PREDICTING NEIGHBOURHOOD TO LIVE IN

COURSERA CAPSTONE PROJECT

PROBLEM

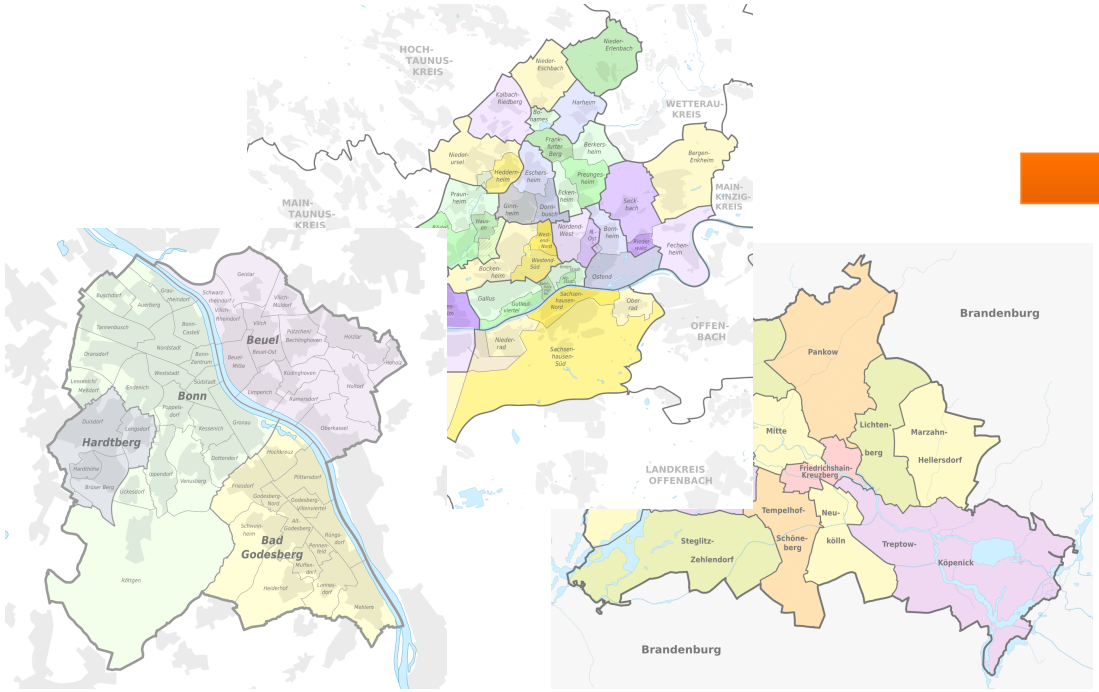
- You are moving to a new city and have to decide where to look for a flat

- Certain neighbourhoods of other cities you like a lot (‘favourite’) and other you find ‘boring’ or ‘too busy’

→ How can you determine in which neighbourhood you should look for a flat?

AIM

Find the neighbourhood to live in the new city based on data from neighbourhoods in other cities that you know



DATA ACQUISITION

- Neighbourhoods latitude and longitude was extracted from the Wikipedia pages of the known neighbourhoods in Amsterdam, Bonn, Berlin, Frankfurt, Rotterdam and Wuppertal (total of 29 neighbourhoods) and the 12 neighbourhoods of the unkown city Zürich
- Venue data in 500 m radius of neighbourhood centre extracted using FourSquare API

