Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

How are decisions made in `decidim.barcelona`?

Jaume Solà

Supervisor: Vicenç Gómez

July 2018

Universitat
Pompeu Fabra
Barcelona

Contents

Abstract

Decidim Barcelona is the citizen engagement platform that Barcelona's city hall put in motion on February 2016. One of its first processes has been the development of the "Plà d'Actuació Municipal" (PAM) which has led to the creation of $10,860$ proposals that eventually resulted in $1,467$ actions included in the plan.

In this project we build a statistical model with the aim to predict whether an individual proposal would be accepted or not. The objectives of this work are two-fold: first, we want to gain understanding in the processes that took place during the development of the PAM and second, we want to test to what extent this type of statistical modeling can capture the decision making process in order to better aid future deliberative processes in the platform.

We first analyze the data generated by citizens in the city of Barcelona that participated actively on the platform. After a preliminary statistical analysis of the features that characterize each proposal, we proceed to build a model that is able to predict if a proposal would be accepted or not from that data. We consider the logistic regression model because its computational simplicity as well as its potential interpretability. We are be able to extract conclusions from the parameters and unveil the decision process which resulted the acceptance/rejection of each proposal in the platform. We show that such a model is able to characterize some particularities of the process, but also how this classifier compares to other methods like random forests, and what do the differences we find between them mean.

Keywords: Logistic Regression Classifier; Collective Action; e-Democracy

Chapter 1

Introduction

Decidim Barcelona is the digital platform for citizen participation that the City Council Barcelona started in 2016. The first step of this platform included the development of the 'Municipal Action Program' (Pla d'Actuació Municipal, hereafter PAM), which has produced 10.860 proposals that resulted in 1.467 municipal actions. The general aim of this project is to develop computational and statistical methods to extract knowledge from the data collected in the platform in form of digital user activity. We aspire to build a predictive model able to forecast the acceptance or rejection each proposal while we gain knowledge of how each of the feature's proposal explains the outcome of the model.

The PAM is elaborated at the beginning of each mandate and it establishes the main objectives and the actions of the city hall government. It is, indeed, the principal road map that guides which model we want for our city. Barcelona's city council opened a citizen participation process to build, think, and discuss the policies and priorities of their term, trying to take a step forward in making a more fair and democratic city in a collective way. The participative process for PAM that happened in Barcelona was a hybrid method between meetings and digital spaces. This singularity has impulsed the birth of the platform `"decidim.barcelona"` [1]. Thanks to the platform, a space for citizens participation was created, thus making the process visible, transparent, and traceable.

Motivation & Objectives

In spite of the digitalization and democratization of citizens' involvement in local policies, there is still a long path to be followed to achieve digital democratization. First steps are through portals like "`barcelona.decidim`", thus is important to understand how petition platforms work and what data they contain. There is huge potential on exploring this type of social generated data, where modern statistical techniques can provide insights embedded on it. The approach of the problem raised the following questions:

- Can we gain understanding of the PAM process by looking at the data generated in the platform during the PAM period?
- Can we build a statistical model that is able to predict acceptance/rejection of the proposals just from the observational data from the platform?

The first question has been addressed in several studies using different methodologies [2, 3, 4]. In this work, we mainly focus on the second claim. We show evidence that such a model also helps in gaining understanding of the PAM process.

This project considers the decision process through which the proposals were accepted of denied. The course of action was executed by the government team and consisted of the revision, classification and re-elaboration of the proposals coming from the citizenry (organized and unorganized), conjointly with in-person meetings, and the local government.

The objective of the project is to automatically extract the knowledge of the data in the "`barcelona.decidim`" platform in order to better understand the mechanism of decision-making upon the acceptance/rejection of the submissions. The approach is based on modeling explicitly the process that assigned to each proposal the decision using analytical methods based on data [5]. More precisely, we formalize the challenge of knowledge extraction as a supervised-structured machine-learning problem. In this definition, each proposal has associated a set of descriptive characteristics or

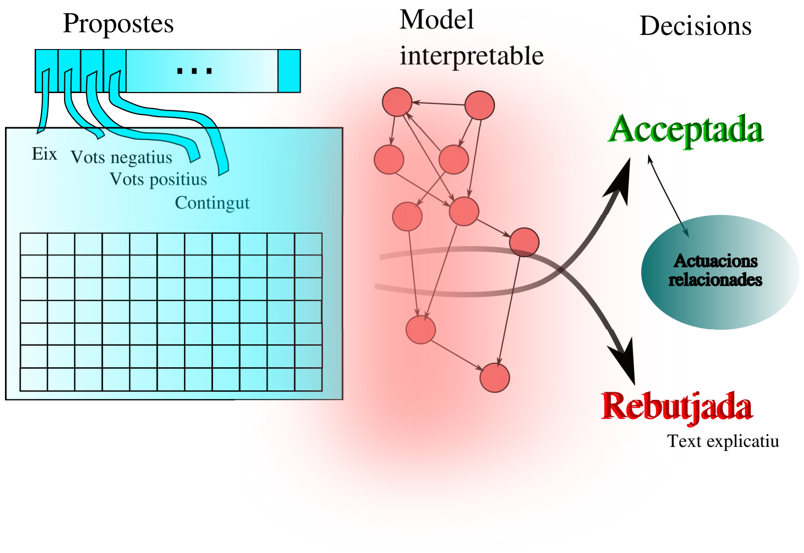


Figure 1: Learning the association between proposals and decisions. The red central part corresponds to the predictive model that we learn from the data observed in the platform `decidim.barcelona`, which are the list of proposal features for each proposal (in blue, left) and whether the proposal was accepted or not (decision, right).

predictors. In general, any information associated with the proposal that could be measured can be used. On the other hand, we have the response variable, which on our case is the decision made by the government to move forward with a proposal or not, and create a action plan to complete the accepted ones. Figure 1 schematically illustrates the approach.

The model that we aim to develop is relevant in the modern society for several reasons. First, generating mathematical models that are able to explain in an objective way which characteristics have more influence on the decision making process of a government team which advocates for transparency, and second, addresses one of the main challenges of giving a good use of the large amounts of data that our society generates.

Structure of the Report

First, we will introduce the dataset we will work with and perform a global statistical analysis of all features selected to to gain knowledge about the PAM process. Later on, we will move into applying this knowledge into an interpretable machine learning algorithm, and how we proceeded methodologically. Further on results will be presented and discussed, analyzing the performance of the methods, the information that we are able to retrieve and how insightful conclusions we can draw.Finally, a discussion about the results and what can be improved as well as some further work ideas.

Chapter 2

Methods

Before starting to build a model that is able to predict whether a proposal is going to be accepted or not, it is necessary to explore the dataset to better understand the structure of the data, and gain the first insights. Afterwards, predictive modeling will be carried out and the particularities of the models developed will be discussed. Lastly, the predicted outcomes will be assessed to draw conclusions.

We begin by analyzing the numerical features in subsection 2.1.1. Then we continue by evaluating the categorical features on subsection 2.1.2. After this preliminary analysis, we detail in section 2.2 how we built the model and explained in detail each of the challenges we faced in subsections 2.2.1 and 2.2.2. Lastly we have an overview of the evaluation methods used in the classifier in the last subsection 2.2.3.

The dataset we will work with was obtained after gathering the data generated by citizens in the city of Barcelona that participated actively on the platform during a period of approximately three months (from the 31st of January 2016 to the 8th of April of the same year). This dataset is open and can be found on-line on the PAM's website [1].

Statistical Analysis

Our dataset is composed of a total of $n = 10,860$ proposals. Each one of them has a total of 46 features of distinct kinds: numerical and categorical variables, other variables that link proposals, and also text, which can be very informative but harder to work with. From those features we focused on eleven, which are presented on Table 1 and will be described briefly in the following sections.

As mentioned earlier, in this work model interpretability is a key concern -the absence of the ability to introspect on a model's decision making is a serious legal mandate in public policy-, the selection of features was driven by the explainability of the given variables. An algorithm that has great accuracy but is impossible to extract conclusions from is not going to be useful, we are not using this for decision making of legal mandates in public policies.

Table 1: List of features considered to explain the resolution of a proposal, with a brief description and data type

Feature	**Description**	**Type**
u	Total number of users	Numeric
c	Total number of comments	Numeric
cp	Total number of positive comments	Numeric
cm	Total number of neutral comments	Numeric
cn	Total number of negative comments	Numeric
v	Total number of votes/supports	Numeric
t	Time active (in days) in the platform	Numeric
s	Source of the proposal	Categoric
d	District of the proposal	Categoric
k	Category of the proposal	Categoric
sk	Subcategory of the proposal	Categoric

Numerical features

The first group of numerical features describe the amount of activity -where activity is defined as the total number of interactions during the time frame of the process-, that a proposal had on the platform. Under this category, and for a given proposal i, $i = 1, \dots, n$, we have the following variables: the total number of *users* u_i that were active on the thread of the proposal (either commenting or voting), how many comments a proposal received c_i classified into *positive* cp_i, *neutral* cm_i, and *negative* comments cn_i. Besides, the number of *votes* v_i that a proposal obtained, and, last but not least, the *time* t_i in the platform as a numerical variable. t_i is defined as the lifespan of a proposal: we check how many days a proposal was active from formulation until the end of user interaction.

