Report: Neighbourhood Classification

1. Introduction

1.1 Background

Each neighbourhood of a city has distinct properties that determine the atmosphere of the neighbourhood and determine whether it is considered to be a good or bad neighbourhood for distinct purpose (family life vs. shopping). These properties can be for example the different shop types or number of parks that, in addition, also attract a certain kind of customer. It can be therefore helpful when opening a new business in a new city or moving to a new city to investigate which neighbourhood will be best for ones purpose.

1.2 Aim and Interest

The aim of this project is to find neighbourhoods in a new unkown city similar to neighbourhoods in a known city in order to choose a neighbourhood to live in. This can be also useful for companies to investigate e.g. where to open a certain type of venue based on the neighbourhood data of a successful venue of the same type in another city.

In this particular project data from familiar neighbourhoods in Amsterdam, Berlin, Bonn, Frankfurt am Main, Rotterdam and Wuppertal categorised as 'boring', 'favourite' and 'too busy' is used in order to classify neighbourhoods in Zürich (unfamiliar city).

2. Data acquisition

2.1 Data Sources

As training data neighbourhoods were chosen with which I was familiar and which I categorized in three classes ('boring', 'favourite', 'too busy'). In total 29 neighbourhoods from 6 different cities (Amsterdam, Berlin, Bonn, Frankfurt am Main, Rotterdam and Wuppertal) were chosen. Latitude and longitude data of the neighbourhoods were extracted from the Wikipedia pages of the neighbourhoods as well as the longitude and latitude data of the central train station and central library of each city. The distance from the neighbourhood centre to the central train station and central library was calculated using haverine function.

The venue data was acquired using the FourSquareAPI and data for all venues in a 500 m radius of the neighbourhood centre was extracted using the latitude and longitude data.

2.2 Data Cleaning

The extracted FourSquare venue data was One-hot encoded and number of occurrence of each venue type per neighbourhood were saved in a table.

In the Zürich dataset (test data) venue types that were not present in the training dataset were deleted. Venue types that were present in the training but not in the Zürich dataset were added as columns containing zeros.

2.3 Feature Selection

The following features for each neighbourhood were used:

- Number of presences of a venue type (N=217 venue types)
- Diversity of venue types
- Distance to central train station
- Distance to central public library

The Diversity of venue types was calculated as the number of unique venue types present in a neighbourhood

3. Predictive Modeling

Using the data from the known neighbourhoods two classification models (Random Forest decision trees and K-Nearest-Neighbour) were generated to predict the class of the Zürich neighbourhoods as 'boring', 'favourite' or 'too busy' to determine in which neighbourhood to live.

3.1 Random Forest Decision Tree (RF)

A Decision Tree was trained as first model and, in order to improve the accuracy of the model, the Random Forest algorithm was used. In this algorithm multiple trees are generated which look very different from each other since during tree generation at each decision point only a random subset of features is used to determine the split.

