Carla Aullón 29117
Rafael Bardissa 231142

Visual Analytics: Final Project Report

Description Use Case

For our final project, we have decided to make an analysis of the results of the elections in Spain by municipalities. In that sense, our main objective has been to analyse the tendency of change of the political parties in the different regions of Spain which means analysing the most common parties and how would different successful campaigns would affect the results.

Process and Methodology

The first thing we have done has been to obtain the data needed for the project. This has included the data of different election results from the early 2000s until now.

To make the corresponding analysis we have divided it into three parts. The first one consists of visualizing the data and its main characteristics using tableau and tools from python notebooks. The second one consists of applying machine learning to the data that we have, meaning that based on which is the most common political party per municipality, we can try to make a prediction of what would be the results of the last elections in Spain. The third consists of location intelligence, and more concretely on situating the results in a map so therefore we can analyse which are the most useful meetings in the different areas of Spain.

Main Results

Data Visualization

We used Tableau to understand the evolution of the congress election results and the distribution of these in a more spatial way. Concretely, we were able to look at the changes people made between April and November 2019. We have analysed different things such as the distribution of popular political parties, or the number of voters for common parties in different communities as can be seen in Figure 1. We have noticed curious things in the different time moments. There has not been many changes, but we have found that most cities voted similarly in both elections, with the notable exceptions of Madrid, the province of Girona, coastal cities from Andalusia, the region of Murcia and the province of Teruel.

We saw fit to only look at the results of both the five most voted parties in April and the most voted regional parties that significantly altered the voting results.

Aside from these, we also saw two nationwide trends: the collapse of CS in favour of PP and VOX (these are also explained below), and the sharp rise in popularity of VOX, who gained votes in every autonomous community.

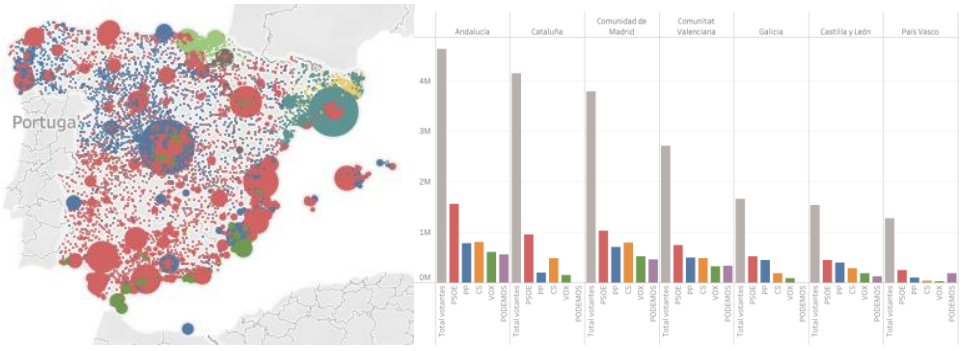

Figure 1. Results from Tableau

Apart from the previous analysis using Tableau, we have also made a superficial analysis of the results of elections in 2019, concretely the results in April versus the ones in November. If we plot a histogram of the percentage of people for each municipality that has voted for a concrete political party, we obtain the following results. In figure 2, we can see the tendency of change that has occurred in people and how many people have changed their minds to vote for the right-mind parties, and how some parties have multiplied their force or reduced it. We can also see that in the vast majority of municipalities there is no party that predominates, meaning that even though we can find the party that has been most voted it has not been voted by 50 percent of more of the population in that municipality.