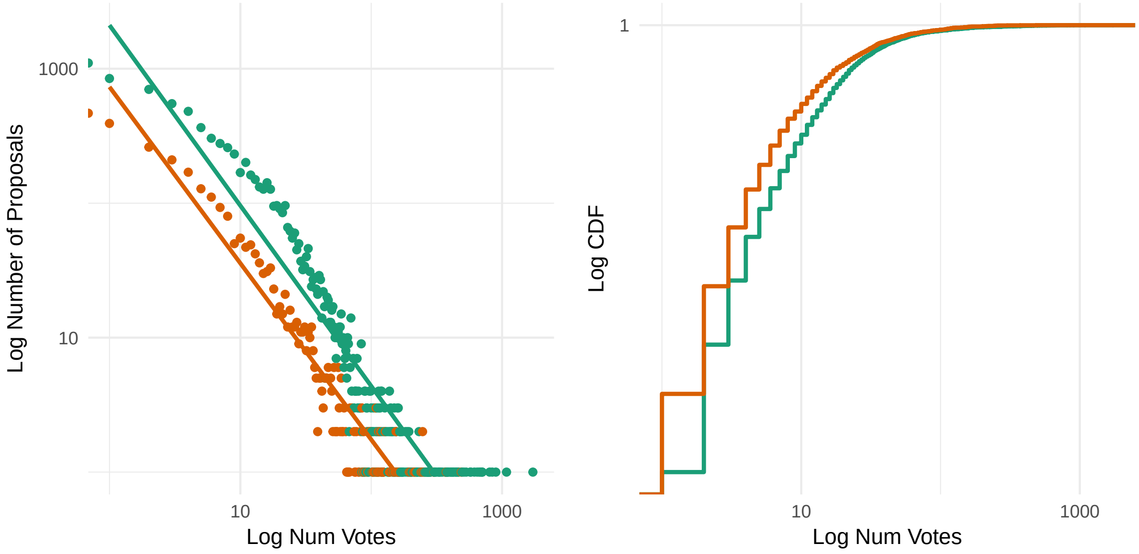
Table 2: Means of each feature per proposal status

Status	u	c	cp	cm	cn	v	t
Accepted	1,01	1,76	0,50	1,07	0,05	16,38	46,02
Rejected	0,99	1,41	0,42	0,55	0,13	11,68	40,64

Table 2 shows the mean values corresponding to each of the features, averaged for the whole dataset and split by proposal status. When we compare the mean of those distributions for accepted proposals against the mean for rejected ones, we observe a similar average number of interactions with users, a larger mean for comments overall (with more positive and neutral, but less negative), a higher average number of votes, and an longer life in the platform. From those differences on those distributions, we see how accepted and rejected proposals have distinct patterns an thus those variables can be used for building the classifier.

Figure 2 shows the probability distribution of the number of votes per proposal in log-log scale. As expected, these distributions are very skewed (lots of proposals with little activity and few popular proposals with huge interactions), which is typically the case for human behavior data [6]. In this particular case, the feature values span up to three orders of magnitude.

Building a predictive model from data following such a distribution, which differs substantially form other probability distributions such as the Normal or exponential, may pose a problem. In subsection 2.2.3 we will describe the preprocessing steps to deal with them.



(a) Probability Density Function (b) Cumulative Distribution Function

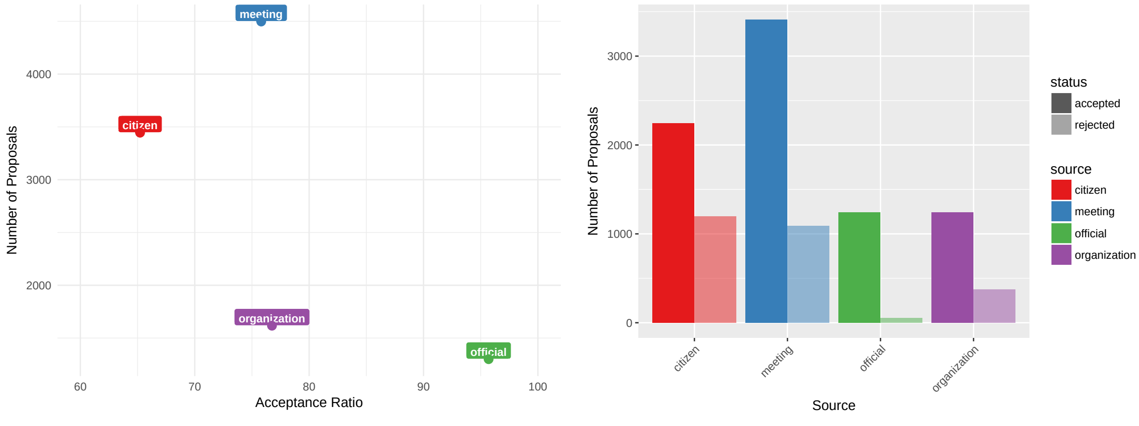
Figure 2: Number of votes on a proposal (green stands for the accepted ones and orange for the rejected): (a) on top of the density functions of votes we have fitted linearly each power law distribution. Both fits have the same slope, but the accepted proposals one is shifted to the right.

Categorical features

The second subset of features to analyze refers to the categorical variables. These contain qualitative information, which ease the understanding of the proposals. These features are *source* s_i, *district* d_i, *category* k_i, and *subcategory* sk_i.

The *source* of the proposal s_i determines which was the origin of the proposal. It takes four values: *official* which corresponds to Barcelona's city council, *citizen* corresponding to a single individual, *organization* referring to an entity or group of people that represent a group of citizens, and finally the *meeting* source, which corresponds to proposals originated *in person* meetings organized by the city council. The proposals under this source are the the translation of the outcome of those reunions, where all citizens could attend and discuss some ideas.

If we observe the acceptance ratio of a proposal based on the source (Figure 3), we clearly see how *official* proposals originated by the city council have a high chance of being accepted compared with the other ones, e.g., citizen proposals. The platform initially contained more than $1,000$ petitions from the electoral program of the ruling party. This is highly influenced by the fact the initial proposals were built and thought following the strategic lines that the government would want to follow in the mandate. Therefore, they are proposals that the government already wanted to implement, and moreover, they where already viable proposals and within the scope of action of the city hall.



(a) (b)

Figure 3: Accpetance / rejection of the proposals according to their origin. (a) Relation between the acceptance ratio and the number of proposals for each different source. (b) Absolute number of accepted/rejected proposals for each different source.

Besides the *source*, we have the *district* d_i category. Whitin this specific feature, proposals fall into two main categories, the ones where the scope is the whole city (are not bounded to an actual neighborhood) and the ones that are more tight to a specific local area. From what we know, the evaluation of the proposals was different depending on the scope. The ones related to just one district were evaluated by each district council independently of the category of the proposal while the ones where the scope was the city as a whole were split by each category and subcategory for evaluation. Due to this reason, we will analyze them separately. We will consider two types of models: a general one that comprises all the proposals and one for each specific district.

Figure 4 shows the different acceptance rates per borough. We can see areas like *Eixample*, *Horta - Guinardó*, and *Nou Barris* have really high acceptance ratios while proposals coming from *Sarrià - Sant Gervasi* or *Les Corts* are less likely to be accepted. As each district has its own administration, the differences we observe could be caused by a political misalignment and between governments with opposing priorities. Those results diversity are also interesting for the prediction.

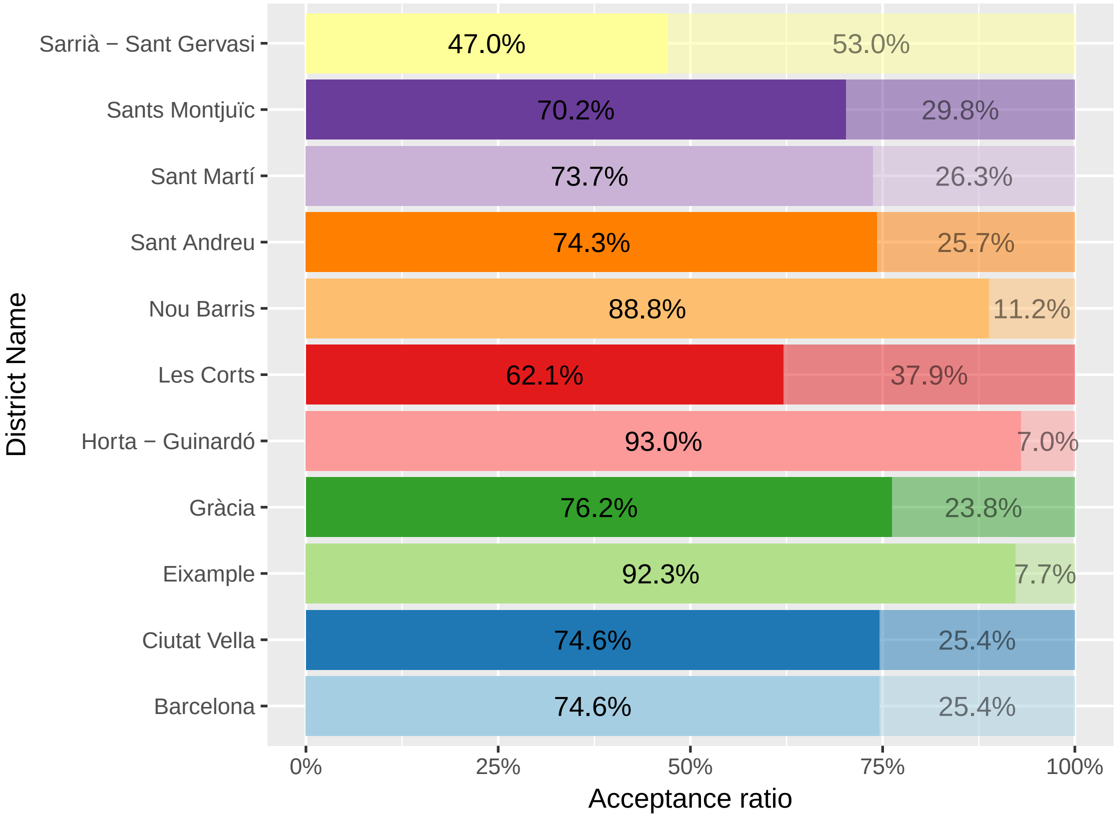


Figure 4: Acceptance ratio per district

The last couple sets of features, *category* k_i and *subcategory* sk_i, classify each proposal in its field of action, which are related to the five principal axis and multiple strategic lines of actuation of the government. Those axis are *Bon Viure*, which aims to improve quality of life, *Transició Ecològocia*, that tries to build a more sustainable city model, *Economia Plural*, that takes care of the economical diversity, *Bon Govern*, which commits to transparency and best governing practices, and *Justícia Global*, which has an overall point of view and tries to fit within the international community. A more detailed overview can be found on the process' website [1].

