contributions in crowdfunding* (Colombo et al., 2015).

6. Bibliography

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). *A Practical Guide to Data Mining for Business and Industry*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118763704

An, J., Quercia, D., & Crowcroft, J. (2014). Recommending Investors for Crowdfunding Projects. *Arxiv - Computers & Society*, 261–269. https://doi.org/10.1145/2566486.2568005

Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM). *Business Process Management Journal*, *9*(5), 672–688. https://doi.org/10.1108/14637150310496758

Chen, S.-Y., Chen, C.-N., Chen, Y.-R., Yang, C.-W., Lin, W.-C., & Wei, C.-P. (2015). Will Your Project Get the Green Light? Predicting the Success of Crowdfunding Campaigns. *PACIS 2015 Proceedings*. Retrieved from http://aisel.aisnet.org/pacis2015/79

Colombo, M. G., Franzoni, C., & Rossi-Lamastra, C. (2015). Internal social capital and the attraction of early contributions in crowdfunding. *Entrepreneurship: Theory and Practice*, *39*(1), 75–100. https://doi.org/10.1111/etap.12118

Crowdsourcing.org. (2015). Crowdfunding industry report - market trends, composition and crowdfunding platforms. *Research Report*, (May), 1–30.

Data Mining and the Case for Sampling Solving Business Problems Using SAS ® Enterprise Miner ™ Software. (n.d.). Retrieved from http://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–1424. https://doi.org/10.1287/mnsc.49.10.1407.17308

Du, Q., Qiao, Z., Fan, W., Zhou, M., Zhang, X., & Wang, A. G. (2015). Money Talks : A Predictive Model on Crowdfunding Success Using Project Description. *Proc. ACIS 2015*, *1*, 1–8. Retrieved from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1048&context=amcis2015

Forbes, H., & Schaefer, D. (2017). Guidelines for Successful Crowdfunding. *Procedia CIRP*, *60*, 398–403. https://doi.org/10.1016/j.procir.2017.02.021

Gordon S. Linoff, & Michael J. A. Berry. (2004). *Data Mining techniques-For Marketing,Sales and Customer Relationship Management*. *Portal.Acm.Org*. Retrieved from http://portal.acm.org/citation.cfm?id=983642

Greenberg, M. D., & Gerber, E. M. (2014). Learning to fail. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 581–590). New York, New York, USA: ACM Press. https://doi.org/10.1145/2556288.2557110

Greenberg, M. D., Pardo, B., Hariharan, K., & Gerber, E. (2013). Crowdfunding support tools. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (p. 1815). New York, New York, USA: ACM Press. https://doi.org/10.1145/2468356.2468682

Hui, J., Greenberg, M., & Gerber, E. (2013). Understanding crowdfunding work: implications for support tools. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (p. 889). New York, New York, USA: ACM Press. https://doi.org/10.1145/2468356.2468515

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. In *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition* (pp. 1–25). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118029145

Kraus, S., Richter, C., Brem, A., Cheng, C.-F., & Chang, M.-L. (2016). Strategies for reward-based crowdfunding campaigns. *Journal of Innovation & Knowledge*, *1*(1), 13–23. https://doi.org/10.1016/j.jik.2016.01.010

Kuppuswamy, V., & Bayus, B. L. (2018). Crowdfunding Creative Ideas: The Dynamics of Project Backers. In *The Economics of Crowdfunding* (pp. 151–182). https://doi.org/10.1007/978-3-319-66119-3_8

Lu, C.-T., Xie, S., Kong, X., & Yu, P. S. (2014). Inferring the impacts of social media on crowdfunding. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining - WSDM '14*, 573–582. https://doi.org/10.1145/2556195.2556251

Milinski, M. (2016). Reputation, a universal currency for human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20150100. https://doi.org/10.1098/rstb.2015.0100

Mollick, E. R. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, *29*(1), 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005

Retrato financiamento coletivo Brasil - Catarse e Chorus. (n.d.). Retrieved May 30, 2018, from http://pesquisa.catarse.me/

Song, Y., & van Boeschoten, R. (2015). Success factors for Crowdfunding founders and funders. *Proceedings of the 5th International Conference on Collaborative Innovation Networks COINs15, Tokyo, Japan March 12-14, 2015*. Retrieved from http://arxiv.org/abs/1503.00288

