Predicting rental prices in Hamurg

April 20, 2020

Troy Figiel
troy.figiel@gmail.com

Contents

1 Introduction

Hamburg is a large and diverse harbour city in the north of Germany and with a population of over 1.8 million, it is the second biggest city in Germany. This population is spread out over more than a hundred districts which vary widely in their characteristics. From the famous nightlife in St. Pauli to the beautiful harbour side, it can be difficult for someone to find their place in Hamburg. The diversity of the districts is equally reflected in the rental prices for apartments. Our goal is to explore what type of venues correlate with an increase in rental price as to be able to give advice on what to expect from a district depending on its price range. This will allow people looking for a new apartment to make a more informed decision on where to rent an apartment.

More precisely, in this report we will look at the effect of different types of venues in the neighbourhood. We might ask questions such as: Does the amount of food establishments in a neighbourhood increase the average rental price? Or what is the most important predictor for the average rental price? We will answer these questions and identify the most important types of venues affecting the rental price in a neighbourhood in order to make a quantitative prediction.

2 Data acquisition

We will use several datasets. First of all, we will identify different districts in Hamburg by their postal codes. Such geographical data can been downloaded from `https://opendatasoft.com`. This data includes a geoJSON file containing a list of postal codes and their area boundaries.

On top of that, we need the rental price for apartments per postal code. Such information is freely available on `https://miet-check.de` and can be obtained through a minimal amount of web scraping.

The rest of the data we used relating to the type and amount of venues, was obtained through the Foursquare API. See `https://foursquare.com`. By specifying a rectangular area and sending a request to the API, a set of venues including the name, latitudinal and longitudinal coordinates and a venue category are returned in a JSON file.

The venues categories defined by Foursquare follow a four-leveled tree structure. The ten top-level categories are: *Arts & Entertainment*, *College & University*, *Event*, *Food*, *Nightlife Spot*, *Outdoors & Recreation*, *Professional & Other Places*, *Residence*, *Shop & Service* and *Travel & Transport*.

All of the other venue categories fall under one of these umbrella terms. The precise structure can be obtained as a different JSON file by sending a request to the Foursquare API.

Once we find all venues through the Foursquare API, we have to remove duplicates. We do this by using the unique venue id that comes with a request to the API. Finally, using the coordinates and the geoJSON file, we can find the postal codes of the addresses of the venues.

It turns out that we find only a handful of venues in the *College & University*, *Event*, *Professional & Other Places* and *Residence* categories. Therefore, we drop these categories from our analysis completely.

3 A first look at the data

Before we start, we create a new feature known as the venue density by considering the total number of venues per postal code divided by the area. This feature will play an important role in our predictions. Using the total number of venues per postal code, we also define more features by considering the total number of venues in a certain category and dividing by the total number of venues. The ratio of venues in the *Food* and *Shop & Service* categories will play a major role in our predictions. It will be these seven features that we study.

As a first step in the data analysis, we look at the set of features we defined before and consider their correlations with each other and with the rental price. This can be shown succinctly in a heat map of the correlations. See figure 1.