Figure 5 shows the acceptance of each of the main categories against the volume of proposals. Even the range of acceptance is quite small, ranging from 65% in *Justícia Global* to an almost 80% in *Economia Plural.* On the other hand, what we see as not being that balanced is the volume of proposals per category. This reflects both government and citizens priorities in a city in topics, being *Bon Viure* the category with the biggest share of proposals, followed closely by *Transició Ecològocia.* Those two capture the proposals with a direct impact on the city and its inhabitants. The rest of the categories are less represented since the implications are not so clear on the peoples day to day life, making them less popular.

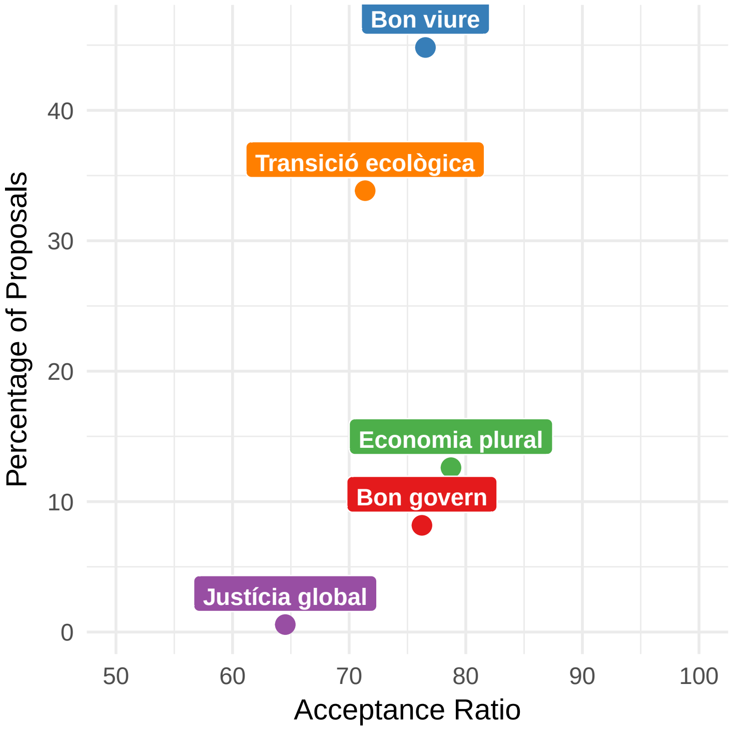
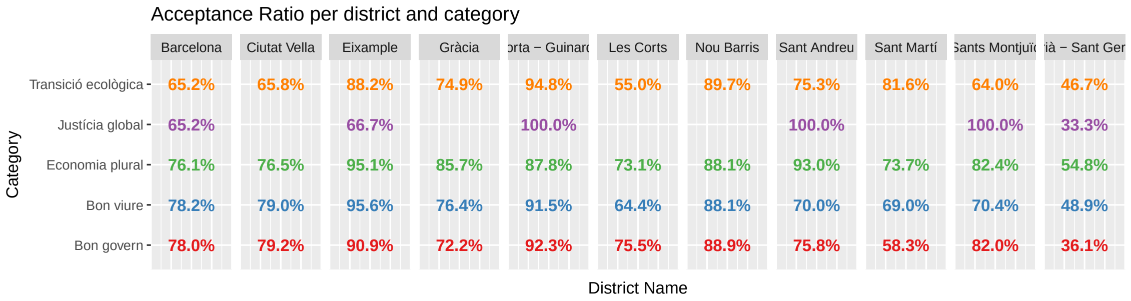


Figure 5: Acceptance ratio per category

Lastly, figure 6 shows the acceptance ratios - which take a wider range of values (from 33% to 100%)-, are not only by category but also per district, and illustrates how priorities differ significantly between neighborhoods. A clear example would be comparing *Sants - Montjuic* and *Sant Martí* districts: on the first on,e the lowest acceptance ratio score is *Transició Ecològocia* with a 64,0%, whereas on *Sant Martí* district is the category with the highest score, an 81,6%. Now in *Sant Martí* we observe how *Bon Govern* has the worst ratio (58,3%) while on *Sants - Montjuic* is one of the highest (82,0%).


Acceptance Ratio per district and category

Category	Barcelona	Ciutat Vella	Eixample	Gràcia	orta – Guinar	Les Corts	Nou Barris	Sant Andreu	Sant Martí	ants Montjuï	ià – Sant Ger
Transició ecològica	65.2%	65.8%	88.2%	74.9%	94.8%	55.0%	89.7%	75.3%	81.6%	64.0%	46.7%
Justícia global	65.2%		66.7%		100.0%			100.0%		100.0%	33.3%
Economia plural	76.1%	76.5%	95.1%	85.7%	87.8%	73.1%	88.1%	93.0%	73.7%	82.4%	54.8%
Bon viure	78.2%	79.0%	95.6%	76.4%	91.5%	64.4%	88.1%	70.0%	69.0%	70.4%	48.9%
Bon govern	78.0%	79.2%	90.9%	72.2%	92.3%	75.5%	88.9%	75.8%	58.3%	82.0%	36.1%

Figure 6: Acceptance ratio per category and district

Building a classifier

After the previous global analysis, we introduce the methodology followed to estimate a statistical model. We have a very particular dataset where we find many typical complications that need to be adjusted for the implementation.

To infer the acceptance or rejection of a proposal i, we have used a simple logistic regression model where the positive class ($y_i = +1$) represents the acceptance of the proposal and the negative class ($y_i = -1$) the rejection. The logistic regression model is widely used; it is the simplest model for classification, since feature values are combined linearly, it can be estimated efficiently and it provides some interpretability as done in similar problems [6].

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as *logit regression*, *maximum-entropy classification* or the *log-linear classifier* [7]. In this model, the probabilities describing the possible outcomes of a single trial are modeled using the logistic function

$$P(y_i = +1) = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j}\right)}, \tag{2.1}$$

where the sum is over the p feature values $x_{i,j}$ of proposal i, and $\beta = \{\beta_0, \beta_1, \ldots, \beta_p\}$

are the coefficients that needed to be estimated from the data. Notice that positive values of β_l point out that x_l has a positive effect on the acceptance of the proposal and, therefore, large values of that variable will make that event more likely.

To estimate the parameters β, we solve the following penalized least squares optimization problem:

$$\min_{\beta,C}\left\{\frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n}\log\left(1+\exp\left(-\beta_0-\sum_{j=1}^{p}\beta_j x_{i,j}\right)\right)\right\}, \tag{2.2}$$

where C represents the inverse regularization strength, i.e., larger values of C result in less regularization and vice-versa.

As the interpretability is one of the main purposes of using this algorithm, we also define a method to quantify how each feature contributes to the model. Importance is measured as the normalized percentage of the t-statistic for each model parameter. Logistic regression does not know how to deal with categorical features, therefore we must transforms each categorical feature with m possible values into m binary features, with only one active. This is known as one-hot encoding.

We will also compare the results obtained from the logistic regression against other classification algorithms to test if our results depend on the actual method implemented. The algorithm we will use specifically is Random Forests [8]. Random forests are an ensemble learning method used in classification, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. Random decision forests correct for decision trees' habit of overfitting to their training set. On figure 7 we show a schema of how the algorithm works, were each tree of the forest gives an output of the response variable and then the final class is defined by majority voting. unlike logistic regression, random forests are able to work with categorical variables, then we will not need to one-hot encode the variables.

After describing the model and objective, we now present some of the challenges that we need to address for this particular dataset of proposals.

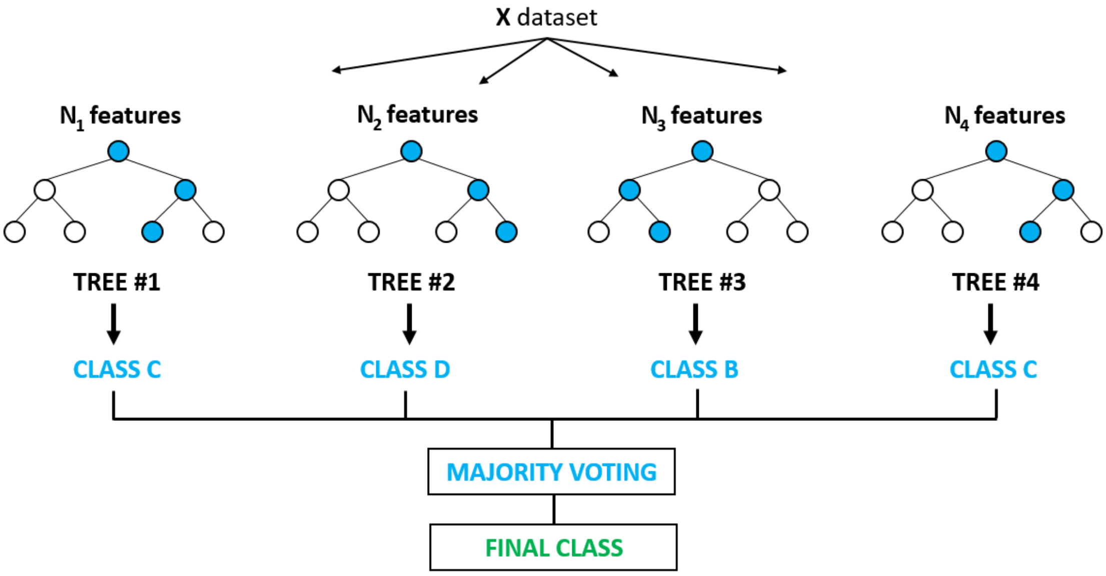


Figure 7: Random Forest Simplification

Skewed feature distributions

The first problem we face is the distribution of our numeric features. In logistic regression classifiers, some kind of normalization is typically applied through transformations. This is specially important when variables have highly skewed distributions as we just seen before. In our case, all numeric features except t_i are heavy-tailed distributed. Hence, we must log-transformed them before using them in our models.