The World Bank. (2013). Crowdfunding's Potential for the Developing World, 5. Retrieved from https://openknowledge.worldbank.org/handle/10986/17626

Xu, A., Yang, X., Rao, H., Fu, W.-T., Huang, S.-W., & Bailey, B. P. (2014). Show me the money! *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, (September), 591–600. https://doi.org/10.1145/2556288.2557045

Yuan, H., Lau, R. Y. K. K., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, *91*, 67–76. https://doi.org/10.1016/j.dss.2016.08.001

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). *A Practical Guide to Data Mining for Business and Industry*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118763704

An, J., Quercia, D., & Crowcroft, J. (2014). Recommending Investors for Crowdfunding Projects. *Arxiv - Computers & Society*, 261–269. https://doi.org/10.1145/2566486.2568005

Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM). *Business Process Management Journal*, *9*(5), 672–688. https://doi.org/10.1108/14637150310496758

Chen, S.-Y., Chen, C.-N., Chen, Y.-R., Yang, C.-W., Lin, W.-C., & Wei, C.-P. (2015). Will Your Project Get the Green Light? Predicting the Success of Crowdfunding Campaigns. *PACIS 2015 Proceedings*. Retrieved from http://aisel.aisnet.org/pacis2015/79

Colombo, M. G., Franzoni, C., & Rossi-Lamastra, C. (2015). Internal social capital and the attraction of early contributions in crowdfunding. *Entrepreneurship: Theory and Practice*, *39*(1), 75–100. https://doi.org/10.1111/etap.12118

Crowdsourcing.org. (2015). Crowdfunding industry report - market trends, composition and crowdfunding platforms. *Research Report*, (May), 1–30.

Data Mining and the Case for Sampling Solving Business Problems Using SAS ® Enterprise Miner ™ Software. (n.d.). Retrieved from http://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–1424. https://doi.org/10.1287/mnsc.49.10.1407.17308

Du, Q., Qiao, Z., Fan, W., Zhou, M., Zhang, X., & Wang, A. G. (2015). Money Talks : A Predictive Model on Crowdfunding Success Using Project Description. *Proc. ACIS 2015*, *1*, 1–8. Retrieved from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1048&context=amcis2015

Forbes, H., & Schaefer, D. (2017). Guidelines for Successful Crowdfunding. *Procedia CIRP*, *60*, 398–403. https://doi.org/10.1016/j.procir.2017.02.021

Gordon S. Linoff, & Michael J. A. Berry. (2004). *Data Mining techniques-For Marketing,Sales and Customer Relationship Management*. *Portal.Acm.Org*. Retrieved from http://portal.acm.org/citation.cfm?id=983642

Greenberg, M. D., & Gerber, E. M. (2014). Learning to fail. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*

- *CHI '14* (pp. 581–590). New York, New York, USA: ACM Press. https://doi.org/10.1145/2556288.2557110

Greenberg, M. D., Pardo, B., Hariharan, K., & Gerber, E. (2013). Crowdfunding support tools. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (p. 1815). New York, New York, USA: ACM Press. https://doi.org/10.1145/2468356.2468682

Hui, J., Greenberg, M., & Gerber, E. (2013). Understanding crowdfunding work: implications for support tools. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (p. 889). New York, New York, USA: ACM Press. https://doi.org/10.1145/2468356.2468515

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. In *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition* (pp. 1–25). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118029145

Kraus, S., Richter, C., Brem, A., Cheng, C.-F., & Chang, M.-L. (2016). Strategies for reward-based crowdfunding campaigns. *Journal of Innovation & Knowledge*, *1*(1), 13–23. https://doi.org/10.1016/j.jik.2016.01.010

Kuppuswamy, V., & Bayus, B. L. (2018). Crowdfunding Creative Ideas: The Dynamics of Project Backers. In *The Economics of Crowdfunding* (pp. 151–182). https://doi.org/10.1007/978-3-319-66119-3_8

Lu, C.-T., Xie, S., Kong, X., & Yu, P. S. (2014). Inferring the impacts of social media on crowdfunding. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining - WSDM '14*, 573–582. https://doi.org/10.1145/2556195.2556251

Milinski, M. (2016). Reputation, a universal currency for human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20150100. https://doi.org/10.1098/rstb.2015.0100

Mollick, E. R. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, *29*(1), 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005