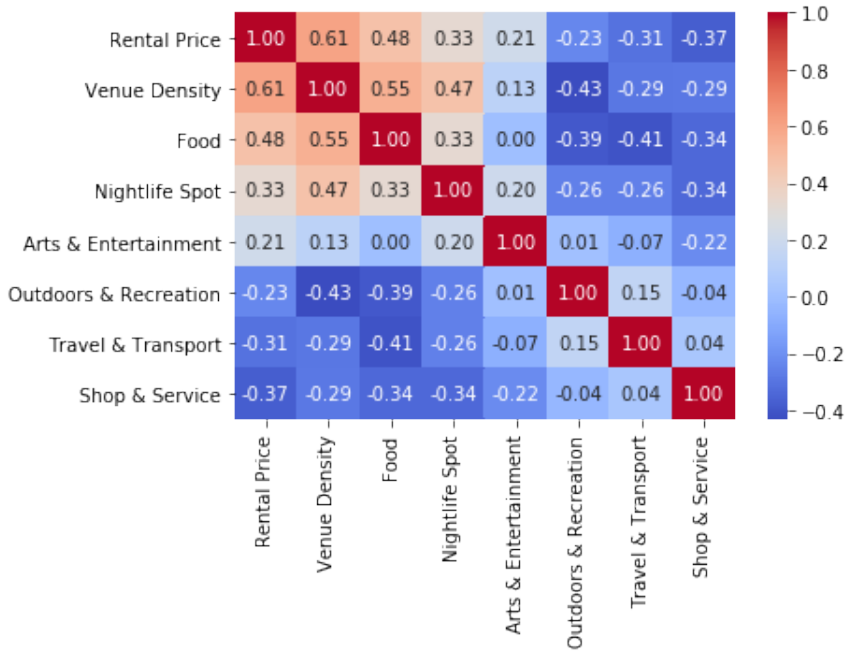


Figure 1: Heat map of correlations between features

The features that are most strongly correlated with the rental price, are *Food*, *Shop & Service* and *Venue Density*. It should be noted that *Food* and *Venue Density* are also strongly correlated.

3.1 Preliminary model

We see that the Kendall correlation between the rental price and the venue density is 0.61 showing a clear positive effect of the density on the rental price. Using a polynomial fit, we find a mean absolute error of approximately 1.08 Euro/m^2 and $R^2 \approx 0.67$. See figure 2. We will improve this model on the mean absolute error when we take into account more features using an ordinary least squares linear regression.

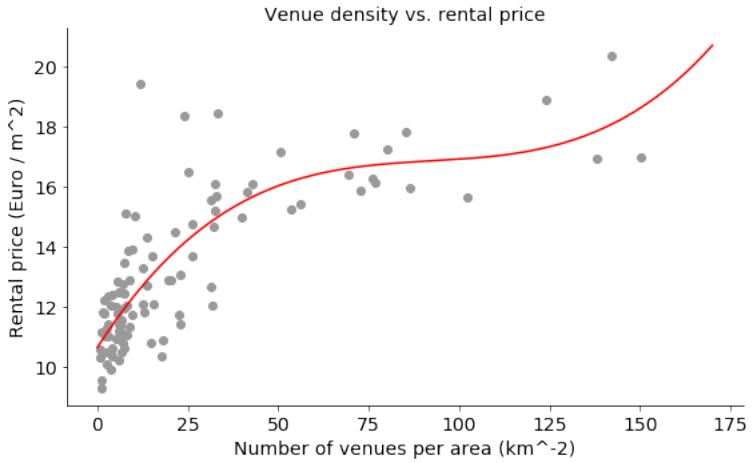


Figure 2: Venue density plotted versus rental price including a polynomial fit

This means that purely based on the venue density, we can already explain the majority of the variance in the rental price.

3.2 Including the main categories in the model

Looking at the fractions of venues in a specific category as in figure 3, we see that four out of six features have many postal codes with more than zero occurences of a venue in these categories.