Figure 8 shows the histograms of the feature values after transforming their values using the logarithm: the first row of charts corresponds to the comment related features (total, positive, neutral, and negative respectively) and the second row shows plots about the total number of users, number of votes, and days active in the platform. We can see that the the distributions of the main variables used in our models are "better behaved" after the transformation. Unlike before, the values are more concentrated and do not span more than two orders of magnitude.

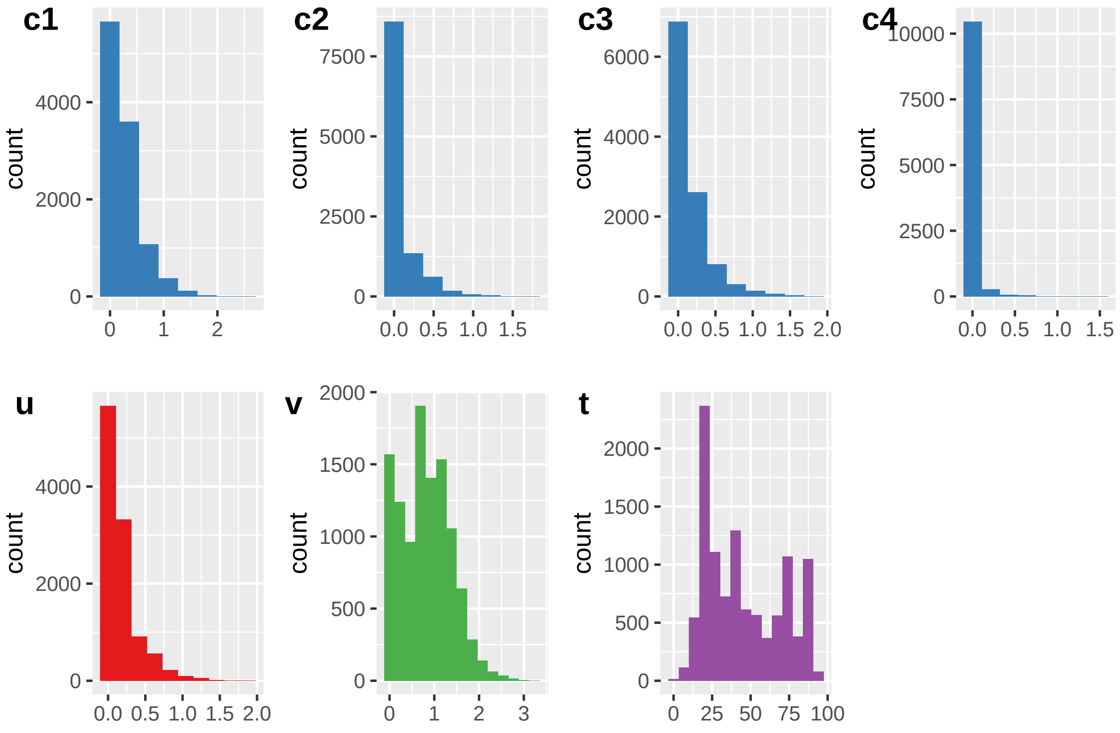


Figure 8: Histograms of numerical log-transformed variables

Highly correlated features

The next problematic aspect that we have considered is the existence of highly correlated features. In these cases, one predictor variable may carry redundant information and the weight of this feature may be *spread* among the other highly correlated features is that, even if all the variables are relevant to the predictive model. Correlated variables can diminish the predicting power and interpretability of a model and we need to understand the explanatory power between them before starting to construct a statistical significant model. The selection of features is addressed using the correlation matrix. As shown in Figure 9, some of the variables have moderate correlation coefficients, which is expected since they all inform us of the activity in the platform.

The largest correlations are found between the total number of users u_i, the total number of comments c_i and the total number of neutral comments. We choose to exclude u_i and c_i from the model.

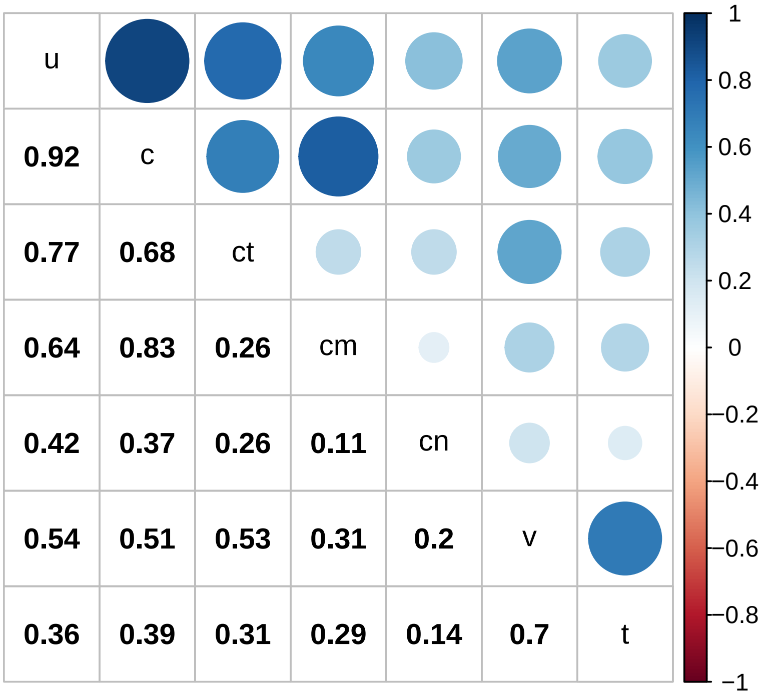
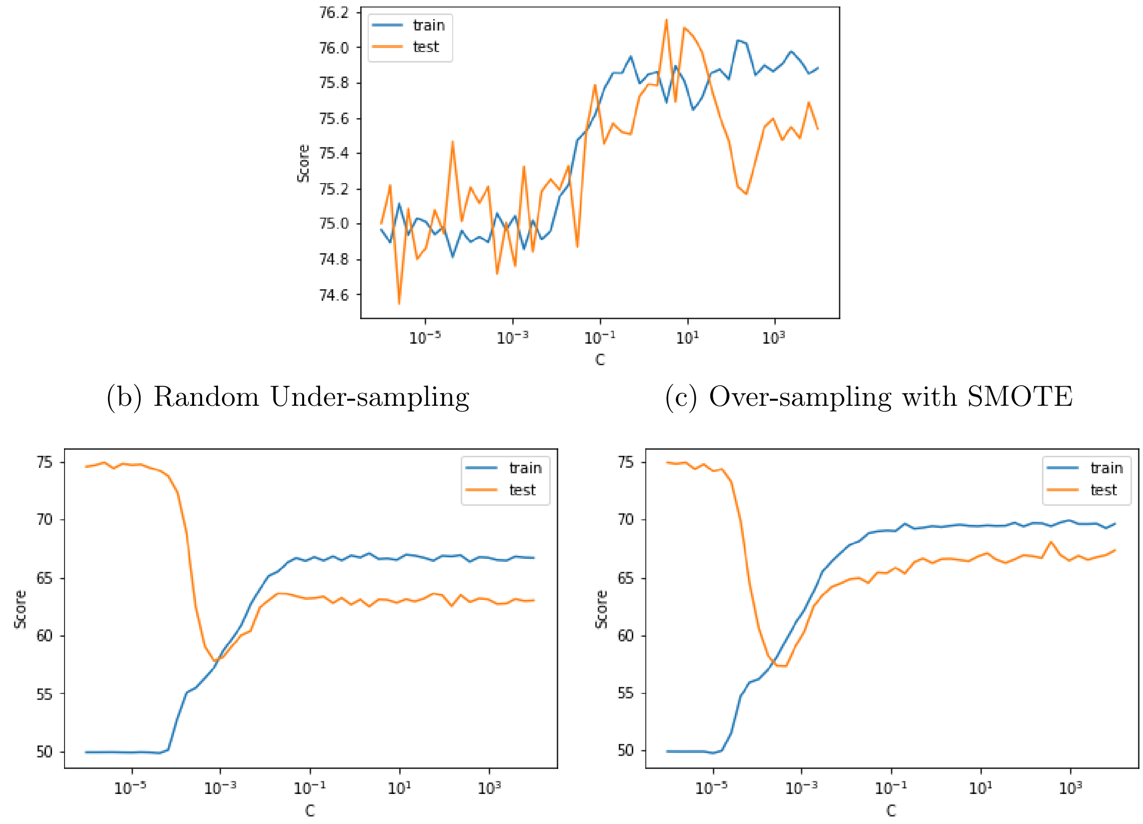


Figure 9: Correlation matrix between features

Evaluation of the classifier - Unbalancedness

After these adjustments, we need to define a way to evaluate the model performance in a robust way in order to analyze its results, and, to do so we need to split our data into training and testing sets. One well-known widely-used method is $k-$fold cross-validation [9]. In $k-$fold cross-validation, the whole dataset is randomly fractioned into k equal sized subsamples. Of the k partitions, a single one is retained as the test data for validating the model, and the remaining $k-1$ partitions are used as training data. The cross-validation process is then repeated k times, with each of the partitions used exactly once as the test data. The results are then averaged to produce a single estimation. In our model, as in the majority of the literature, we used $k = 10$.

The dataset that we are dealing with is unbalanced when it comes to the response variable, where 75% of the proposals were accepted and only 25% were rejected. This may cause problems when training the model so we need to do some sampling in order to equilibrate it. We have two options: we can either reduce the number of samples that represent the majority class or generate synthetic cases for the minority one. We compared both methods, using random under-sampling on one hand, and SMOTE [10] on the other and compared their performances.



(a) No Re-sampling

(b) Random Under-sampling

(c) Over-sampling with SMOTE

Figure 10: Comparison of the re-sampling techniques

On Figure 10 we observe a comparison between the two regularization methods mentioned earlier against not using any re-sampling technique. We plotted the train and test scores (where score is the accuracy of the logistic model fit) against the inverse of regularization strength C. On the left, the random under-sampling method achieves a 67% accuracy on the train subset and 66% on the test one. On the right, the SMOTE method shows a slightly better predicting potential, having a 69% on train set and 67% on test. Here the difference we observe is quite small, $\sim$ 2%, but for some models we want to build where the amount of data we dispose is not large, therefore we select SMOTE oversampling method as a solution to the unbalance of the data.