Retrato financiamento coletivo Brasil - Catarse e Chorus. (n.d.). Retrieved May 30, 2018, from http://pesquisa.catarse.me/

Song, Y., & van Boeschoten, R. (2015). Success factors for Crowdfunding founders and funders. *Proceedings of the 5th International Conference on Collaborative Innovation Networks COINs15, Tokyo, Japan March 12-14, 2015*. Retrieved from http://arxiv.org/abs/1503.00288

The World Bank. (2013). Crowdfunding's Potential for the Developing World, 5. Retrieved from https://openknowledge.worldbank.org/handle/10986/17626

Xu, A., Yang, X., Rao, H., Fu, W.-T., Huang, S.-W., & Bailey, B. P. (2014).

Show me the money! *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, (September), 591–600. https://doi.org/10.1145/2556288.2557045

Yuan, H., Lau, R. Y. K. K., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, *91*, 67–76. https://doi.org/10.1016/j.dss.2016.08.001

7. Appendix

Variables	M_raised	Avg_M_supporters	Avg_reward	N_supporters	N_updates	N_rewards	Duration	PER_raised	N_tag	M_goal
M_raised	1	0.44	0.90	0.82	0.05	0.23	-0.04	0.53	0.06	0.32
Avg_M_supporters	0.44	1	0.42	0.10	0.014	0.15	-0.007	0.26	0.08	0.18
Avg_reward	0.90	0.42	1	0.75	0.05	-0.004	-0.08	0.61	0.05	0.21
N_supporters	0.82	0.10	0.75	1	0.04	0.22	-0.08	0.50	0.03	0.27
N_updates	0.05	0.014	0.05	0.04	1	0.005	-0.007	0.03	0.09	-0.006
N_rewards	0.23	0.15	-0.004	0.22	0.005	1	0.15	-0.07	0.07	0.30
Duration	-0.04	-0.007	-0.08	-0.08	-0.007	0.15	1	-0.25	0.02	0.26
PER_raised	0.53	0.26	0.61	0.50	0.035	-0.07	-0.25	1	0.01	-0.23
N_tags	0.06	0.085	0.051	0.03	0.09	0.07	0.02	0.01	1	0.06
M_goal	0.32	0.18	0.21	0.27	-0.006	0.30	0.26	-0.23	0.06	1

Table 5 - Person correlation

Statistics - Decisions Tree/Train	Model 1	Model 2	Model 3	Model 4
Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.79	0.76	0.72	0.89
Kolmogorov-Smirnov Statistic	0.70	0.68	0.65	0.63
Average Squared Error	0.10	0.10	0.11	0.12
Roc Index	0.91	0.90	0.86	0.82
Cumulative Percent Captured Response	13.99	13.99	12.80	12.49
Percent Captured Response	7.00	7.00	6.40	6.25
ion Criterion: Valid: Misclassification Rate	0.17	0.17	0.18	0.17
Total Degrees of Freedom	958.00	958.00	958.00	958.00
Divisor for ASE	1916.00	1916.00	1916.00	1916.00
Gain	39.65	39.65	27.69	24.67
Gini Coefficient	0.82	0.80	0.71	0.64
Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.69	0.67	0.64	0.61
Kolmogorov-Smirnov Probability Cutoff	0.67	0.67	0.67	0.26
Cumulative Lift	1.40	1.40	1.28	1.25
Lift	1.40	1.40	1.28	1.25
Maximum Absolute Error	0.96	0.96	0.96	0.94
Misclassification Rate	0.12	0.13	0.15	0.14
Sum of Frequencies	958.00	958.00	958.00	958.00
Root Average Squared Error	0.31	0.32	0.33	0.34
Cumulative Percent Response	100.00	100.00	91.43	89.27
Percent Response	100.00	100.00	91.43	89.27
Sum of Squared Errors	183.62	191.00	210.73	227.16

