Regression Analysis of an Electoral Participation Data Set for the 2021 Parliamentary Elections in Catalonia

Elitza Maneva

1. Main objective of the analysis

Our objective is to interpret observed electoral participation with respect to neighborhood characteristics. We want to find what kinds of business venues or combinations of venues are more positively or more negatively correlated with voter turnout. The final goal is to identify characteristics of neighborhoods that lead to low voter turnout in particular, but for the current study we will concentrate on regression analysis, looking at models for the exact voter turnout, and not on classification.

2. Description of the data set

This data set was created for the Capstone Project of the Data Science IBM Specialization on Coursera. There are two main sources of data. The first source is the publicly available data set from the elections held on February 14, 2021 in Catalonia. In particular we use the census and voter participation by electoral sections. The second source of data is the Foursquare API which we queried for venues in the proximity of the electoral sections. In order to create the data set we also had to use geographic data about the electoral sections in Catalonia, which is publicly available on a government website.

Before querying the Foursquare API we removed some outliers, namely sections with census less than 400. These generally have turnout that is much higher than average, and also correspond to geographic areas with very sparse venue data on Foursquare. After removing these sections our table has 4643 rows corresponding to the remaining electoral sections. Each section is identified by its municipality, district and section number, as well as geographic coordinates for the centroid of the area it encompasses.

The venue data was processed as part of the Capstone Project. The venue categories were reduced to 30 types (down from 698 that were provided by Foursquare). We have columns for the number of venues returned by Foursquare for each of these 30 categories. The total number of venues is 175,354. The average number of venues per electoral section is 109 (note that some venues correspond to more than one electoral section, especially in the very dense areas).

Additionally there are columns for the census, for the turnout, and for the population density.

The total number of citizens in the census is 5,368,992, and the number who voted is 2,874,610 for a total turnout of 53.54%. The country occupies an area of 32,108 km^2.

On the map below you can see the geographic locations of the electoral sections. The transparent points are the ones that were removed from the set for having too few names in the census. The red dots correspond to sections with relatively low turnout, and the blue dots to relatively high turnout.

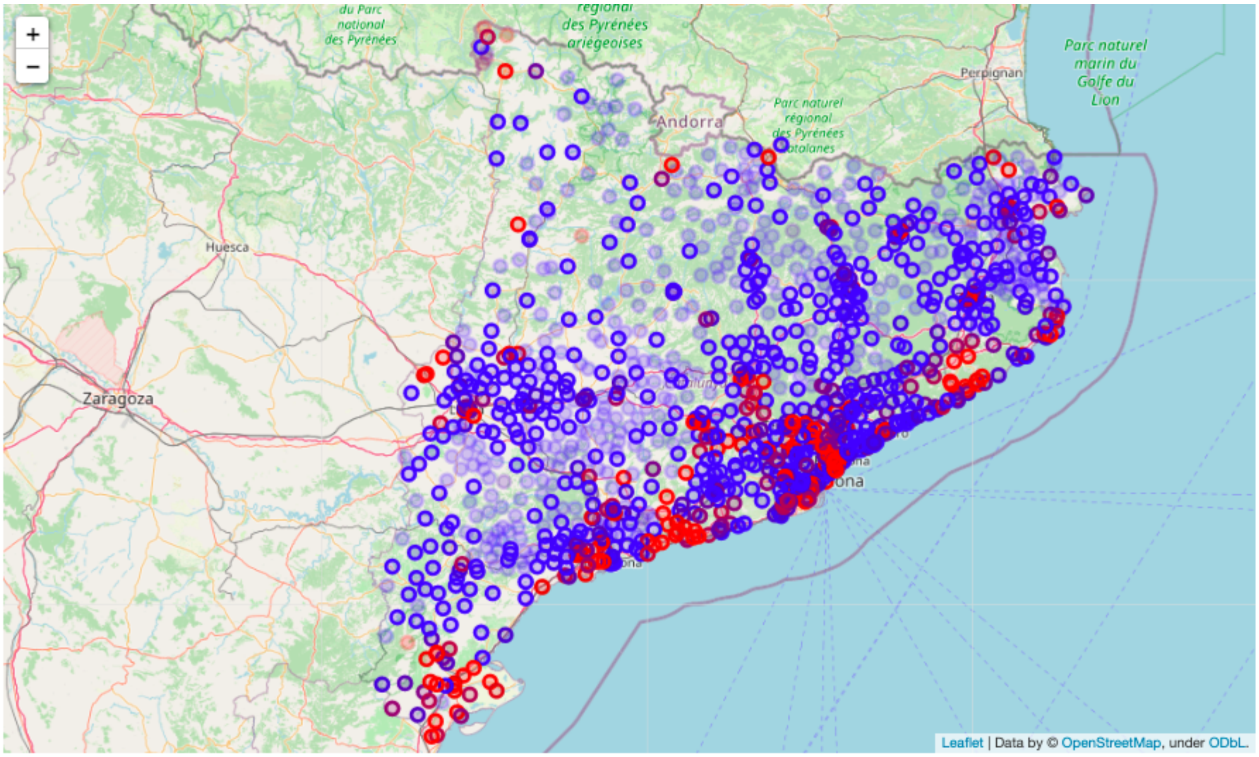
+
−
du Parc national des Pyrénées
régional des Pyrénées ariégeoises
Andorra
Perpignan
Parc naturel marin du Golfe du Lion
Parc naturel régional des Pyrénées catalanes
Huesca
Zaragoza
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 1. Electoral sections in Catalonia, color coded depending on the level of voter turnout in the 2021 elections. Red corresponds to lower turnout.