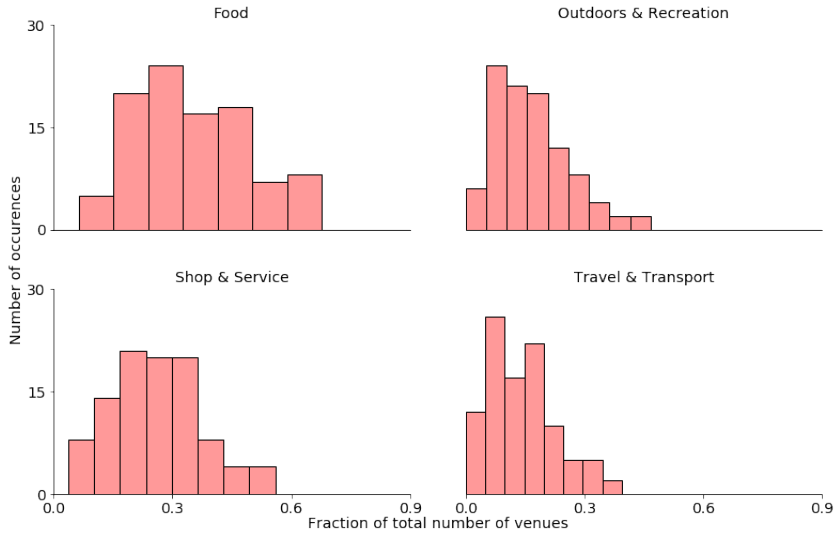


Figure 3: Histogram of categories *Food*, *Outdoors & Recreation*, *Shop & Service* and *Travel & Transport*

On the other hand, the histograms of the features *Arts & Entertainment* and *Nightlife Spot* show that there are a substantial number of postal codes with very few or no venues of these categories. See figure 4.

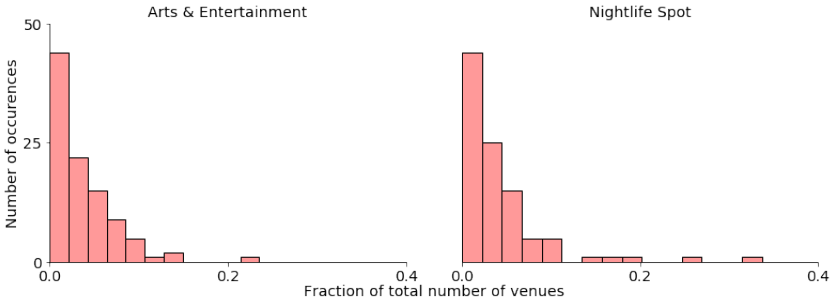


Figure 4: Histogram of categories *Arts & Entertainment* and *Nighlife Spot*

At this point, we have not much to say about these two features. We will focus on the categories that are more ubiquitous.

As we had already seen from the correlations, the most important features are *Food* and *Shop & Service*. We can make this more precise using LASSO regression.

LASSO regression is a regularization technique applied to a predictive model that is performed by adding a term to the loss function. This term is scaled by a parameter α, which has the effect of making some coefficients vanish if we increase it. This drops out the features from the regression completely, allow us to select the most important features in the analysis. The features that drop out at higher values of α are more relevant than those that drop out earlier.

By plotting the coefficients as a function of α, we find behaviour as in figure 5.

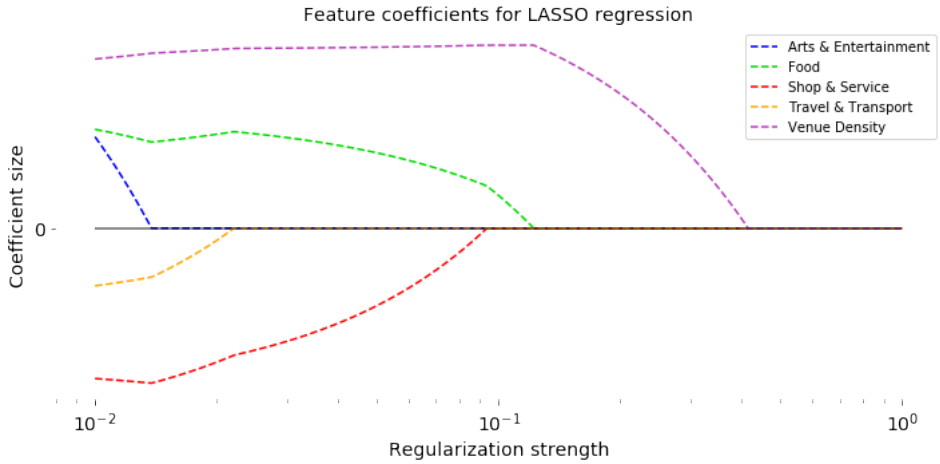


Figure 5: Plot of feature importance through LASSO regression

The features not appearing in figure 5 already vanished for $\alpha < 10^{-2}$. It can be seen that *Venue Density* is indeed the most important feature to consider, followed by *Food* and *Shop & Service*. It should be noted that at this stage it is hard to say whether *Food* or *Shop & Service* is more important.