Table 6 Statistics for train models

Statistics - Decisions Tree/Validation	Model 1	Model 2	Model 3	Model 4
Kolmogorov-Smirnov Statistic	0.606	0.561	0.561	0.556
Average Squared Error	0.126	0.131	0.136	0.138
Roc Index	0.868	0.851	0.808	0.780
Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.889	0.875	0.914	0.893
Cumulative Percent Captured Response	14.049	14.049	12.659	12.363
Percent Captured Response	7.024	7.024	6.330	6.181
Divisor for VASE	824.000	824.000	824.000	824.000
Gain	37.811	37.811	24.184	21.274
Gini Coefficient	0.737	0.703	0.616	0.559
Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.598	0.559	0.559	0.530
Kolmogorov-Smirnov Probability Cutoff	0.880	0.670	0.670	0.260
Cumulative Lift	1.378	1.378	1.242	1.213
Lift	1.378	1.378	1.242	1.213
Maximum Absolute Error	1.000	1.000	0.960	0.938
Misclassification Rate	0.167	0.170	0.180	0.172
Sum of Frequencies	412.000	412.000	412.000	412.000
Root Average Squared Error	0.355	0.362	0.369	0.372
Cumulative Percent Response	99.010	99.010	89.219	87.129
Percent Response	99.010	99.010	89.219	87.129
Sum of Squared Errors	103.941	107.967	112.062	113.771