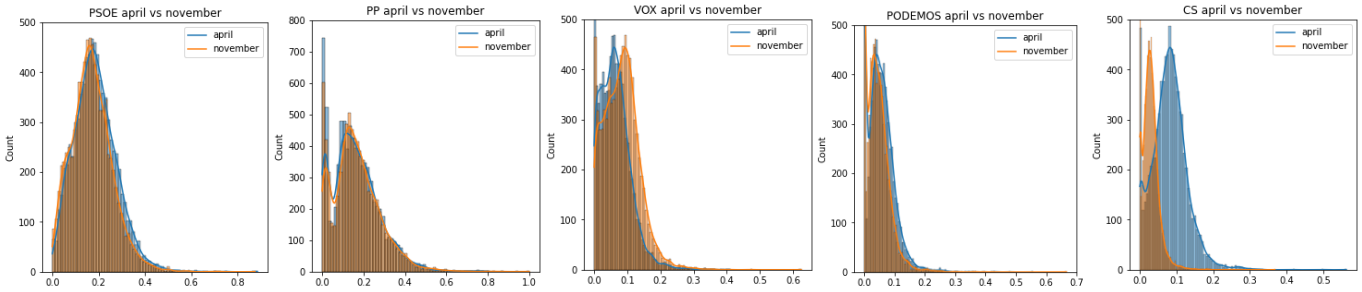

Figure 2. Histograms of percentage results for most common political parties

Predicting elections

Once we have analysed our data and we have visualized our results, we have decided to predict the election results on November 2019. When we say predicting, we mean predicting the number of municipalities in which each party would be the most popular.

To do that we have taken the most popular political party for each municipality. For these, we have performed two different approaches: the first one is taking into account elections during the 2010 decade and the second one is using all the election results from early 2000 until now. In each of them, we have used four different algorithms for prediction: Linear regression, Decision Tree, Random Forest, and Neural Networks. We have used a 75% of the data for testing and 25% for training.

The first option mentioned takes into account results from elections in Nov 2011, Dec 2015, Jun 2016, and Apr 2019. For each, we have fitted the same training dataset and for the testing dataset, the results can be seen in table 1 (left). To analyse the performance of these we have computed the Mean Squared Error that has given the values of 449, 24.532, 28.661, and 1.092 for linear regression, decision tree, random forest, and neural networks, respectively.

For the second option, we have taken into account results from the elections in Mar 2000, Mar 2004, Mar 2008, and the ones mentioned before. For each, we have fitted the same training dataset, different from the previous one, and for the testing dataset, the results can be seen in table 1(right). To analyse the performance of these we have computed the Mean Squared Error that has given the values of 2.837, 26.649, 38.321, and 7.532 for linear regression, decision tree, random forest, and neural networks, respectively.

	Popular_2019_11	Prediciton_linreg	Prediciton_dtree	Prediciton_rforest	Prediciton_nn
CCA-PNC	7.0	25.0	0.0	0.0	18.0
PCTE	0.0	12.0	0.0	4.0	1.0
IU-LV	0.0	12.0	0.0	0.0	0.0
PPSO	1.0	13.0	0.0	4.0	2.0
ERC	622.0	674.0	153.0	114.0	720.0
IU-UPEC	0.0	11.0	0.0	2.0	0.0
MÉS	0.0	11.0	0.0	2.0	0.0
GBAI	1.0	14.0	0.0	0.0	0.0
PRC	31.0	25.0	3.0	3.0	21.0

	Popular_2019_11	Prediciton_linreg	Prediciton_dtree	Prediciton_rforest	Prediciton_nn
FE	0.0	4.0	0.0	0.0	6.0
FAC	0.0	4.0	0.0	1.0	6.0
LV-GV	0.0	4.0	0.0	5.0	6.0
CDN	0.0	4.0	0.0	0.0	6.0
CCA-PNC	7.0	19.0	1.0	1.0	18.0
ERC	622.0	716.0	5.0	107.0	661.0
F.A.	0.0	4.0	0.0	0.0	6.0
UV	0.0	4.0	0.0	0.0	6.0
ARALAR	0.0	4.0	0.0	0.0	6.0
UPYD	0.0	4.0	0.0	1.0	6.0
JXCAT-JUNTS	241.0	105.0	106.0	121.0	137.0
PRC	31.0	19.0	3.0	2.0	25.0
CIU	0.0	121.0	1.0	542.0	316.0
PH	0.0	3.0	0.0	0.0	6.0
ES2000	0.0	4.0	0.0	0.0	6.0

Table 1. Predictions for different algorithms using elections from 2010's (left) and century XIX (right)

From this initial analysis we can see that when we have more previous results, the prediction seems to be more accurate; in a sense that we can capture with more depth the tendency of each political party. We have to mention that even though the results in for each analysis are different we can see a similar disposition.