A common measure the performance of a model is the Area Under the ROC curve (AUC). The ROC curve [11], is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate vs. the false positive rate, at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate. ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger AUC is usually better.

Being that said, ROC curves can present an optimistic view of the classifier performance if the distribution of the response variable is not well balanced, which is our case. For those kind of distributions, Precision-Recall curves [12]. Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. They are therefore are an alternative to ROC because they evaluate the percentage of true positives among positive predictions.

To sum up, we will apply the over-sampling technique just on the train subset to not jeopardize the predictions with synthetic results and we will use the Area Under the Precision-Recall Curve (AUC) as a performance indicator metric.

Chapter 3

Results

Although the previous results show the potential predictive power of the features generated in the platform, to get a complete picture of proposal resolution we build a predictive model based on all the variables introduced in the previous chapters. As mentioned earlier, each district is also building their own action plans with a similar structure but with a territorial point of view, objectives and actuations of their own may differ based on priorities. Therefore we define two different prediction models depending on the scope of the proposal. In the first one, *Model* 1, we used the features selected on the previous chapter and all the proposals, regardless of their origin, to understand if there is a common pattern across all proposals. The amount of data in this case is large enough, and we expect less problems due to the lack of data than in the second model. For the district case we create a second subset of ten models with the same structure, *Model* 2, where we use the *district* feature to group the data and build 10 different models, one for each district.

As mentioned in the previous chapter, we used the simple logistic regression model (LogR) because we can interpret the results with ease. The performance of the models can be seen in Table 3.

Model 1 achieves values around 0.65 for its accuracy, sensitivity, and specificity, showing a good balance of our model detecting both classes but not a great accuracy. The multiple models that we have per district show very heterogeneous results,

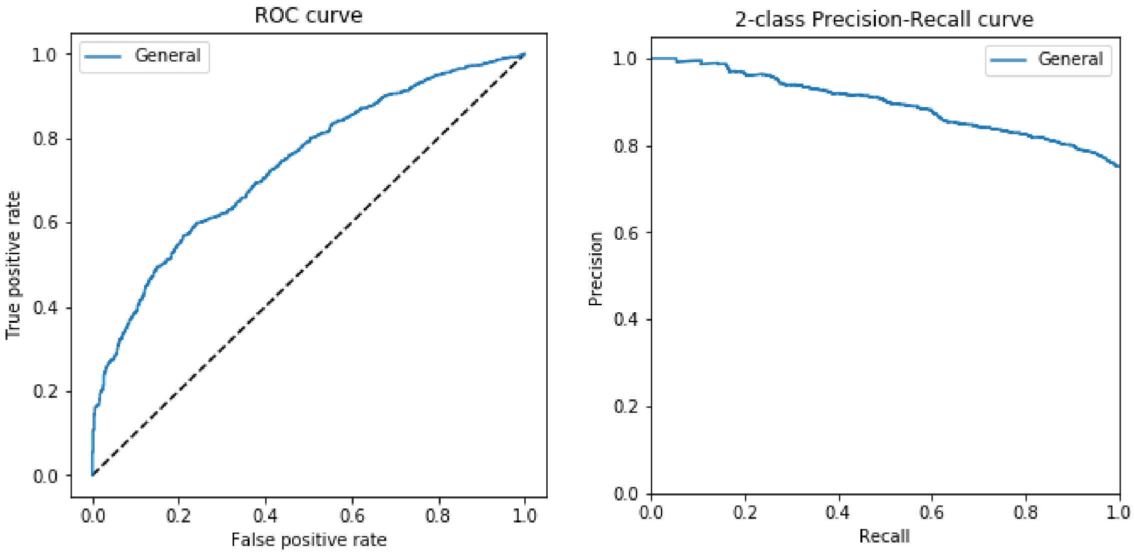
Table 3: Model Performance Comparison

Model	Accuracy	Sensitivity	Specificity	AUC
All Data	0.66	0.68	0.61	0.73
Barcelona	0.58	0.62	0.48	0.61
Ciutat Vella	0.56	0.52	0.65	0.68
Eixample	0.79	0.83	0.33	0.73
Gràcia	0.48	0.48	0.48	0.52
Horta - Guinardó	0.64	0.63	0.71	0.63
Les Corts	0.58	0.48	0.73	0.63
Nou Barris	0.59	0.61	0.50	0.55
Sant Andreu	0.63	0.59	0.75	0.66
Sant Martí	0.65	0.68	0.61	0.55
Sants Montjuic	0.54	0.52	0.58	0.60
Sarrià - Sant Gervasi	0.61	0.63	0.62	0.67

but none of them seem to be an improvement to the general model in terms of performance.

Figure 11 shows on the left the ROC curve and on the right the precision-recall curve for *Model* 1, and its a more visual representation of the model performance. The ROC curve shows how the model performs better than random chance since its above the dashed line, but there is room for improvement since a good performing model shows AUC around 0.90. On the right side, the precision-recall plot shows the trade-off between both parameters, and that we can achieve a high recall with also great precision. This means we can find a lot of positive cases (accepted proposals) without losing a lot of precision (tagging rejected proposals as accepted) which would be the case where we accept every proposal in the platform, and still get a decent accuracy because of the distribution of proposals.

After that, figure 12 shows the same graphics than the previous one but for each district's model. Here we see, compared to the general model plot on figure 11, how both ROC and PR curves are less stable due to the lack of data we have per district. Some of the districts like *Gràcia*, *Nou Barris* or *Sant Martí*, show the incapability to predict proposal acceptance with the data we are using, which points out that the platform itself was not helpful in the process on those cases, or that the data



(a) ROC Curve

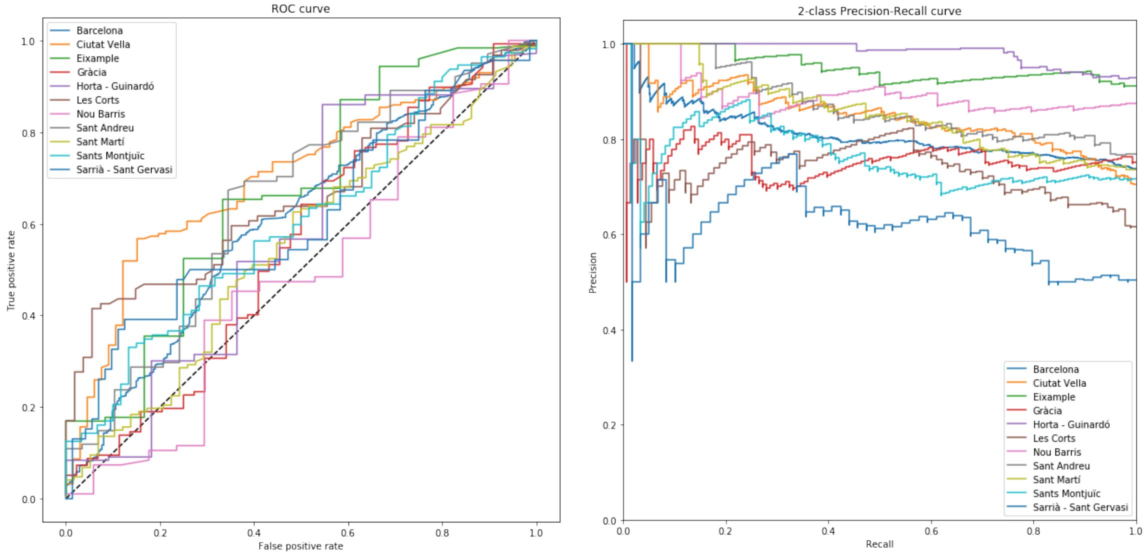
(b) Precision-Recall Curve

Figure 11: Results of *Model* 1 - General Model

captured with the platform was not used in the decision making process. On the Precision-Recall curves, we see how the districts with best AUC are simply the ones where the acceptance ratios where higher as expected, since they don't have that many rejected proposals.

The detailed results for the different models are presented in section 3.1: Figure 16 shows the coefficients for the general model. There we can observe that variables coming from an official source or having neighborhoods like *Eixample*, *Horta* o *Nou Barris* have a positive effect in the prediction. In a similar way, the opposite happens when the proposal comes from a citizen, it has a large number of negative votes or belongs to districts like *Sarrià - Sant Gervasi* or *Les Corts.*

However, not all the variables have equal importance in the classification model. As mentioned in subsection 2.2.3, we measured importance as the normalized percentage of the t-statistic for each model parameter. We have an importance for each feature value because we discretized our categorical variables with one-hot encoding. Figure 13 shows that importance is very spread across a lot of variables, thus not showing a clear driver of the decision-making process of proposal acceptance. To analyze it we added up the importance per group of features.



(a) ROC Curve per district

(b) Precision-Recall Curve per district

Figure 12: Results of *Model* 2 - Per District

First we observe how all *subcategory* features represent a $\sim 36\%$ of the model importance, they are the largest one since they break down in a lot of variables. This is an expected result since they classify each proposal very specifically and can be easily related with PAM's priorities. On the other hand, proposal's *category* variable represents only the $\sim 6\%$ of the importance, which can be explained by two factors, first there were small differences when it came to the acceptance ratios per each category as seen during the analysis (figure 5 in section 2.1.2), and second, we also included *subcategory* features in the model, which already divides the proposal scope in more concrete areas, thus, making category information redundant. For those reasons, *category* variable could be removed from the model without losing much information, and we would reduce 5 coefficients to fit.

The feature set that comes next in relevance corresponds to the *district* features, representing around $\sim 25\%$ of the model importance. Here we observe very different results depending on each neighborhood. The districts with a similar split between accepted and rejected proposals to the whole dataset are not helpful (and have almost none importance), whereas the ones with higher values are mainly because they differ from the norm. We believe that this can be explained by the differences between the ruling parties in city and district halls or the misalignment on priorities.