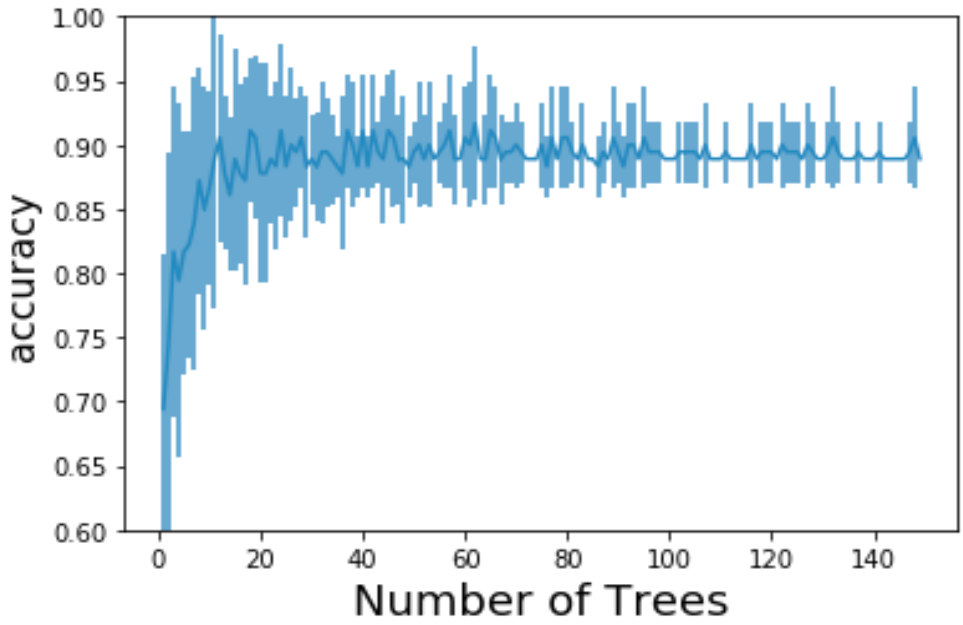

Figure 1: Accuracy of Random Forest models for different number of trees in model. Mean accuracy with standard deviation (20 repetitions with different training-test-data splits).

To train the Random Forest model the data was first split into 'training' and 'test' data (70% of data for training and 30% for testing). This was repeated 20 times and each time RF models with different number of trees (1-150 trees) were trained. The accuracy of each model was tested using the test data (Figure 1). Based on this analysis a model with 70 trees was chosen, since at this number the accuracy and standard deviation had stabilized.

In order to get an idea which features are important to determine the class of a neighbourhood the feature importances were extracted (Figure 2). 49 features showed an importance larger than 0, with 'distance to library' showing the largest importance (0.09). This shows that only subset of the data is crucial for determine the class of a neighbourhood.