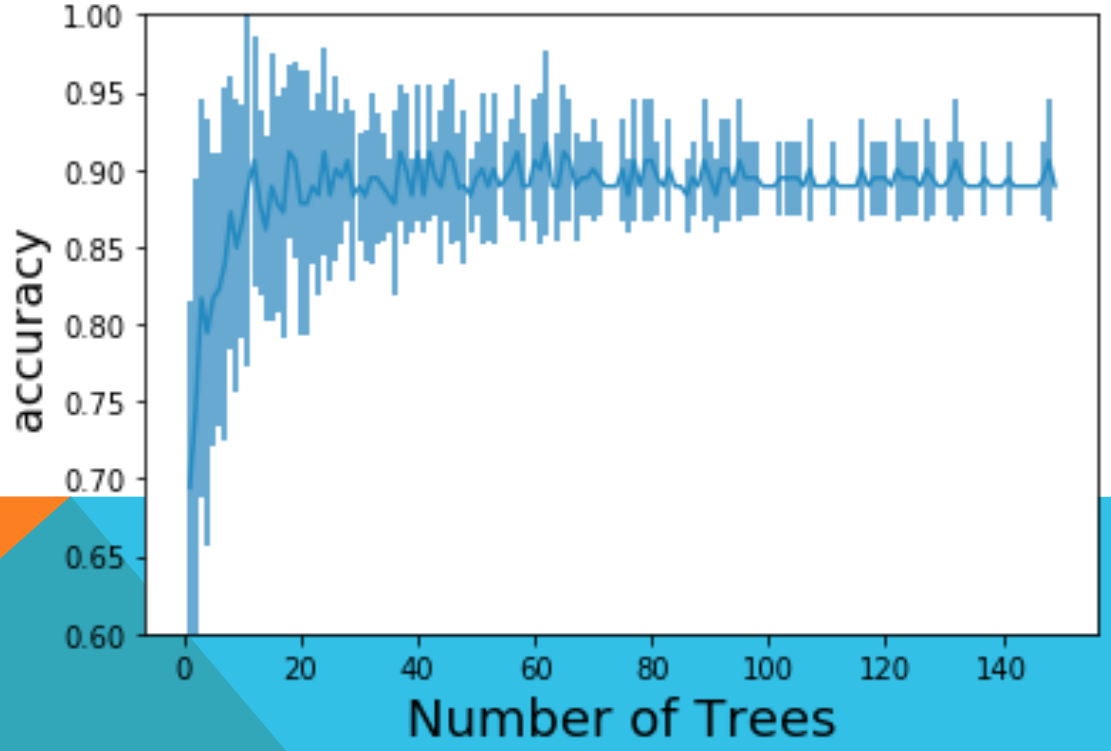
Features:

- number of each venue type present in neighbourhood (217 different venue types)
- number of different venue types per neighbourhood (diversity of venues in neighbourhood)
- distance to important venue types (central train station and central public library)

METHODS

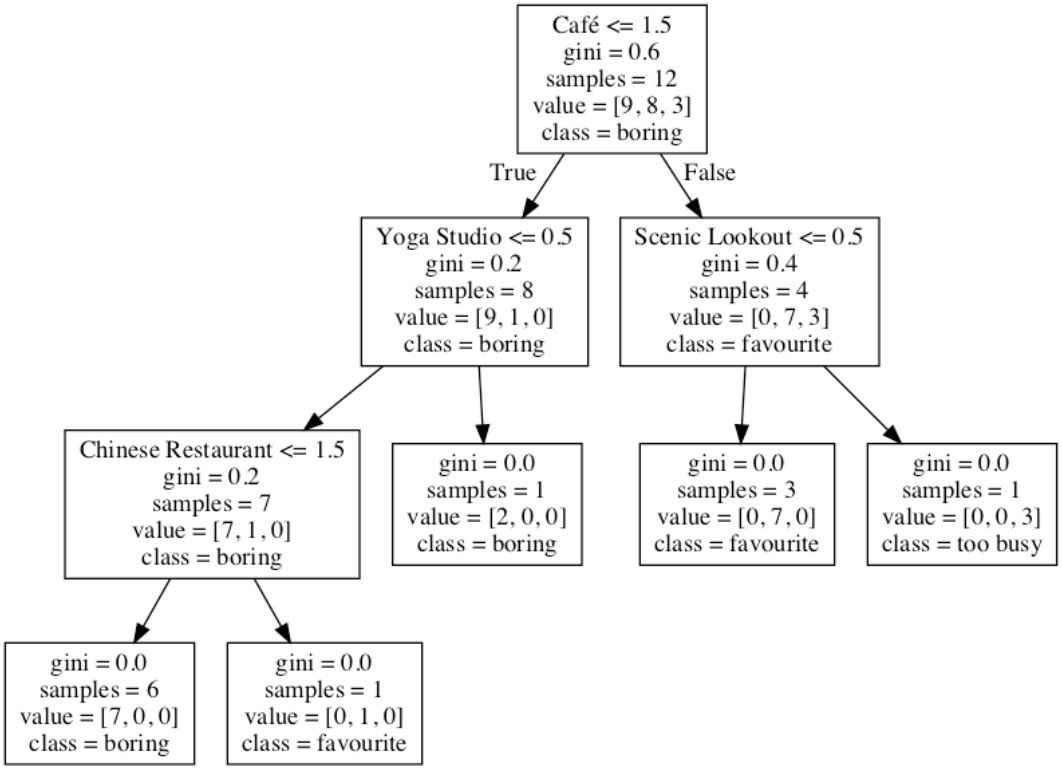
Random Forest Decision Tree (RF)

- A random forest of 70 trees was build based on analysis of the accuracy of different forests sizes taking into account that models with less trees are faster
- Accuracy =89%