As an example, below we have a map with the locations of the venues associated with 10 different electoral districts in the municipality of Esplugues de Llobregat.

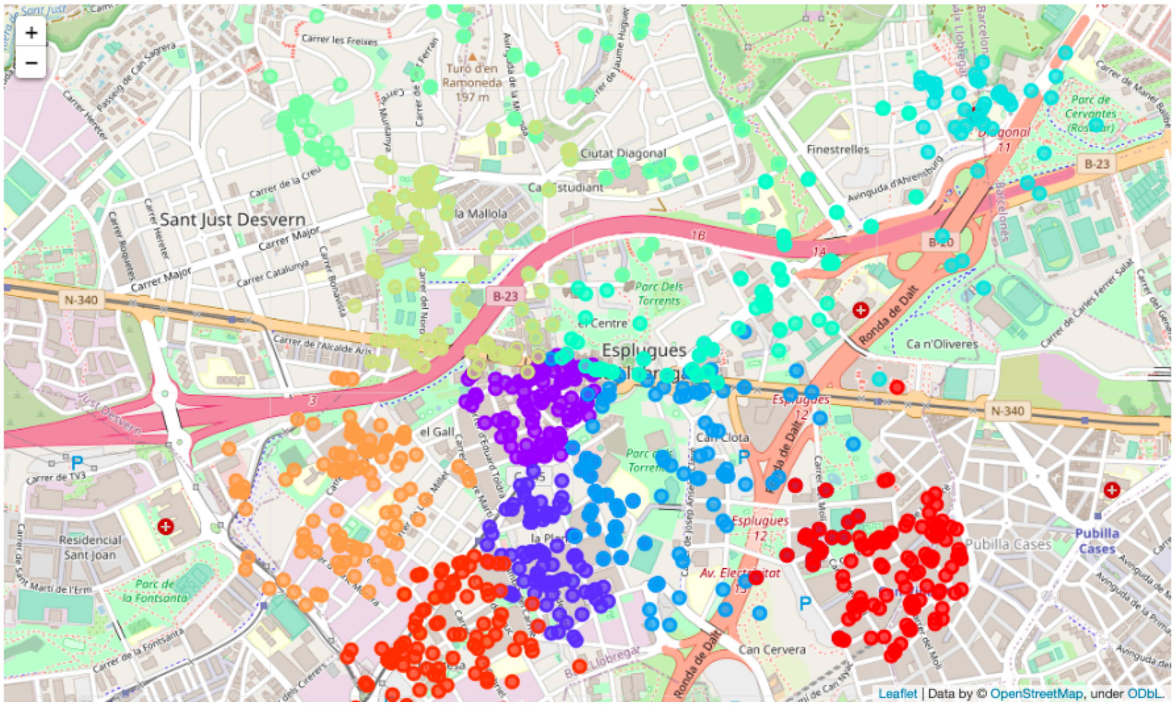

Figure 2. Locations of the venues returned by the Foursquare API for the 10 electoral districts in the municipality of Esplugues de Llobregat.

3. Data cleaning and feature engineering

3.1. Distribution of voter turnout

The distribution of our target variable is shown in Figure 3.

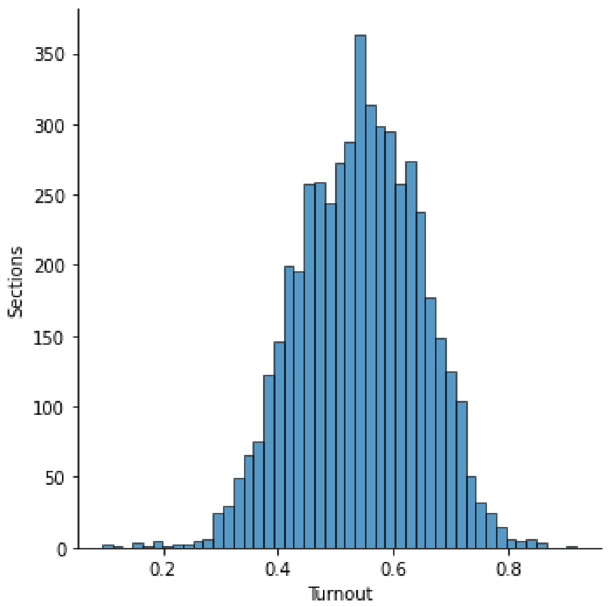

Figure 3. Distribution of voter turnout by sections.

In order to give more importance to the sections with low turnout, which are the ones we are more interested in identifying, for our regression analysis we will consider as a target variable the logarithm of the absenteeism, which is *log (1 - turnout)*. Its distribution is shown in Figure 4.