Although decision trees are considered as an easy tool to use and understand, they are not a good algorithm to predict these values. This might be because we cannot predict the results based on asking a question and deciding if our answer is one or the other, so in this sense we consider that decision trees are a useful tool for classification. Similarly, we have random forests which have an alike structure as decision trees, but now we have more than two options at each step. Even they can capture non-linear relations, we consider that it is still not a good option. Talking about linearity we have linear regression model, this one captures linear relations between the features and predicts the results based on the hyperplane. Knowing that typically the relations don't follow a linear structure, the results are accurate as is shown in mean squared error and more concretely for higher values (but not for lower ones). Finally, we have neural networks, which seem to give a general good results. The problem of this method is that it acts as a black box and we don't really know how much it has learned from the raining dataset, but for the small dataset it

seems to work pretty well. However, we need to mention that neural networks have a results of being random, and this can affect to its performance.

Taking into account the results previously mentioned, we have decided to make a further analysis on linear regression and neural networks for both options. For linear regression, we have plotted the prediction results in comparison to elections in Apr 2019 as scatter plots shown in Figure 3. For neural networks, we have computed the shap values for each options mentioned above and the results are shown in Figure 4. From these we can see that the majority of the values tend to have a low impact on the model output, which means that for those samples that have not been influenced by the concrete feature they have also low values relative to other instances. Moreover, only concrete parties had a huge positive impact in the further prediction. We can also see that in both cases last elections had a negative impact but being the more influential one.

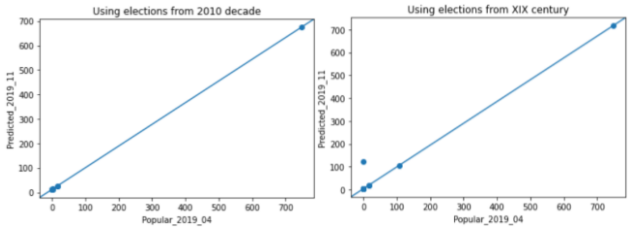

Figure 3. Scatter plots from test sets in elections in April and November 2019

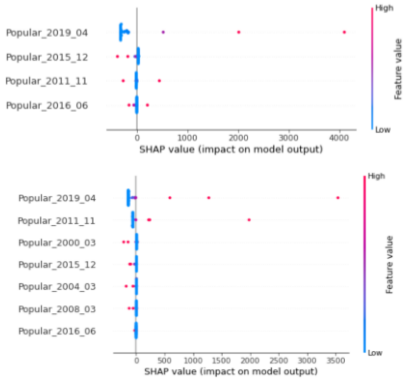

Figure 4. Results from shap values from two options explained

Remark that we consider that no algorithm works fine for our data since the results can depend on previous elections but this is not the unique or the main characteristic that influences on the results. Moreover, the algorithms mentioned cannot distinguish the performance of the different political parties, meaning that if one party crashes, the votes typically are taken by other parties; so when some values decrease others increase. However, the general tendency is captured in a good way in linear regression.

Convenient meetings

We used CARTO in conjunction with the geo spatial location of all cities (provided by RTVE) to create a small web application where the different cities are shown in a 3d map as can be seen in figure 5. We can also see the counts the number of PP voters in the viewport.

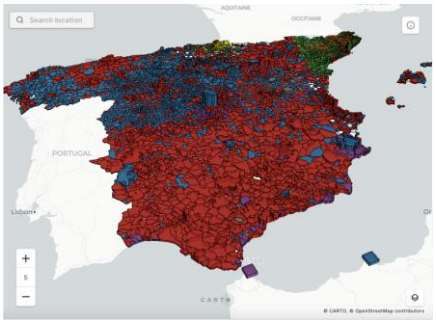

Figure 5. Carto results in 3D

Conclusions

Based on the results, we can conclude that the elections are something more complex than what we initially planned. We know that there are a lot of factors that can make someone vote for a concrete political party. Taking into account that, we consider that the analysis performed for these datasets are nearly the best that we can obtain.

We consider that the data we've gathered is as good as it can be. Unfortunately, we could not afford to carry on a larger project, mainly due to time constraints. In particular, the project could be expanded with a more complex predictive model and a more in-depth analysis of the changes in voting trends over time. This can be done by taking into account more features for each municipality, for example, the PIB, its area of it, or the unemployment rates; or maybe also political scandals.