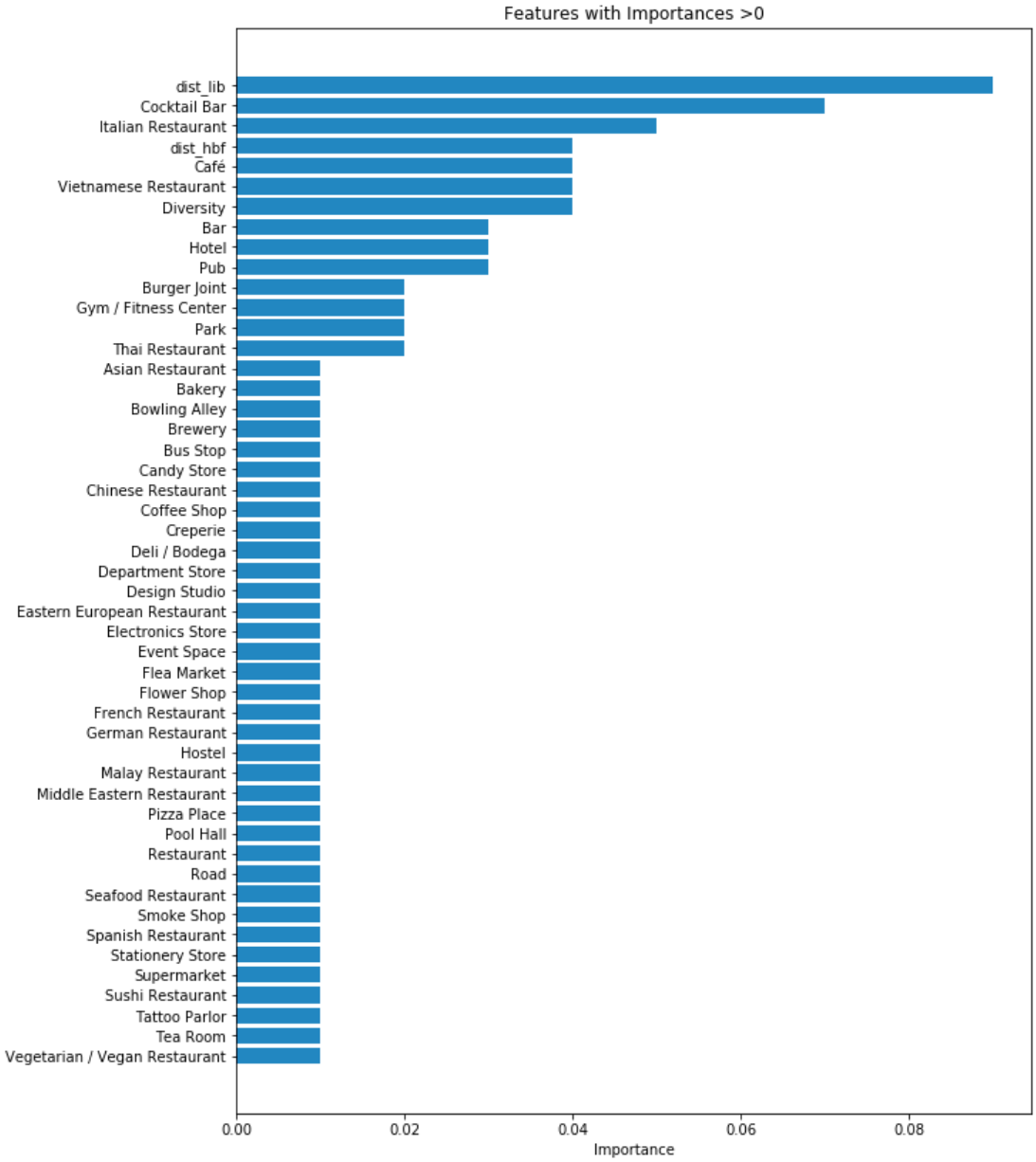

Figure 2: Feature importance of the 49 features that showed an importance of >0.

3.2 K-Nearest-Neighbour (KNN)

As a second model a K-Nearest-Neighbour Classifier was trained. For this the data was first normalized. First KNN models with different Ks were trained including all features showing a maximum average accuracy of 80% (Figure 3).

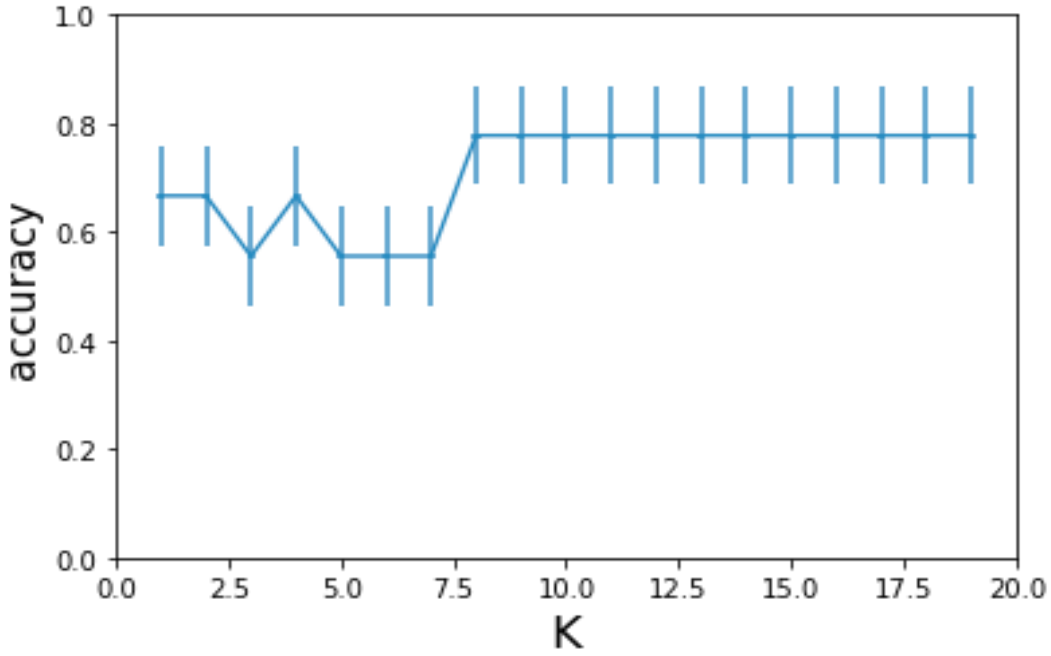

Figure 3: Accuracy of KNN models for different numbers of K using all features. Mean accuracy with standard deviation (20 repetitions with different training-test-data splits).

To improve the accuracy a new KNN model was trained using only the features that showed a >0 importance in the RF model (49 features). The accuracy improved up to a maximum of 100% accuracy when using a K of 9, 11 or 12 (Figure 4). For the model K=7 was chosen since it gave a good accuracy (89%). Smaller Ks could lead to overfitting of the model to the data and larger Ks are problematic due to a small dataset size (29 samples) with unequal number of samples per class.