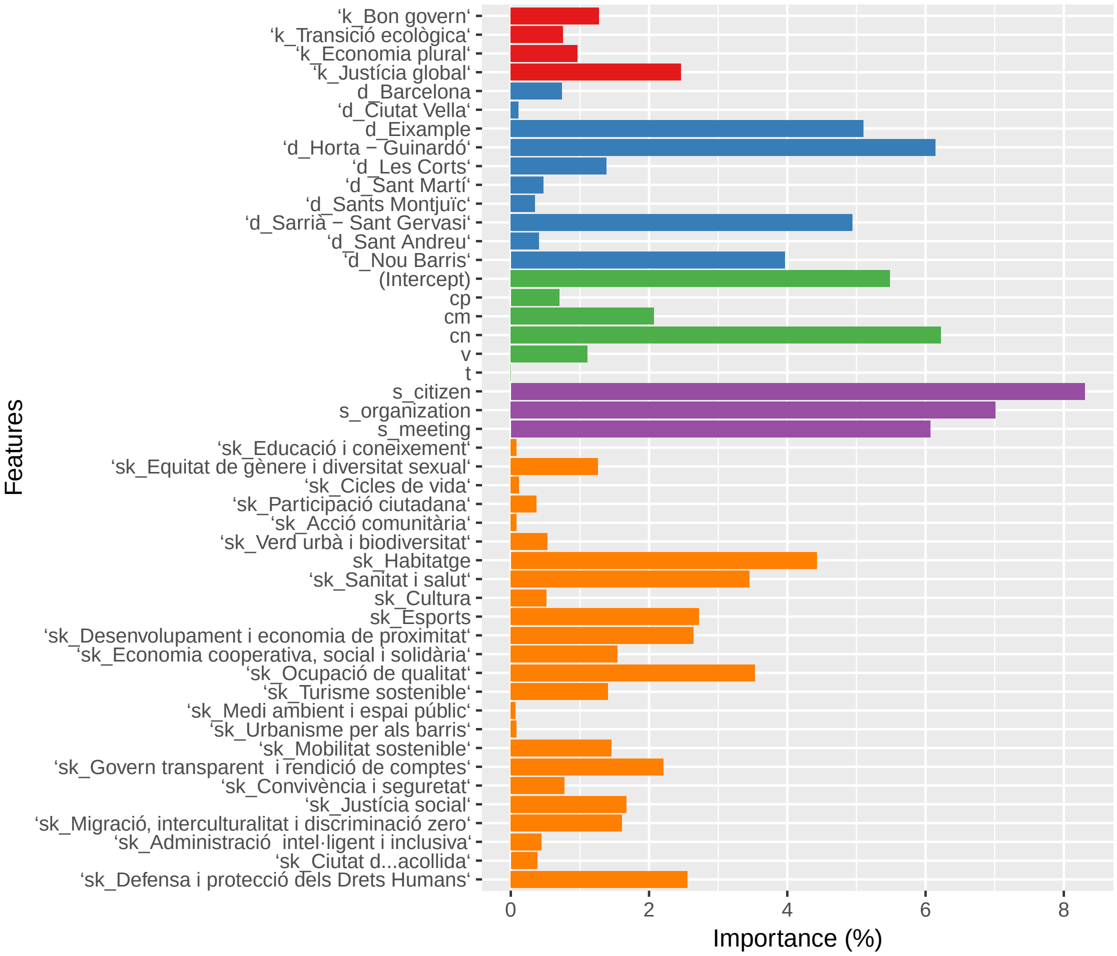


Figure 13: Feature importance in proposal acceptance in *Model* 1, calculated using model's coefficients normalized and color coded grouping types of features.

Next we have the *source* variable giving $\sim 23\%$ of the importance, as we also saw on the analysis, this was also expected from the high difference in acceptance ratios depending on the source of the proposal. Finally, the group of numerical features that represent proposal activity on the platform, only represents $\sim 11\%$ of the total importance. Those are the variables that contain the information about users interaction in the platform, and being the least important would mean the process failed to include citizens feedback into PAM.

Its remarkable how the importance of the model is widely spread across a lot of variables, not showing a clear set of features that influence the response variable outcome. Similar dispersion of importance is found in most of the district models, as we can see on figure 14, where we show *Eixample*'s district model. We chose only

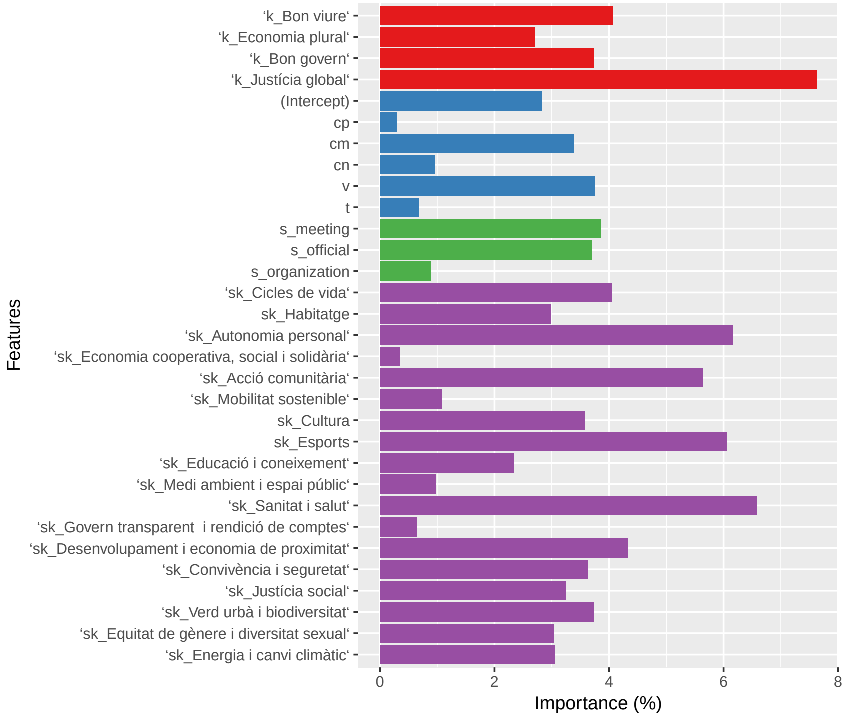


Figure 14: Feature importance in proposal acceptance in *Model* 2 in *Eixample* district, calculated using model's coefficients normalized and color coded grouping types of features.

to show this example more because results are very similar in terms of distribution of importances under the groups of features presented on the general model.

Once provided with the logistic regression results, we wanted to test that those results were not due to the particular algorithm used to classify the proposals. That is why we have also used another prediction model for this binary classification problem for comparison purposes. As presented in the methods chapter, random forests is the algorithm selected to compare with. The main downfall of random forests is that we are not able to clearly identify the reasons why each proposal is classified into each class since they are an ensemble of a lot of decision trees.

Figure 15 shows the comparison of the importance of the variables for *Model 1* using Random Forest algorithm against the Logistic Regression. Results are quite different in terms of importance. While the logistic regression model generally gives more importance to the categorical variables, random forest most important feature is the time a proposal was on the platform. This can be explained by the fact that