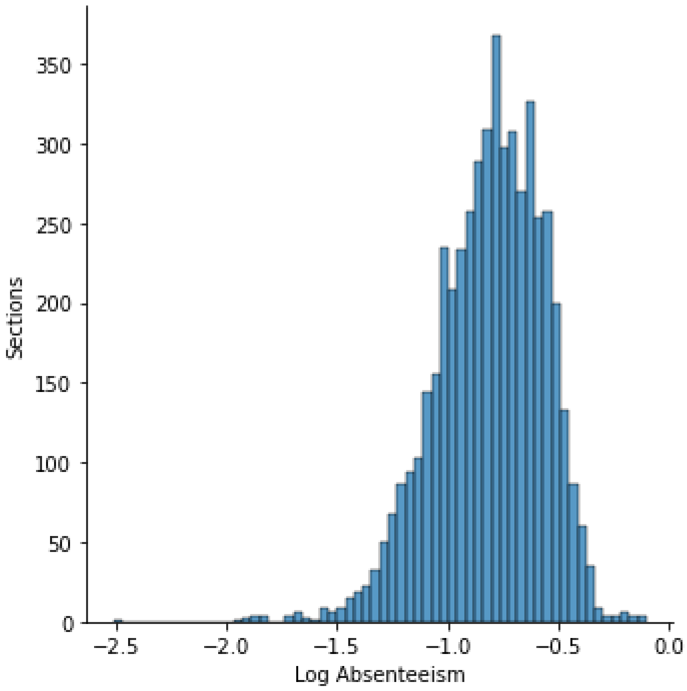

Figure 4. Distribution of the log of voter absenteeism by sections.

3.2. Sizes of sections (census)

As part of the data cleaning process we removed 90 sections for which we didn't have geographic coordinates. Next, we removed sections with a very small census as they can be considered outliers. The distribution of the sizes of the sections is shown in Figure 5. The small sections generally have high turnout as shown in Figure 6.

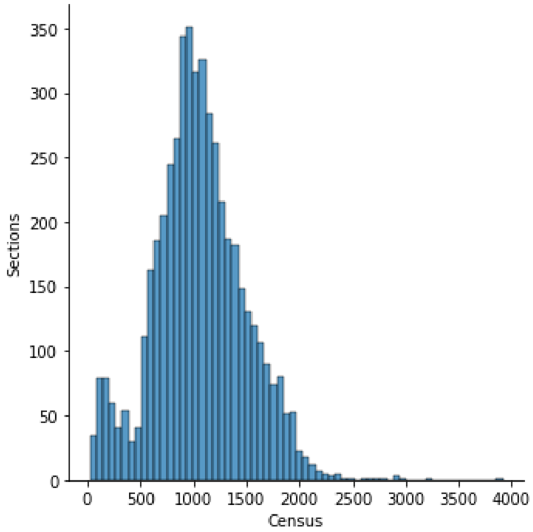

Figure 5. Distribution of the census size of the sections.

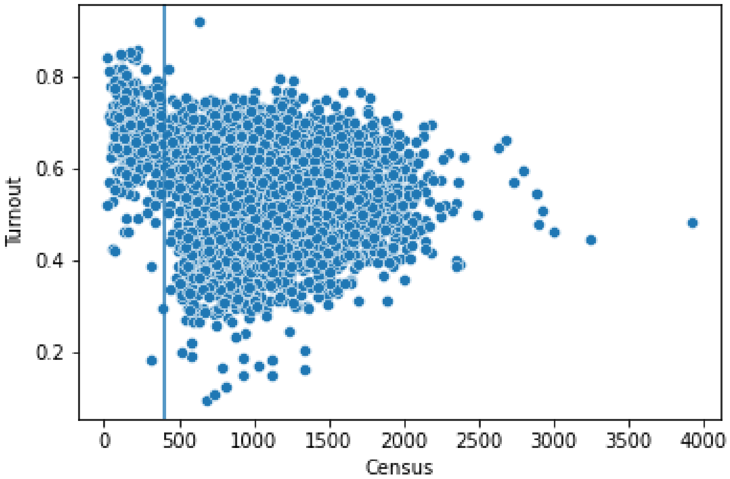

Figure 6. Voter turnout with respect to census size. The small sections that we remove from the set before querying Foursquare are the ones to the left of the vertical line.

3.3. Population density

The distribution of the logarithms of the population density and its relationship with voter turnout are plotted below.

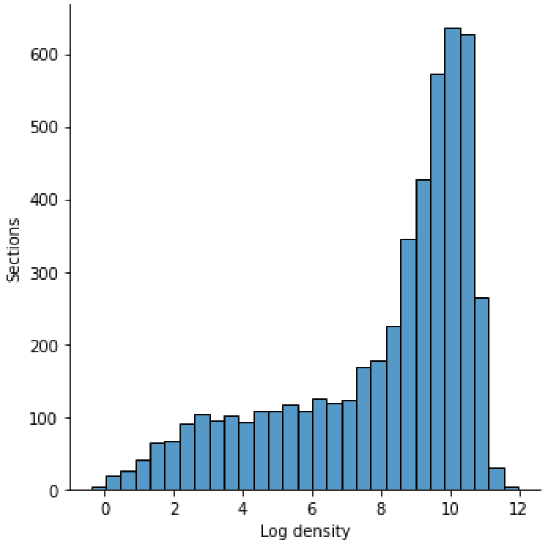

Figure 7. Distribution of the logarithm of the population density by electoral sections.