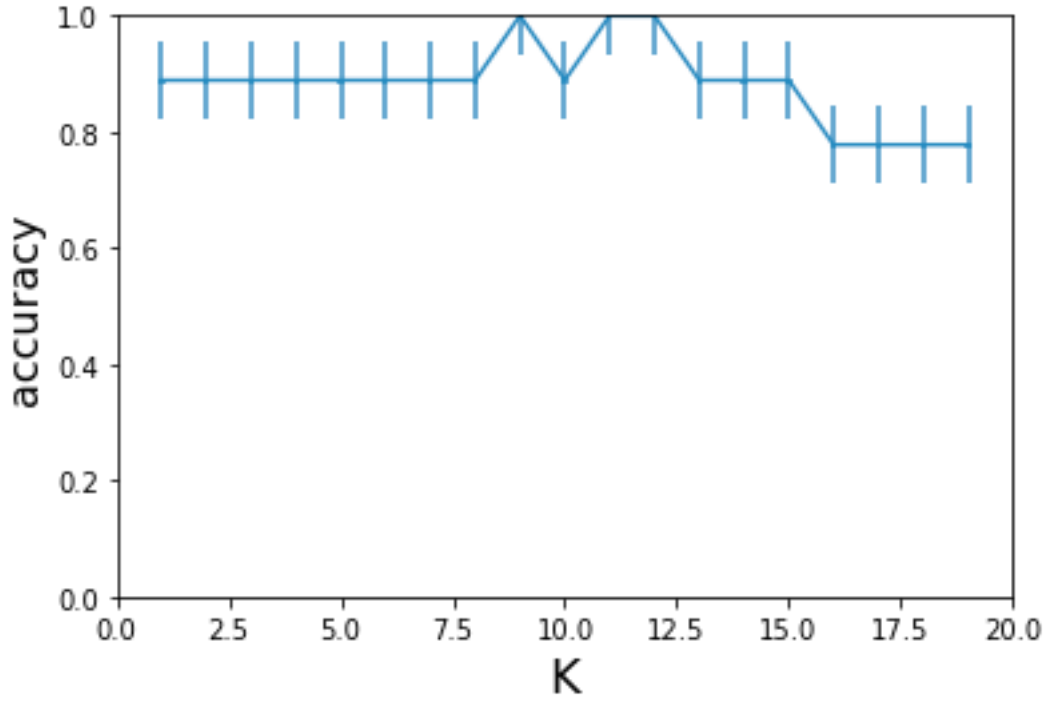

Figure 4: Accuracy of KNN models for different numbers of K using only features with an importance >0 in the RF model. Mean accuracy with standard deviation (20 repetitions with different training-test-data splits).

4. Results

The KNN and RF models were used to predict the neighbourhood class of the 12 neighbourhoods in Zürich. The RF model predicted that all neighbourhoods belong to the class 'boring', whereas the KNN model predicted Kreis 1, 4 and 5 as 'favourite' and therefore as neighbourhood to choose to live in (Table 1). None of the neighbourhoods were classified as 'too busy' by neither of the two model types.

Table 1: Classification of Zürich neighbourhoods by KNN and RF models. RF classifies all neighbourhoods as 'boring', whereas KNN classifies Kreis 1,4 and 5 as 'favourite'.

Neighbourhood	KNN prediction	RF prediction
Kreis 1	favourite	boring
Kreis 2	boring	boring
Kreis 3	boring	boring
Kreis 4	favourite	boring
Kreis 5	favourite	boring
Kreis 6	boring	boring
Kreis 7	boring	boring

Kreis 8	boring	boring
Kreis 9	boring	boring
Kreis 10	boring	boring
Kreis 11	boring	boring
Kreis 12	boring	boring

Taking a closer look at the probability with which the models predicted the neighbourhoods belong to a certain class, shows that the RF model was not as certain for Kreis 1, 4 and 5 as it was for the other neighbourhoods (Table 2A). The KNN model was in general more certain about the classifications of the neighbourhoods, showing for example a probability of 1 for Kreis 1 belonging to class 'favourite' (Table 2B).

Table 2: Class probability for Zürich neighbourhoods for Random Forest (RF) and K-Nearest-Neighbour (KNN) model.