Table 7 - Statistics for validation models

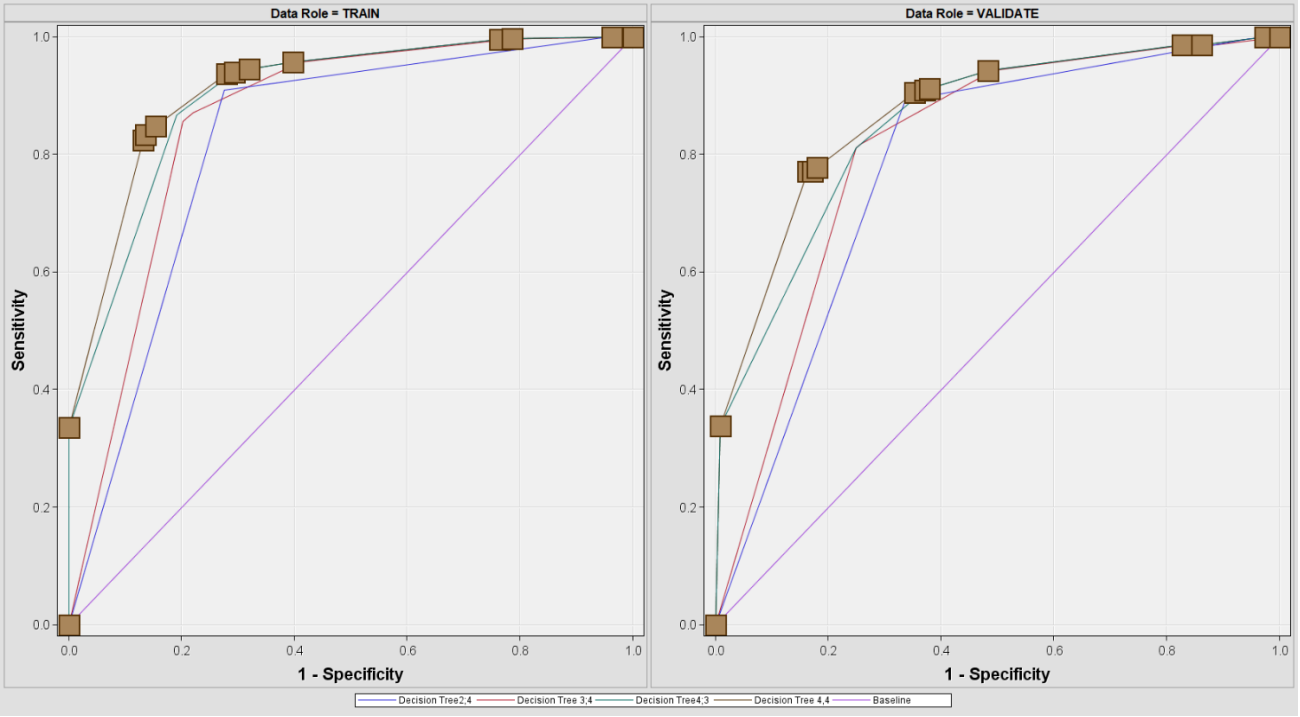

Figure 4 - ROC chart

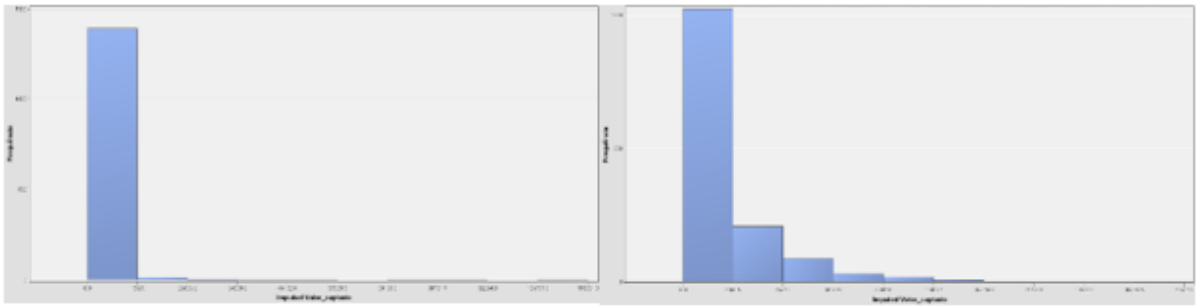

Figure 7 – M_raised variable before and after outliers

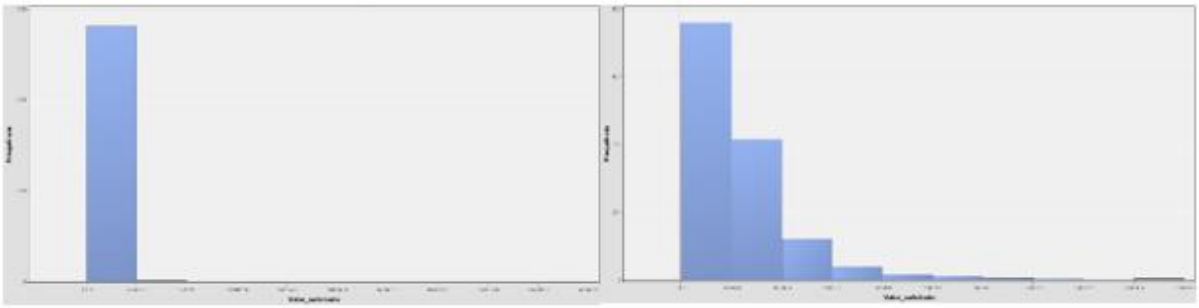

Figure 6 M_goal variable before and after outliers