If we fit the polynomial model from before with either *Food* or *Shop & Service* as an additional feature, we find that a larger increase in predictive power is provided by including *Shop & Service*, giving a value $R^2 \approx 0.78$. This should be compared to only including *Food*, which leads to a value $R^2 \approx 0.70$.

4 Building predictive models

4.1 Improving on the polynomial regression

We split up the total set of postal codes with features into a training set and testing set with a 90%-10% split. We consider a simple linear regression minimizing the squared error on the space of all seven features *Arts & Entertainment*, *Food*, *Nightlife Spot*, *Outdoors & Recreation*, *Shop & Service*, *Travel & Transport* and *Venue Density*.

This model leads to a mean absolute error of 0.98 Euro/m^2 on the training set. If we look at the test set, we find a mean absolute error of 1.66 Euro/m^2. This is a large difference, so it is worthwhile to look into the errors made by our model in more detail. See figure 6.

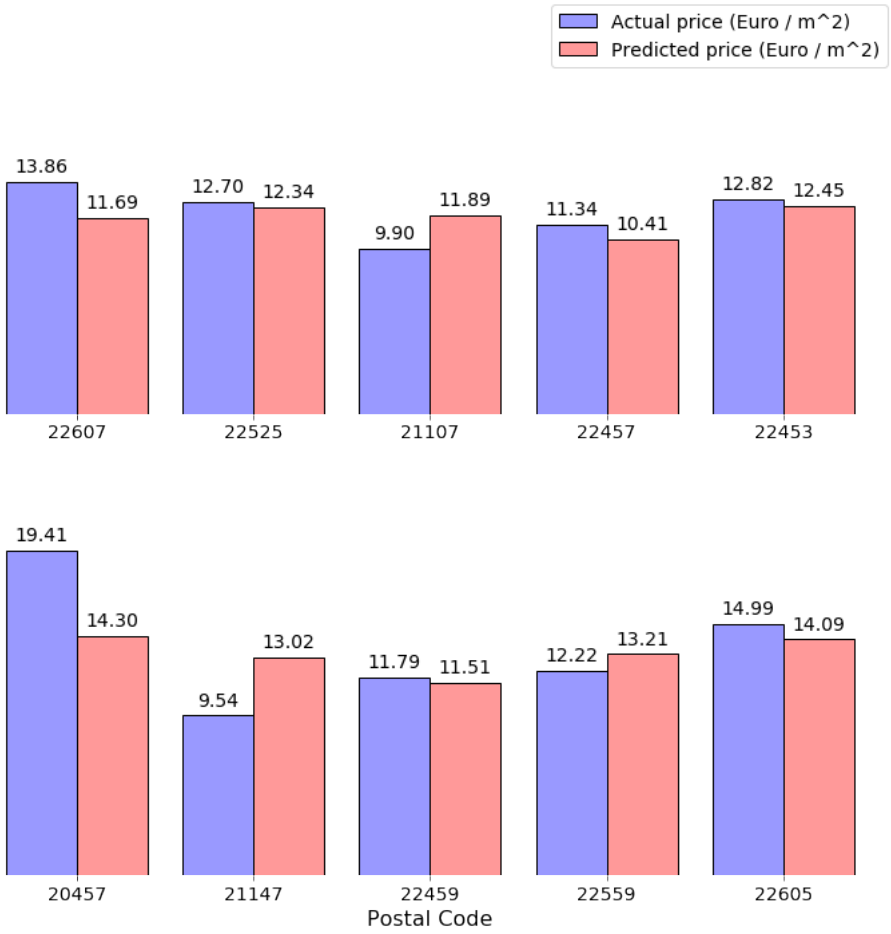


Figure 6: Actual price versus predicted prices for ten test cases

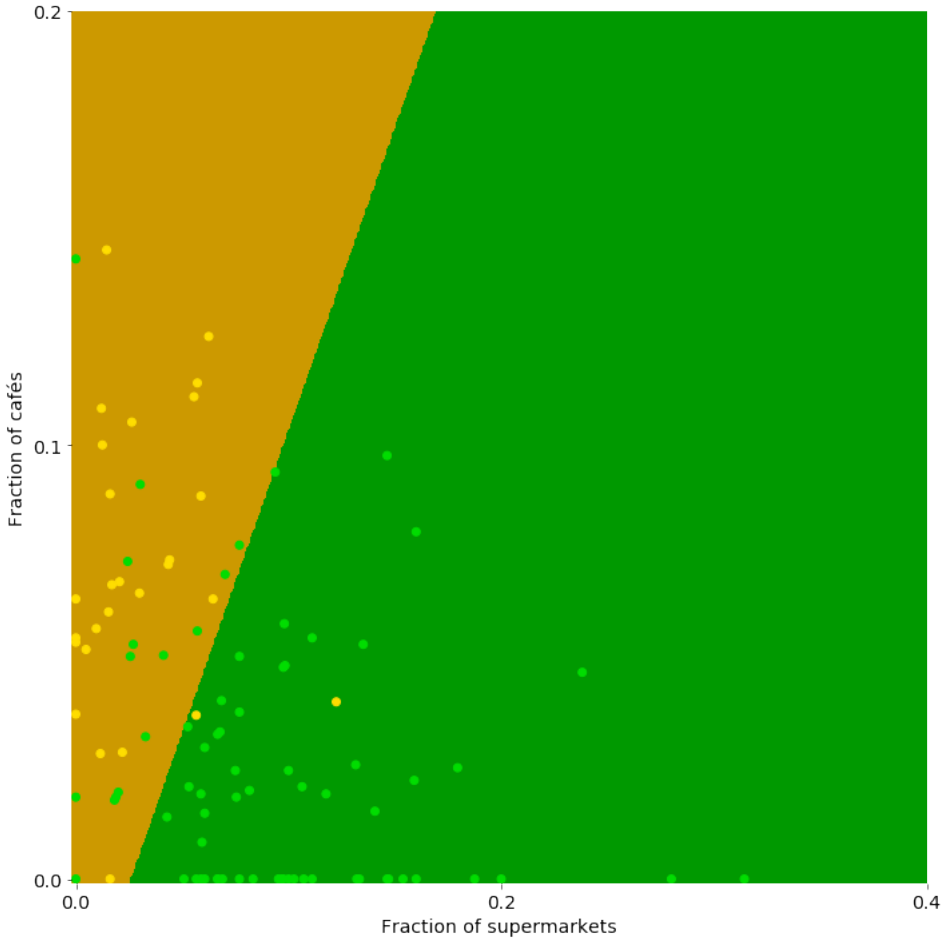
If we remove the entry with postal code 20457, the mean absolute error shrinks to 1.27 Euro/m^2, which is more in line with the error on the training set. This means there might be a couple of outliers that could heavily skew the results.

4.2 Including more categories?

Up until now, only the main categories have been considered. Could we improve our predictions further by including more categories that are not one of the main categories? As a toy example, we consider all other categories which appear at

least once in fifty or more postal codes and have a reasonable correlation with the rental price. Two of the strongest correlators with the rental price in this class are the ratio of cafés with correlation 0.39 and the ratio of supermarkets with correlation -0.47. The former belong to the *Food* category, while the latter belong to *Shop & Service*.

To get cleaner results for our toy model, we will split the rental prices into two categories: A high price category and a low price category where the cut off is $15\,\text{Euro}/m^2$. We use logistic regression to classify postal codes into one or the other category based purely on the ratio of cafés and supermarkets. We may plot the decision boundary of the logistic regression as in figure 7.



ix

Figure 7: Decision boundary logistic regression

To ensure the regression weighs each of the two categories equally heavy, we balance the class weights. In this case we are able to classify the prices correctly with an accuracy of 83%. Although this accuracy is high, the probability that any given postal code falls in one or the other category always lies between $40\% - 60\%$. We obtain our accuracy by placing the decision threshold at 50%. The uncertainty of our model reflects the fact that we have a number of outliers as can be seen in figure 7.