A RF

Neighbourhood	boring	favourite	too busy
Kreis 1	**0.54**	**0.34**	**0.11**
Kreis 2	0.96	0.04	0.00
Kreis 3	0.70	0.21	0.09
Kreis 4	**0.53**	**0.30**	**0.17**
Kreis 5	**0.47**	**0.39**	**0.14**
Kreis 6	0.71	0.24	0.04
Kreis 7	0.83	0.17	0.00
Kreis 8	0.79	0.19	0.03
Kreis 9	0.97	0.03	0.00
Kreis 10	0.97	0.03	0.00
Kreis 11	0.96	0.04	0.00
Kreis 12	0.96	0.04	0.00

B KNN

Neighbourhood	boring	favourite	too busy
Kreis 1	0.00	1.00	0.0
Kreis 2	0.79	0.21	0.0
Kreis 3	0.59	0.41	0.0
Kreis 4	0.00	1.00	0.0
Kreis 5	0.13	0.87	0.0
Kreis 6	0.57	0.43	0.0
Kreis 7	0.69	0.31	0.0
Kreis 8	0.60	0.40	0.0
Kreis 9	0.67	0.33	0.0
Kreis 10	0.71	0.29	0.0
Kreis 11	0.79	0.21	0.0
Kreis 12	0.67	0.33	0.0

5. Conclusion

Two different classification models were used to classify the neighbourhoods of Zürich. The results from these two models show that most neighbourhoods belong to the class 'boring' and none to the class 'busy'. Only three neighbourhoods (Kreis 1, 4 and 5) were classified as 'favourite' by the KNN model and are therefore recommended when choosing a place to live in Zürich.

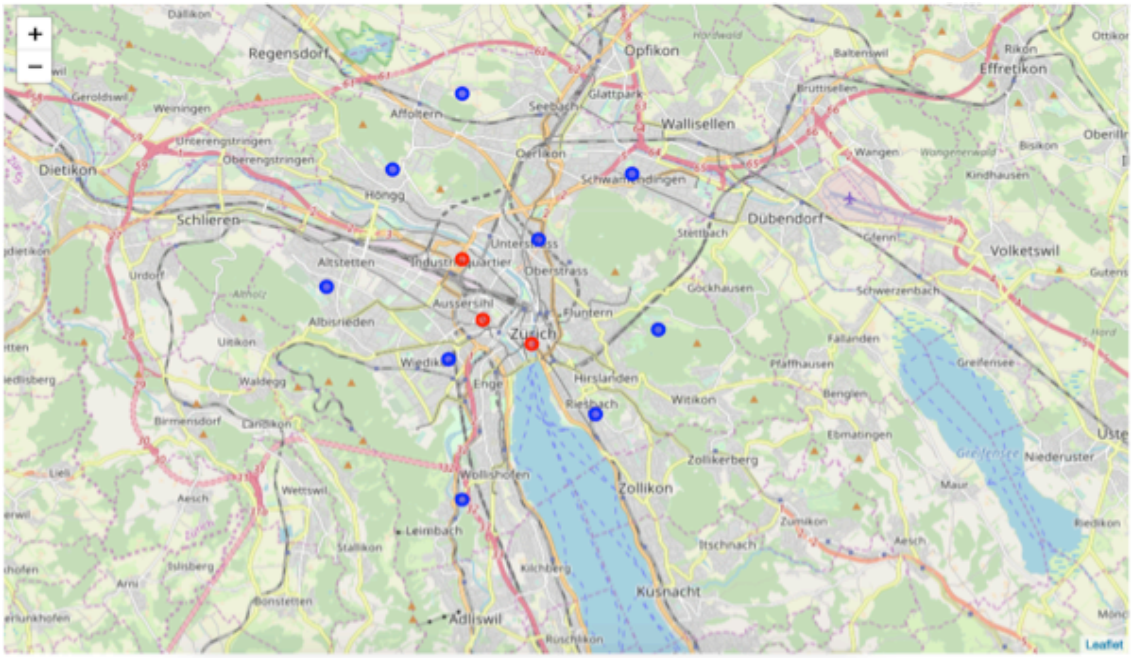

Figure 5: Map of Zürich with 'favourite' classified neighbourhoods in red and 'boring' classified neighbourhoods in blue.

6. Outlook

In order to improve the models one should acquire more data, especially for the class 'too busy'. In addition, more specific feature could be selected comparable to the features 'distance to library' or 'diversity', since these had a high importance score in the RF model. Furthermore feature selection already showed an improvement for the KNN model and it would be therefore recommended to further reduce the feature space e.g. by Principal Component Analysis to extract the most descriptive/important features of the data.