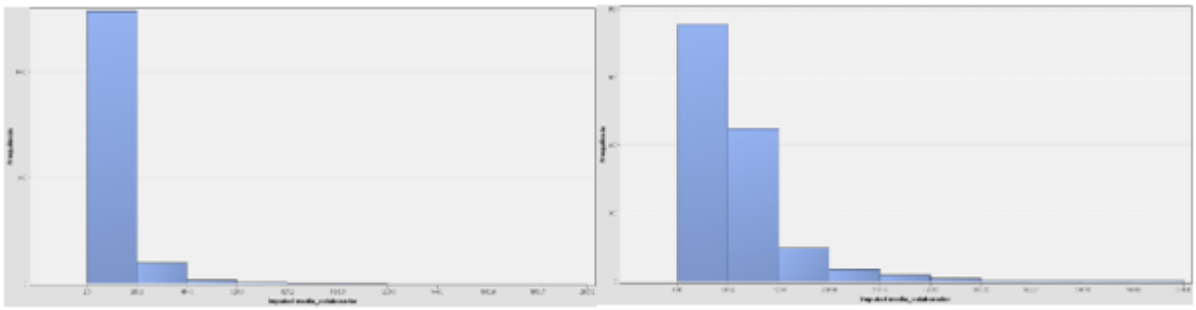

Figure 5 Avg m supporters - variable before and after outliers

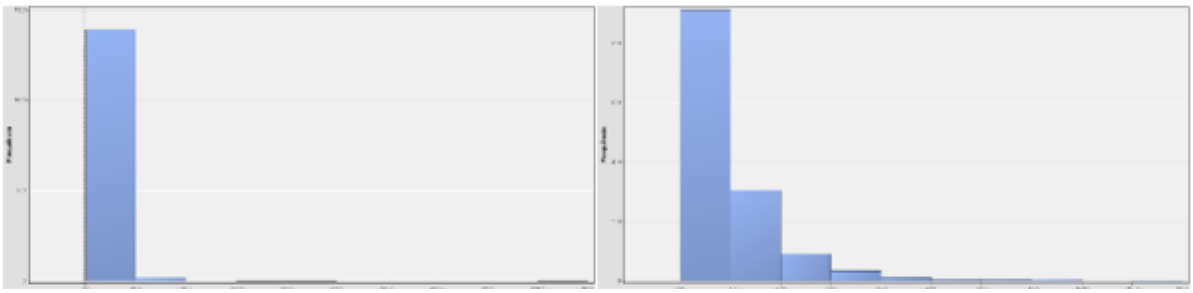

Figure 8 – N_ supporters variable before and after outliers

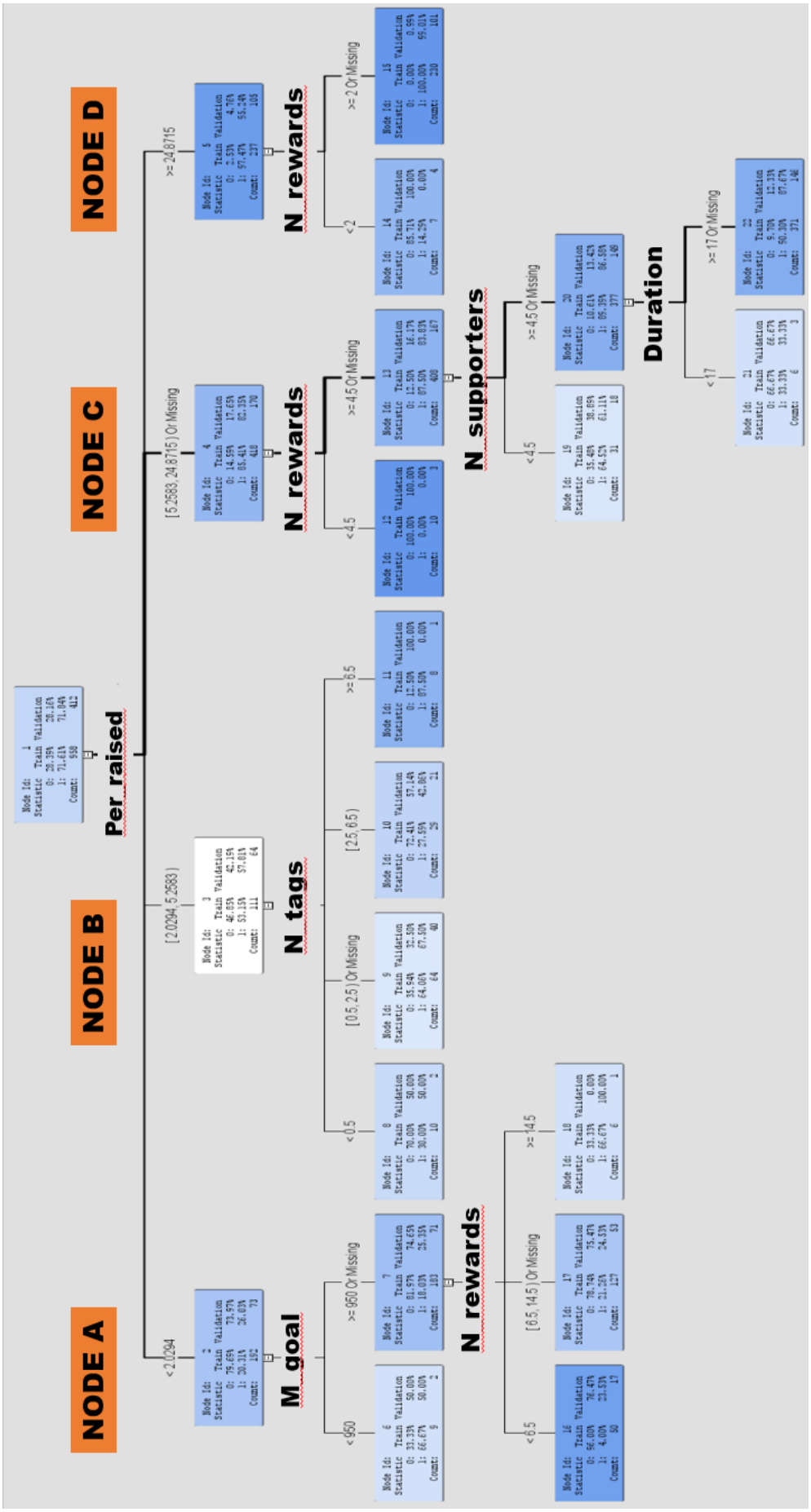

Figure 9 Decision Tree of the best model