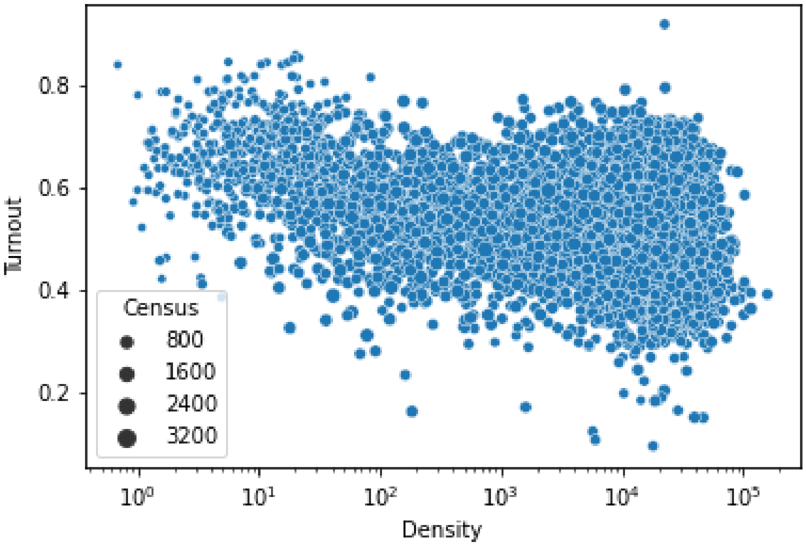

Figure 8. Turnout versus population density by electoral sections.

3.4. Categories of venues

We reduced the categories of venues to 30 down from 698 by merging some and by removing low frequency ones. Figure 9 shows a Word cloud of the categories before and after summarizing them.

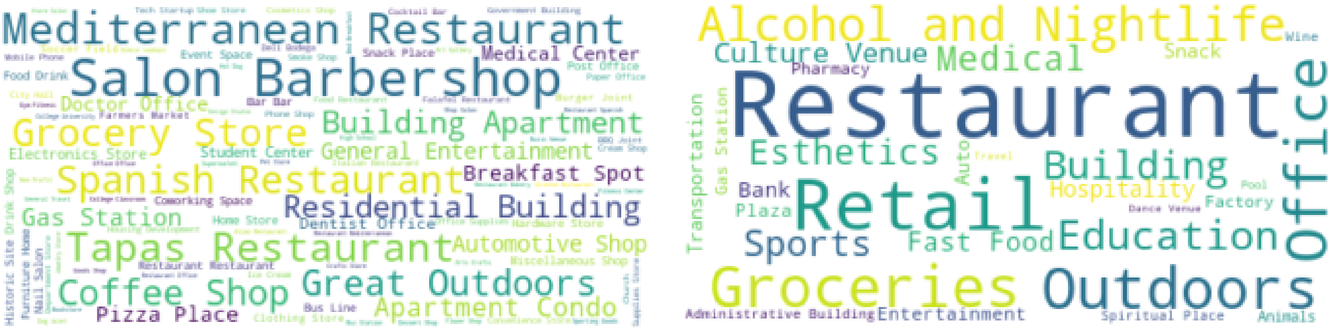

Figure 9. Word cloud of categories before and after summarizing the categories.

3.5. Selecting independent observations

We went over the electoral sections in alphabetical order and only included in our final data set those that have at most 4 venues in common with the sections that were already selected. This way we avoid dependencies between the observations included in the final data set. At the end of this selection process we have 1049 electoral sections in our data set (down from 4643).

3.6. Correlations

In Figure 10 we have a pairwise correlation plot for the 15 venue categories most strongly correlated with voter turnout, as well as the population density. The categories are sorted by their correlation with the turnout (more precisely with the logarithm of the absenteeism).

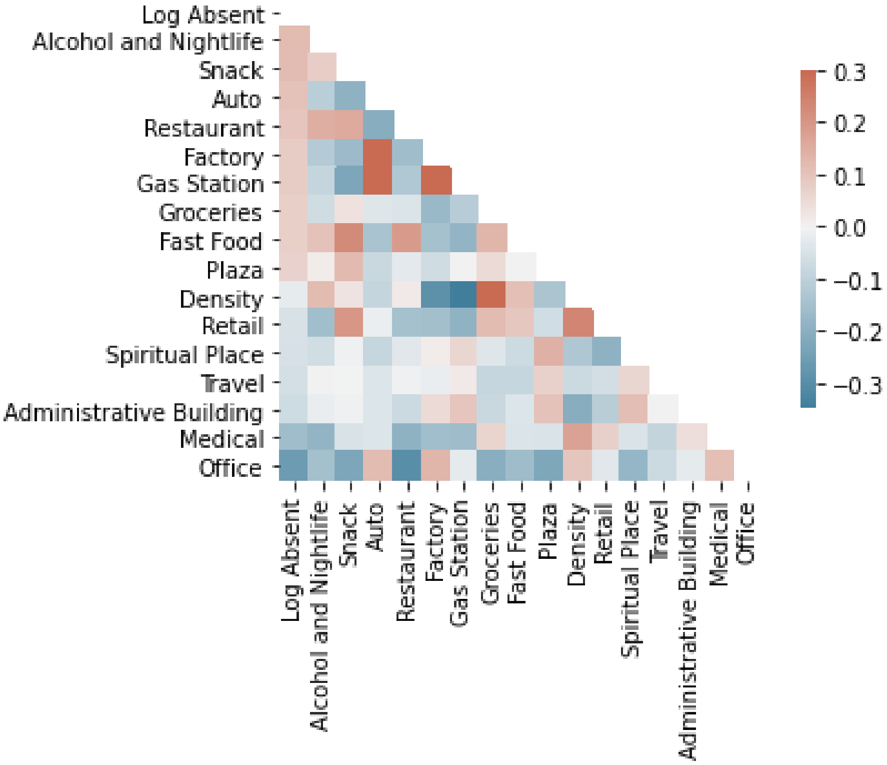

Figure 10. Heatmap of the correlations between the top 15 categories most correlated with the logarithm of the absenteeism.