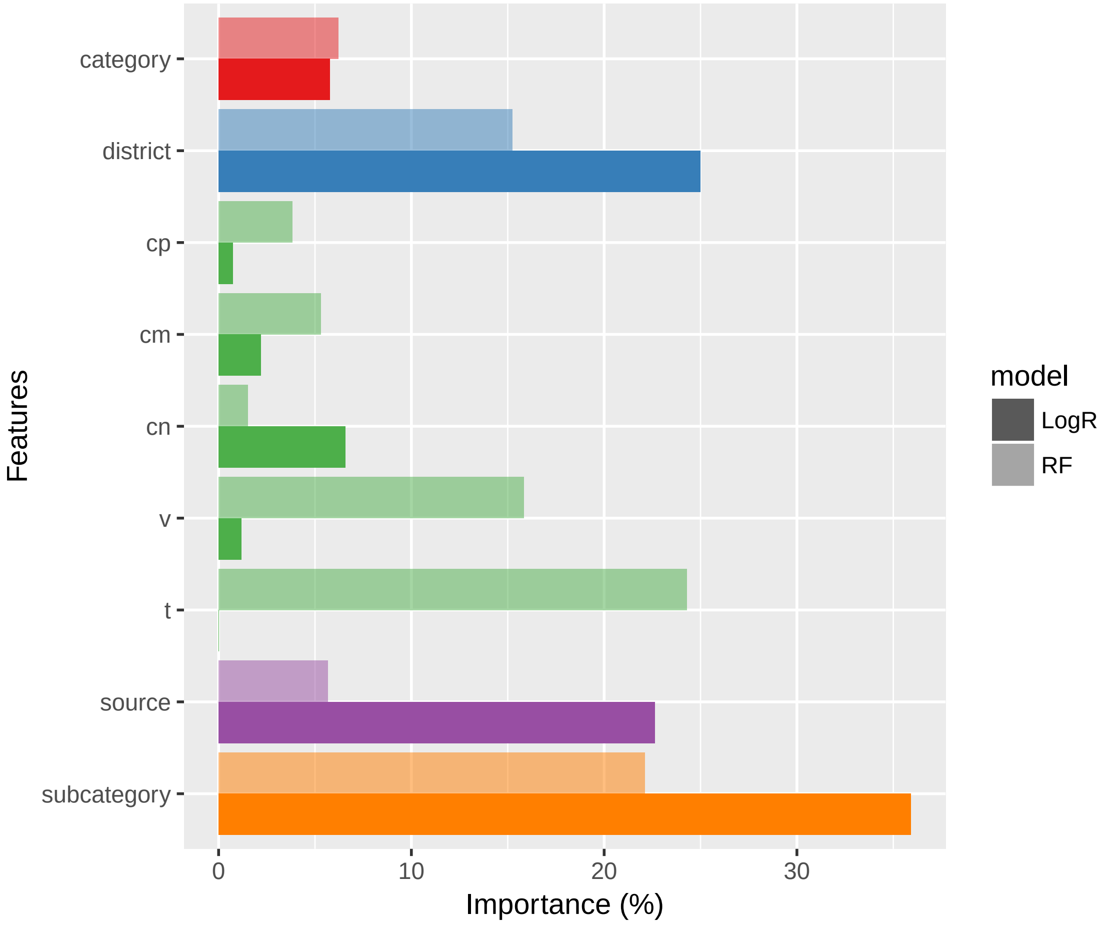
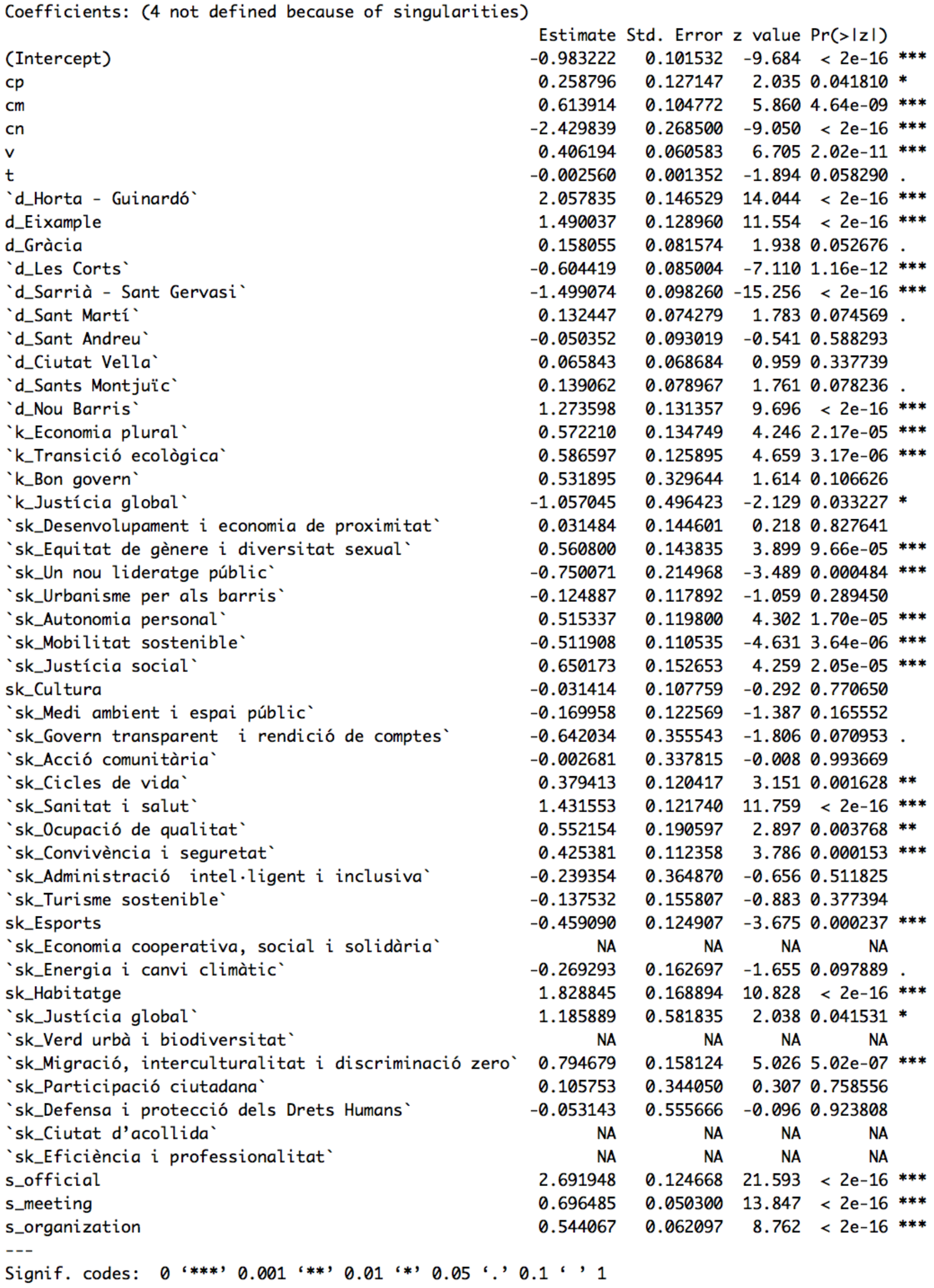


Figure 15: Feature importance comparison of *Model* 1 with Random Forests, color coded by group of features.

having categorical variables with a large amount of possible values, which need to be one-hot encoded for the logistic regression, increases the number of variables and thus the dimensionality of the problem while this transformation is not needed on random forest. On the other hand, seeing that the time on the platform is a very important variable on random forest model makes sense since the oldest proposals were the ones introduced by the government at the beginning of the process, and, as seen in the analysis, are the ones more likely to be accepted. Also accuracy of the random forest, which is around 0.73, is bigger when we compare it with the logistic regression.

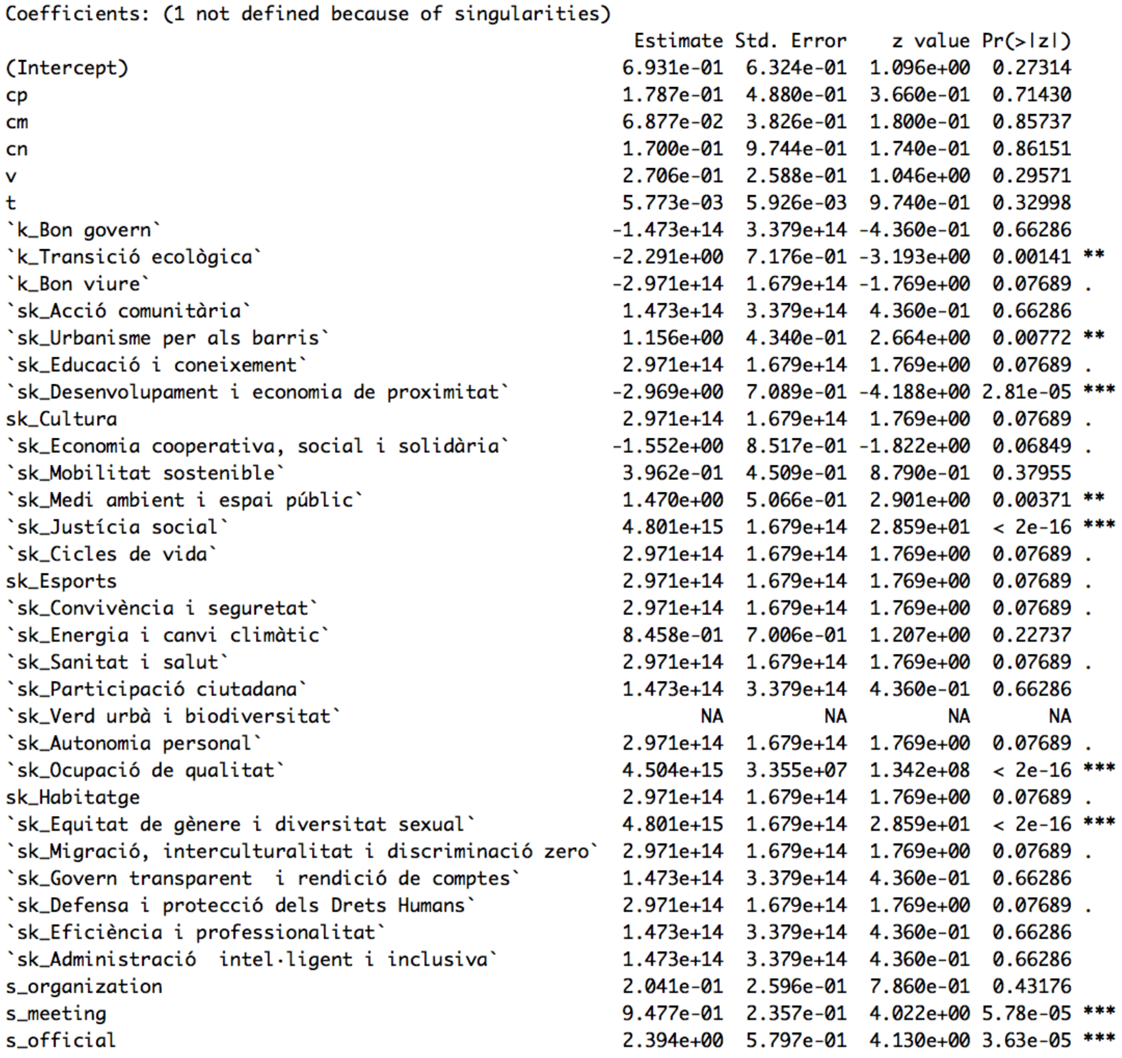
We have analyzed two possible methods for building the classifier, and on both of them we observe a similar predicting power, 0.66 vs 0.73 respectively, that is not really high. This result shows that the data we are using may not be sufficient for achieving the results we expected and further features which can potentially be more informative may need to be introduced. On the other hand, the differences we observe between models come mainly from the way we ensemble them. On the logistic regression, the large amount of categorical variables diminishes the focus on the data gathered on the platform, where the importance is really low, but on the random forest those numerical features have a larger weight on the prediction. This is closer to what we expected, but this model is less transparent to interpret compared to the main one.