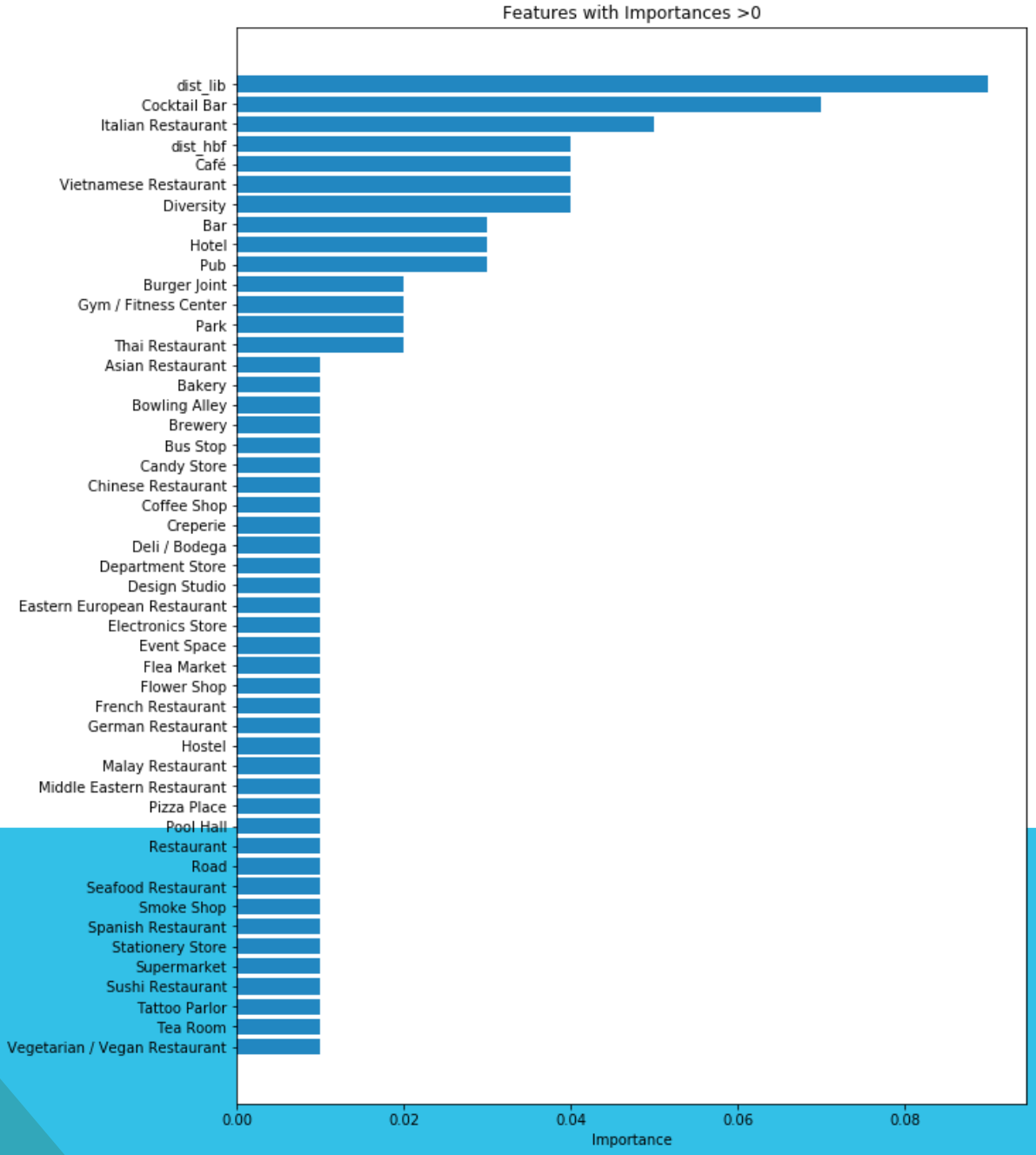
METHODS

Example Tree from Forest



METHODS

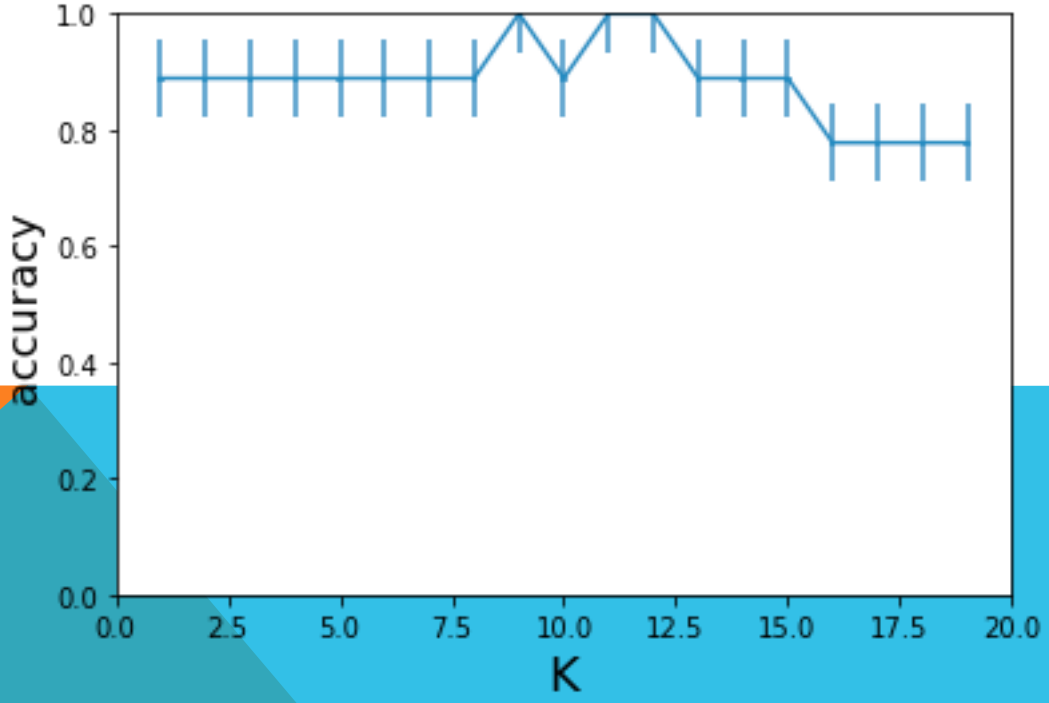
- Feature importances where extracted from the Random Forest model .