3.7. Features from inverse frequencies

In addition to the frequencies of a particular kind of venue it would make sense to consider the ratios between frequencies, for example *how many offices there are per cultural venue in the neighborhood*. In order to consider such ratios in the polynomial regression model we generate new features that are 1/(# venues +0.1). (Notice that the +0.1 in the formula serves only to avoid division by 0).

4. Regression Analysis of the Data Set

4.1. Linear regression

We fit a linear regression model to our data of 1049 electoral sections, with the frequencies of the top 30 categories as features, as well as the population density as an additional feature. As a target variable we used the logarithm of absenteeism. This resulted in R2 score of 0.165. In Figure 11 the predicted turnout is plotted versus the real turnout. (Notice that the model was trained on the data set of 1049 independent sections and the plot is for the same set.)

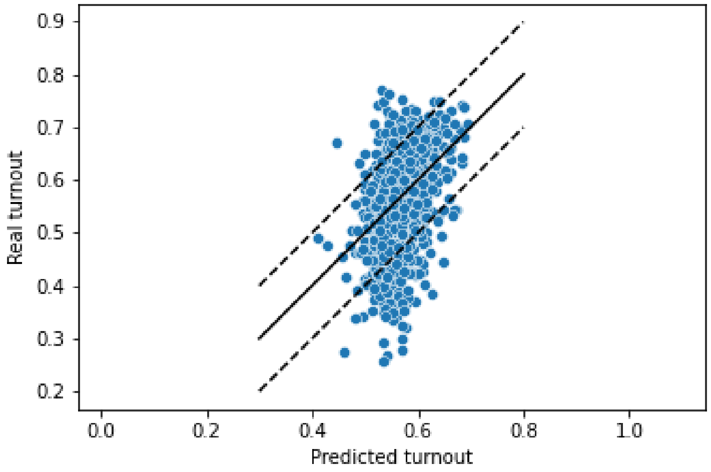

Figure 11. Predicted versus real voter turnout for the linear regression model on 31 features.

The coefficients of this linear regression model are shown in Figure A.2.

4.2 Polynomial regression

We generated degree-2 features out of the top 15 venue categories and the top 14 inverse frequencies (in both cases we chose the ones with largest correlation with voter turnout), as well as the population density. This resulted in 496 features. We built the linear regression model based on those features for the log of the absenteeism. The R2 score we obtain in this case is 0.42. In Figure 12 you can see the real versus the predicted turnout for this model.

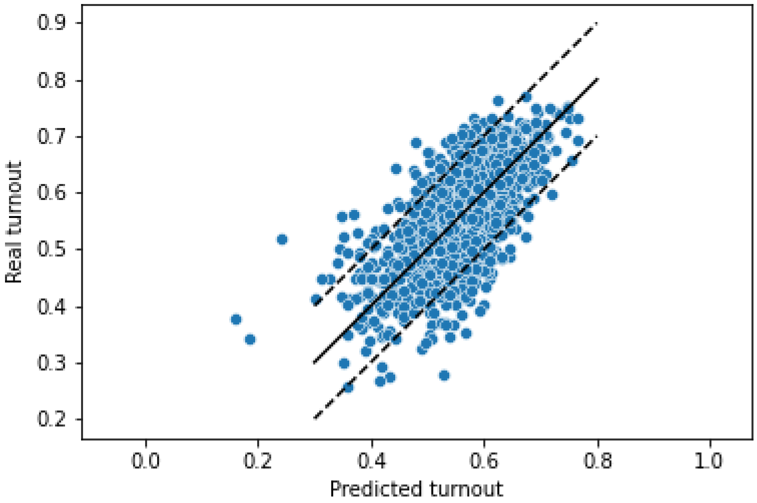

Figure 12. Predicted versus real voter turnout for the polynomial regression of degree 2.

4.3 Regularization with LASSO

Since we are interested in models for interpretation and not prediction, we applied LASSO to the degree-2 features from the previous section. We tried it with different values for the regularization parameter, and plotted the R2 score with respect to the number of non-zero coefficients in Figure 13.

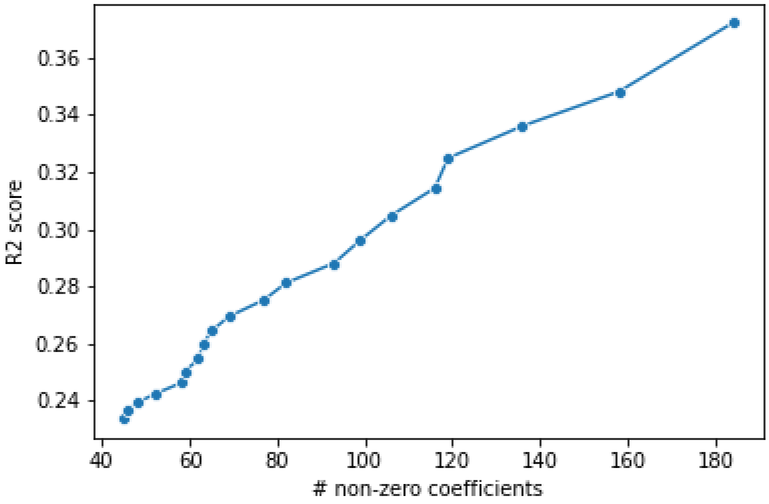

Figure 13. The dependence of R2 score on the number of non-zero coefficients of the LASSO model.