Model Results



```
Coefficients: (4 not defined because of singularities)
                                                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                           -0.983222   0.101532  -9.684  < 2e-16 ***
cp                                                     0.258796   0.127147   2.035 0.041810 *  
cm                                                     0.613914   0.104772   5.860 4.64e-09 ***
cn                                                    -2.429839   0.268500  -9.050  < 2e-16 ***
v                                                      0.406194   0.060583   6.705 2.02e-11 ***
t                                                     -0.002560   0.001352  -1.894 0.058290 .  
`d_Horta - Guinardó`                                   2.057835   0.146529  14.044  < 2e-16 ***
d_Eixample                                             1.490037   0.128960  11.554  < 2e-16 ***
d_Gràcia                                               0.158055   0.081574   1.938 0.052676 .  
`d_Les Corts`                                         -0.604419   0.085004  -7.110 1.16e-12 ***
`d_Sarrià - Sant Gervasi`                             -1.499074   0.098260 -15.256  < 2e-16 ***
`d_Sant Martí`                                         0.132447   0.074279   1.783 0.074569 .  
`d_Sant Andreu`                                       -0.050352   0.093019  -0.541 0.588293    
`d_Ciutat Vella`                                       0.065843   0.068684   0.959 0.337739    
`d_Sants Montjuïc`                                     0.139062   0.078967   1.761 0.078236 .  
`d_Nou Barris`                                         1.273598   0.131357   9.696  < 2e-16 ***
`k_Economia plural`                                    0.572210   0.134749   4.246 2.17e-05 ***
`k_Transició ecològica`                                0.586597   0.125895   4.659 3.17e-06 ***
`k_Bon govern`                                         0.531895   0.329644   1.614 0.106626    
`k_Justícia global`                                   -1.057045   0.496423  -2.129 0.033227 *  
`sk_Desenvolupament i economia de proximitat`          0.031484   0.144601   0.218 0.827641    
`sk_Equitat de gènere i diversitat sexual`             0.560800   0.143835   3.899 9.66e-05 ***
`sk_Un nou lideratge públic`                          -0.750071   0.214968  -3.489 0.000484 ***
`sk_Urbanisme per als barris`                         -0.124887   0.117892  -1.059 0.289450    
`sk_Autonomia personal`                                0.515337   0.119800   4.302 1.70e-05 ***
`sk_Mobilitat sostenible`                             -0.511908   0.110535  -4.631 3.64e-06 ***
`sk_Justícia social`                                   0.650173   0.152653   4.259 2.05e-05 ***
sk_Cultura                                            -0.031414   0.107759  -0.292 0.770650    
`sk_Medi ambient i espai públic`                      -0.169958   0.122569  -1.387 0.165552    
`sk_Govern transparent  i rendició de comptes`        -0.642034   0.355543  -1.806 0.070953 .  
`sk_Acció comunitària`                                -0.002681   0.337815  -0.008 0.993669    
`sk_Cicles de vida`                                    0.379413   0.120417   3.151 0.001628 ** 
`sk_Sanitat i salut`                                   1.431553   0.121740  11.759  < 2e-16 ***
`sk_Ocupació de qualitat`                              0.552154   0.190597   2.897 0.003768 ** 
`sk_Convivència i seguretat`                           0.425381   0.112358   3.786 0.000153 ***
`sk_Administració  intel·ligent i inclusiva`          -0.239354   0.364870  -0.656 0.511825    
`sk_Turisme sostenible`                               -0.137532   0.155807  -0.883 0.377394    
sk_Esports                                            -0.459090   0.124907  -3.675 0.000237 ***
`sk_Economia cooperativa, social i solidària`                NA         NA      NA       NA    
`sk_Energia i canvi climàtic`                         -0.269293   0.162697  -1.655 0.097889 .  
sk_Habitatge                                           1.828845   0.168894  10.828  < 2e-16 ***
`sk_Justícia global`                                   1.185889   0.581835   2.038 0.041531 *  
`sk_Verd urbà i biodiversitat`                               NA         NA      NA       NA    
`sk_Migració, interculturalitat i discriminació zero`  0.794679   0.158124   5.026 5.02e-07 ***
`sk_Participació ciutadana`                            0.105753   0.344050   0.307 0.758556    
`sk_Defensa i protecció dels Drets Humans`            -0.053143   0.555666  -0.096 0.923808    
`sk_Ciutat d'acollida`                                       NA         NA      NA       NA    
`sk_Eficiència i professionalitat`                           NA         NA      NA       NA    
s_official                                             2.691948   0.124668  21.593  < 2e-16 ***
s_meeting                                              0.696485   0.050300  13.847  < 2e-16 ***
s_organization                                         0.544067   0.062097   8.762  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 16: *Model* 1: General Model Coefficients



```
Coefficients: (1 not defined because of singularities)
                                                         Estimate Std. Error    z value Pr(>|z|)
(Intercept)                                             6.931e-01  6.324e-01  1.096e+00  0.27314
cp                                                      1.787e-01  4.880e-01  3.660e-01  0.71430
cm                                                      6.877e-02  3.826e-01  1.800e-01  0.85737
cn                                                      1.700e-01  9.744e-01  1.740e-01  0.86151
v                                                       2.706e-01  2.588e-01  1.046e+00  0.29571
t                                                       5.773e-03  5.926e-03  9.740e-01  0.32998
`k_Bon govern`                                         -1.473e+14  3.379e+14 -4.360e-01  0.66286
`k_Transició ecològica`                                -2.291e+00  7.176e-01 -3.193e+00  0.00141 **
`k_Bon viure`                                          -2.971e+14  1.679e+14 -1.769e+00  0.07689 .
`sk_Acció comunitària`                                  1.473e+14  3.379e+14  4.360e-01  0.66286
`sk_Urbanisme per als barris`                           1.156e+00  4.340e-01  2.664e+00  0.00772 **
`sk_Educació i coneixement`                             2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Desenvolupament i economia de proximitat`          -2.969e+00  7.089e-01 -4.188e+00 2.81e-05 ***
sk_Cultura                                              2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Economia cooperativa, social i solidària`          -1.552e+00  8.517e-01 -1.822e+00  0.06849 .
`sk_Mobilitat sostenible`                               3.962e-01  4.509e-01  8.790e-01  0.37955
`sk_Medi ambient i espai públic`                        1.470e+00  5.066e-01  2.901e+00  0.00371 **
`sk_Justícia social`                                    4.801e+15  1.679e+14  2.859e+01  < 2e-16 ***
`sk_Cicles de vida`                                     2.971e+14  1.679e+14  1.769e+00  0.07689 .
sk_Esports                                              2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Convivència i seguretat`                            2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Energia i canvi climàtic`                           8.458e-01  7.006e-01  1.207e+00  0.22737
`sk_Sanitat i salut`                                    2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Participació ciutadana`                             1.473e+14  3.379e+14  4.360e-01  0.66286
`sk_Verd urbà i biodiversitat`                                 NA         NA         NA       NA
`sk_Autonomia personal`                                 2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Ocupació de qualitat`                               4.504e+15  3.355e+07  1.342e+08  < 2e-16 ***
sk_Habitatge                                            2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Equitat de gènere i diversitat sexual`              4.801e+15  1.679e+14  2.859e+01  < 2e-16 ***
`sk_Migració, interculturalitat i discriminació zero`  2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Govern transparent  i rendició de comptes`          1.473e+14  3.379e+14  4.360e-01  0.66286
`sk_Defensa i protecció dels Drets Humans`              2.971e+14  1.679e+14  1.769e+00  0.07689 .
`sk_Eficiència i professionalitat`                      1.473e+14  3.379e+14  4.360e-01  0.66286
`sk_Administració  intel·ligent i inclusiva`            1.473e+14  3.379e+14  4.360e-01  0.66286
s_organization                                          2.041e-01  2.596e-01  7.860e-01  0.43176
s_meeting                                               9.477e-01  2.357e-01  4.022e+00 5.78e-05 ***
s_official                                              2.394e+00  5.797e-01  4.130e+00 3.63e-05 ***
```

Figure 17: *Model* 2: *Eixample* Model Coefficients

Chapter 4

Discussion

Conclusions & Future Work

In this thesis, we have shown the procedure followed in order to build a classifier using a logistic regression model. The main objective of the work was to mine knowledge from the on-line platform *Decidim Barcelona* and the process that took place to decide which proposals were going to be accepted and which not. To do so first an analysis was performed to the whole dataset to get a better understanding at the building of the model. This prepossessing already showed some indicators of which variables could influence the decision. Then when we moved on to the building and understanding of the model itself, we saw the same patterns that were already visible in the previous analysis, but when we compared it against a more complex model, our classifier showed to work not as good, but the most remarkable fact was that each model had different features as main drivers of the classification. On our model, categorical variables had a larger weigh in importance while the Random Forest performed had as primary features the numerical ones tied to the proposal. We learned that the variables used lacked some key information to the decision making process, not just because a proposal is really commented and praised by the community this will imply an action will come out of it and will be accepted. Here political competences and limitations take over, and is where our models are blind. This could be tried to overcome by introducing richer text features, mining the information from title, text and comments from proposals. The results we saw where all interaction features had low importance on the logistic regression model versus more structural ones like the district or the origin of the proposals could be discouraging since it disconnects users' interaction from the decision of acceptance / rejection. Nevertheless, this process was the first of its kind in the city of Barcelona, and its a promising step towards e-democracy. The results we observe are quite biased since most of the proposals are coming directly from the city hall, dumped into the platform and accepted a posteriori. This does not mean that in the near future, if the platform works as its intended to, with a more pure origin of proposals, more insights could be draw out from it repeating a similar analysis and model.

List of Figures

List of Tables

Bibliography

[1] Ajuntament de barcelona: Plà d'actuació municipal, 2016.

[2] P. Aragón, A. Calleja-López, V. Gómez, A. Kaltenbrunner, D. Laniado, M. Manca, A. Monterde, and F. Bria. Networked Models of Democracy. Technical report, Decentralised Citizens ENgagement Technologies (DCENT). Specific Targeted Research Project Collective Awareness Platforms, 05 2015.

[3] P. Aragón, A. Kaltenbrunner, A. Calleja-López, A. Pereira, A. Monterde, X. E. Barandiaran, and V. Gómez. Deliberative platform design: The case study of the online discussions in decidim barcelona. In *Social Informatics: 9th International Conference (SocInfo) (vol II)*, pages 277–287. Springer, 2017.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] C. D. Brown and H. T. Davis. Receiver operating characteristics curves and related decision measures: A tutorial. 80:24–38, 01 2006.

[7] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[8] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

[9] K. P. Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[10] H. Navarro, G. Miritello, A. Canales, and E. Moro. Temporal patterns behind the strength of persistent ties. *EPJ Data Science*, 6(1):31, 2017.

[11] I. Peña López. decidim.barcelona, Spain. Voice or chatter? Technical report, Case studies. Bengaluru: IT for Change. Open University of Catalonia, 2017.

[12] J. S. Cramer. The origins and development of the logit model. 01 2003.