5 Discussion

5.1 Including extra categories

Our linear regression model was built on only seven features. To improve our model, we could take into account many more categories besides the top-level categories. Our toy example shows that these categories do hold a certain amount of predictive power. However, in these cases linear regression cannot be used as the most suitable model anymore, because the data becomes sparse. In other words, we would be including many more categories for which a lot of postal codes have zero or very few venues. The distributions of these categories would be comparable to the distributions in figure 4. One way to overcome this problem, is to regularize our results using LASSO regression or perform subset selection.

Furthermore, by one-hot encoding our data, we find a high-dimensional feature set. Of course, we can perform dimensionality reduction on the set of features, but another approach is to change our model. Models like gradient boosting for decision trees are more suitable to work with high-dimensional data.

In principle, including extra categories should have a positive effect on the predictive power of the model. We may use statistical tests to determine if adding a feature improves the model or not.

5.2 Improving the quality of the data

From our results, it is clear that the most improvements to our models can be made by improving the quality of the data. For example, let us consider the postal code 20457 in more detail.

This postal code belongs to the harbour city in Hamburg. The area spanned by this postal code includes both water and the harbour such that the residential area is significantly smaller by a factor of $\sim 5 - 10$. Taking such factors into account, would already bring significant improvements to the model, since the venue density is the strongest predictor of the rental price. See figure 8.

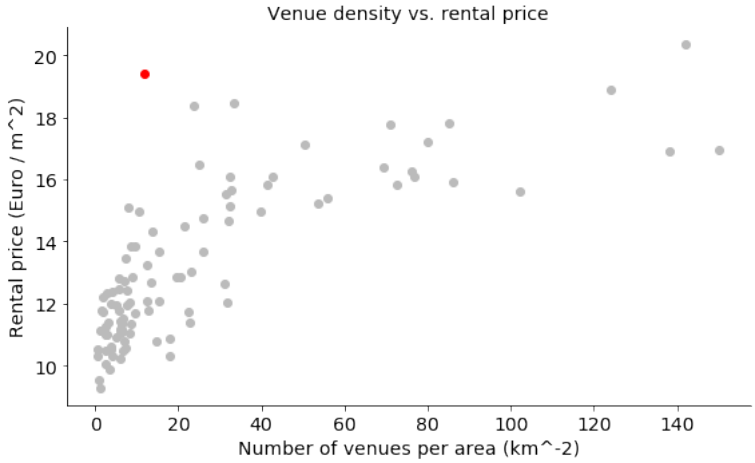


Figure 8: Venue density versus rental price including postal code 20457

There are of course other features that affect the rental price of a district. For example, the harbour city has a view on the Elbe. It would be expected that this would also increase the rental price.

Another way to improve the predictions, would be to include the cost of each venue. We would expect cheap restaurants to lie in cheaper neighbourhoods for example, and more expensive restaurants in the more expensive parts of the city.

5.3 Inherent variability of the rental price data

The rental prices in a given area might vary significantly. Especially when the areas are larger, this could become a limiting factor in our predictions. To get a better result, we would need to know the standard deviation of the average price. Moreover, it is possible that the geographical locations cluster based on rental price in a different way than the actual borders of the postal code areas. We might find slight improvement by taking these factors into account.

6 Conclusion

Most of the data analysis we have performed, is exploratory in nature. We have observed that from the features we considered, the venue density bears the strongest correlation with the rental price. Not far behind follow the ratio of food venues and the ratio of shops.

In a similar exploratory fashion, we have shown that it could prove useful to use more categories. For example, we have seen that only using the ratio of cafés and supermarkets is enough to classify neighbourhoods into expensive and cheaper neighbourhoods with an accuracy of 83%.

We conclude that a person wishing to move apartments in Hamburg would benefit greatly asking what type of neighbourhood or district they are willing to live in, because the density of venues as well as the type of venues, have a significant impact on the rental price in an area.