In Figure 14 you can see the predicted versus real turnout for the LASSO model with 65 non-zero coefficients. The 20 coefficients of largest absolute value are shown in Figure A.3.

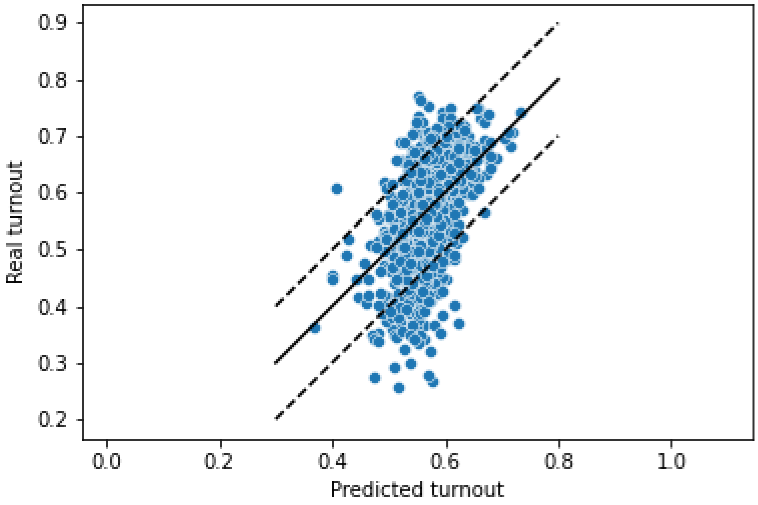

Figure 14. Predicted versus real voter turnout for the LASSO regression with regularization parameter 0.0024, 65 non-zero coefficients and R2 score 0.26.

5. Key findings, insights and recommendations

There are four important observations that we want to emphasize.

1. The venue categories that consistently come out as more influential in the regression model are those of *Office*, and *Medical*. This is a bit unexpected, as it goes against the initial intuition of the map of low turnout areas, which indicated that the dense city areas are the ones with lower turnout.

2. The predicted turnout is much more concentrated around the mean than the real voter turnout both for the linear regression model and for the Lasso model. This is an indication that our models have a large bias towards the mean voter turnout.

3. Both the linear regression and the Lasso models seem to fit better the high turnout areas then the low turnout ones, which are the ones we would be more interested in identifying. Perhaps we need to apply a classification model, such as logistic regression and concentrate specifically on identifying low-voter-turnout sections.

4. Although the population density has a small coefficient in the linear regression model, when features of degree 2 are added it is present in the majority of features with largest coefficients in the Lasso model. This indicates that we should differentiate between high-density and low-density population areas and build different models for each subset.

6. Suggestions for next steps

We find that the main obstacle to getting a better model is the small size of the data set, after removing dependent observations. We suggest as a next step using the data from an election for Spanish Government, which would be 5 times bigger.

Furthermore, we suggest clustering the sections according to both population density and venue characteristics, with the goal to identify rural versus urban areas. We suspect that if we build a different regression model for each of the two clusters separately we would find coefficients that are easier to interpret.

7. Data quality

In addition to the size of the data, which we commented on in the previous section we have to reflect on the quality of the Foursquare data. Unfortunately, we have no control over how the venues are chosen by Foursquare. We can only hope that the venues returned are a good sample of the businesses in each area.

We noticed some misclassification in the categories of the business venues. It may be possible to use the names of the venues together with a Catalan-English dictionary in order to infer their true categories.

Another way to improve the data is to query Foursquare specifically for particular types of venues for which we have some indication that are more highly correlated with voter turnout. Currently we may be looking at a lot of data that is largely irrelevant to the target variable and we may be looking for patterns where there are few.

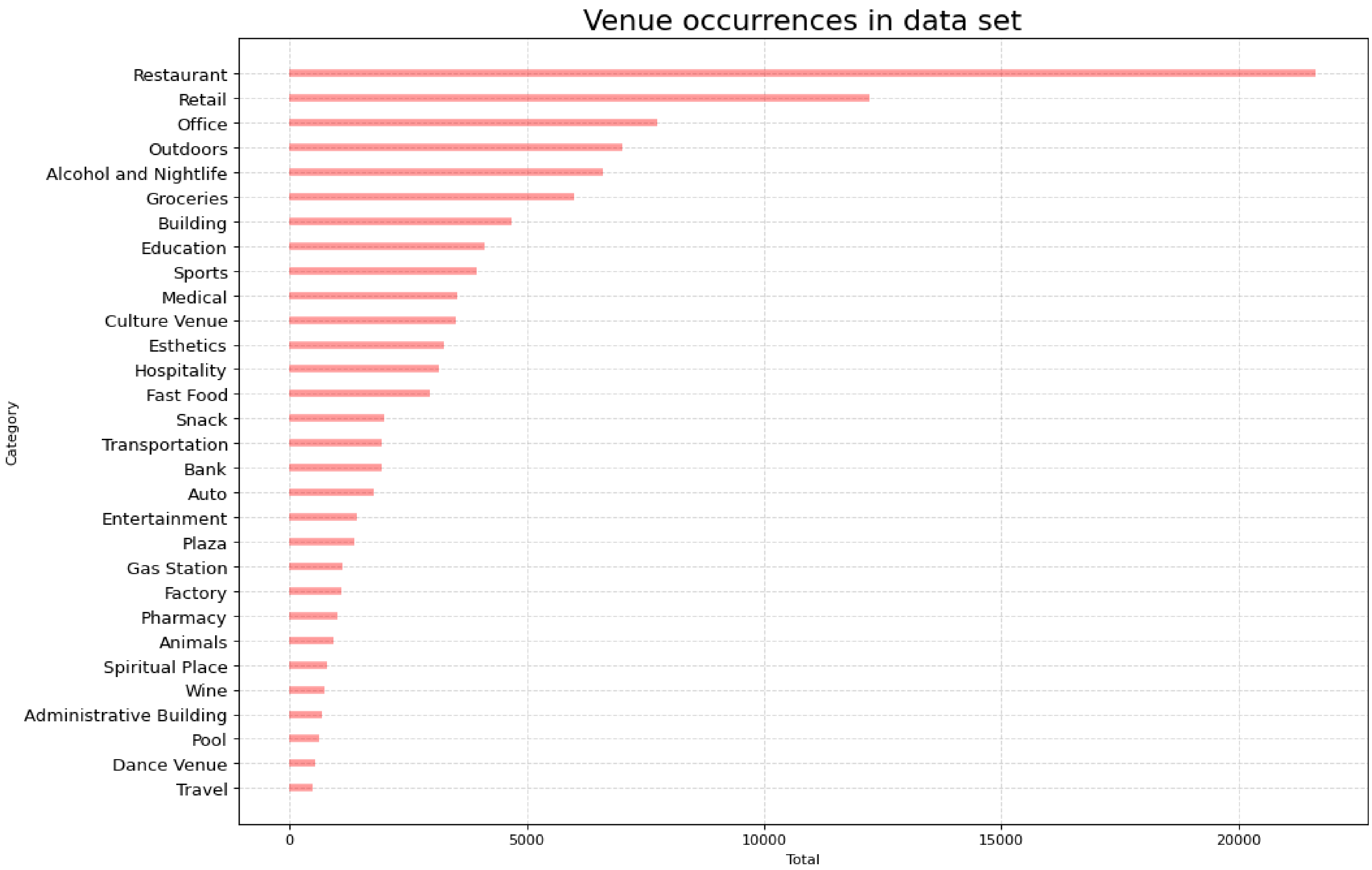

Figure A.1. Total number of venues of each kind for the 1049 selected electoral sections.

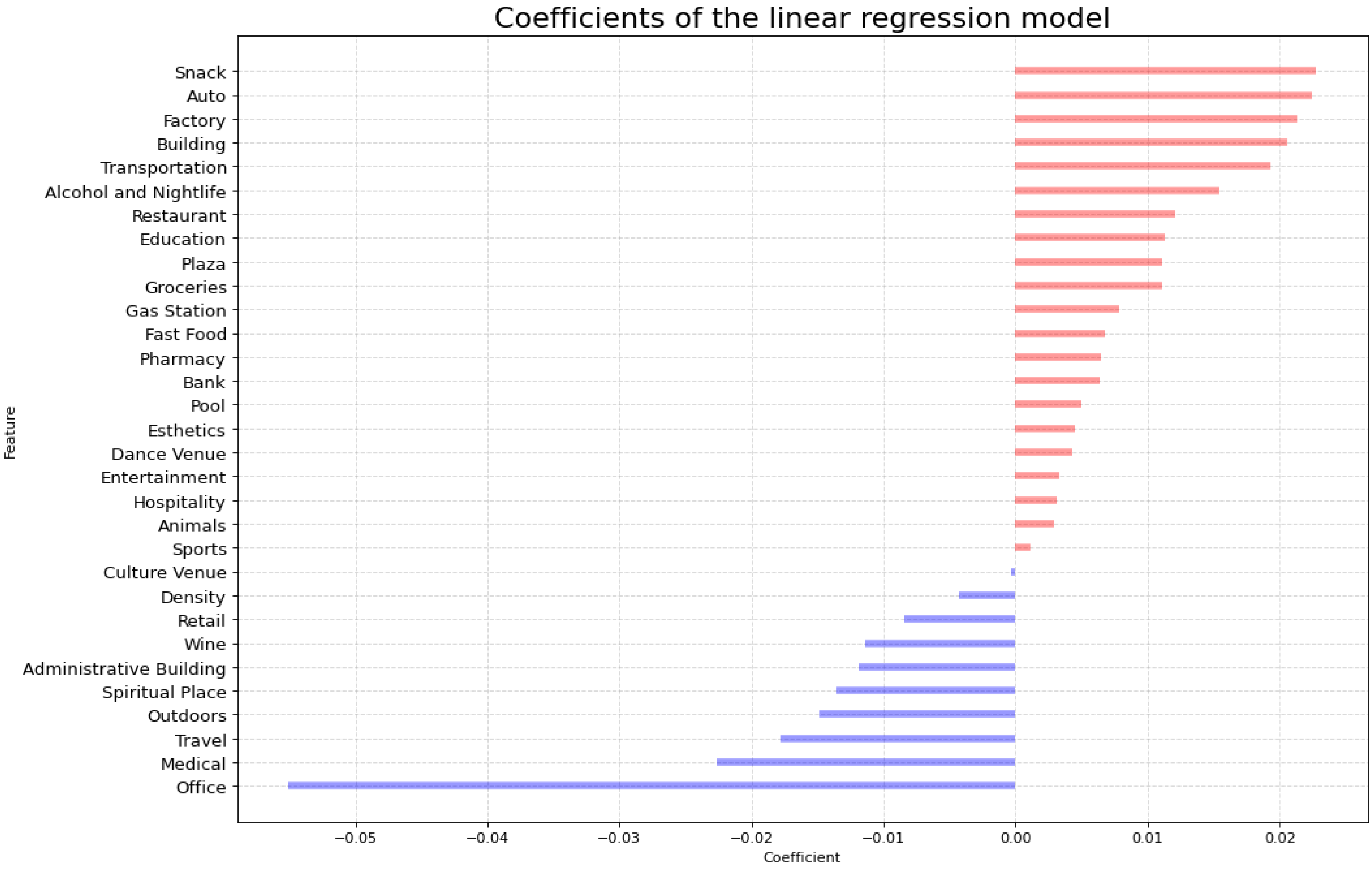

Figure A.2. Coefficients of the linear regression model.

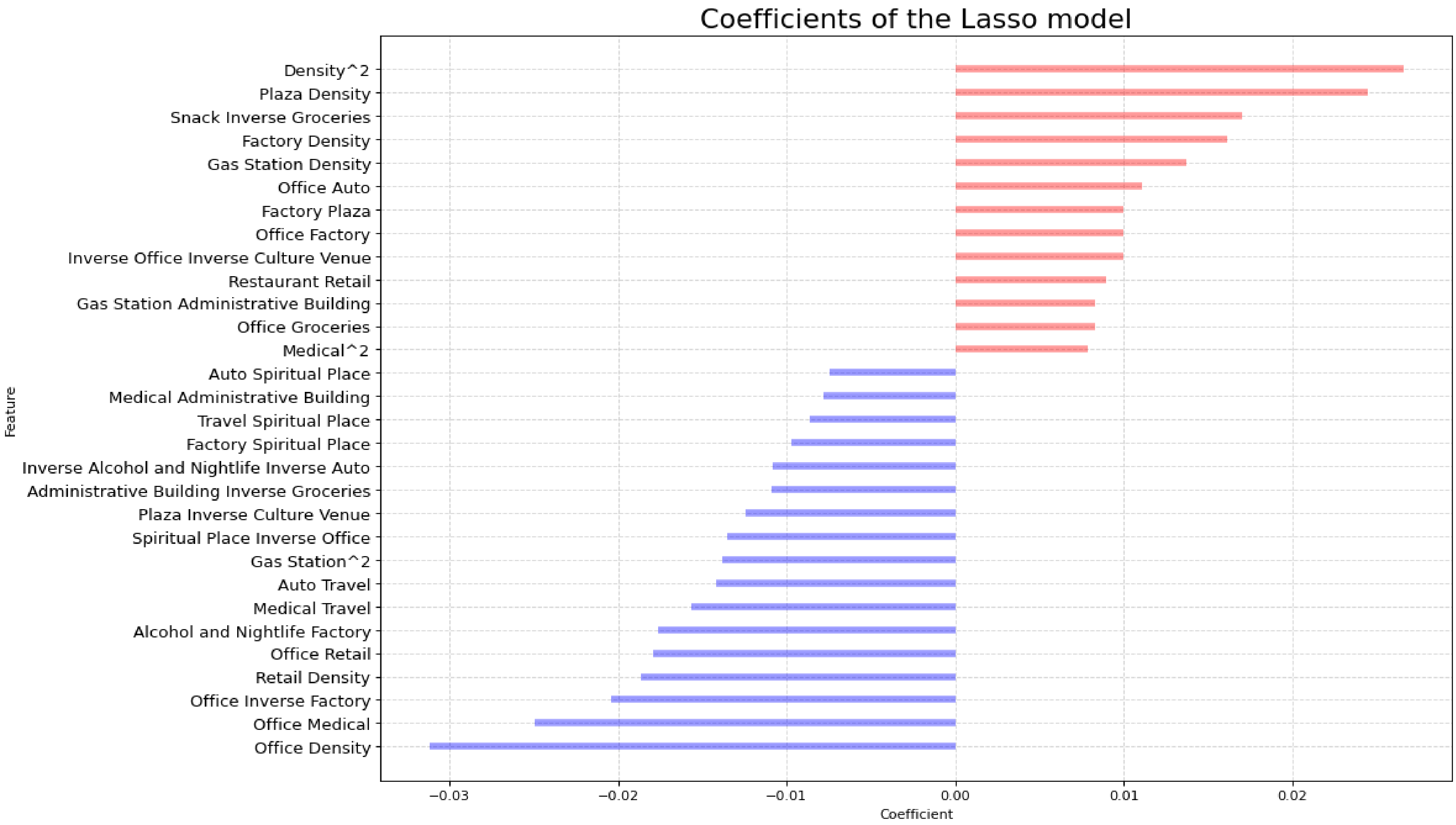

Figure A.3. Largest and smallest coefficients of the LASSO regression with score 0.26 and 65 non-zero coefficients.