- 49 features showed an importance >0



METHODS

K-Nearest-Neighbour (KNN)

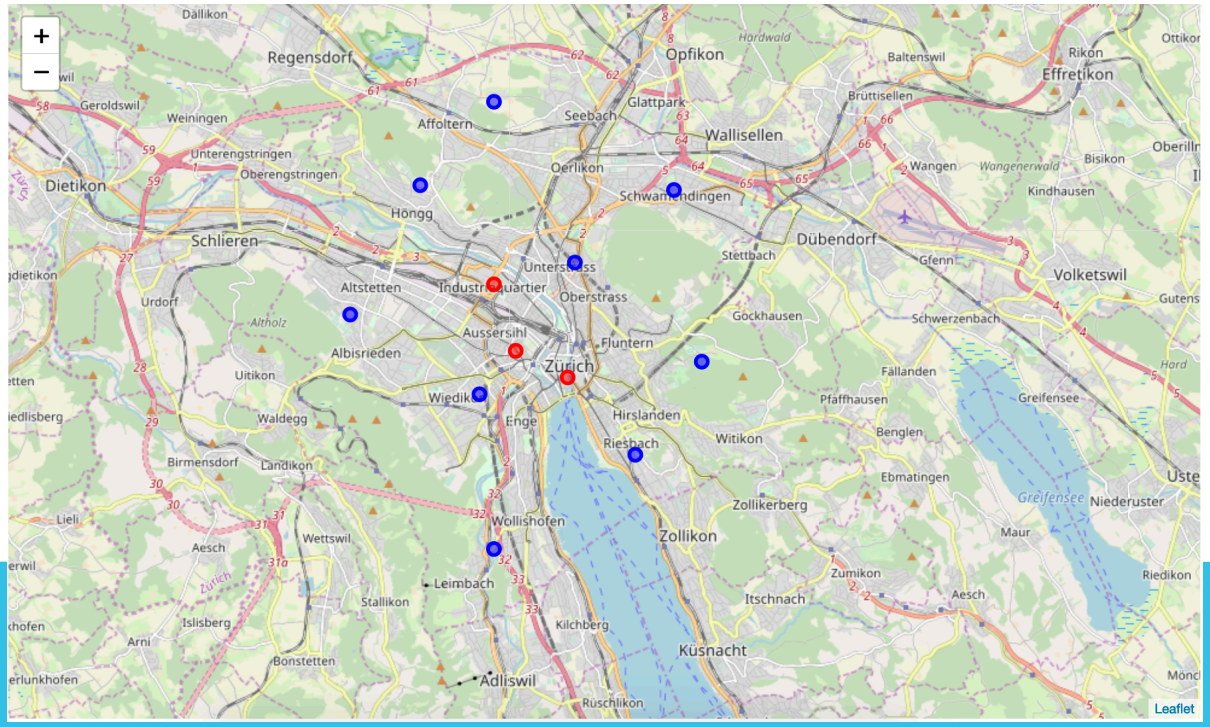
- To reduce the number of features used in the KNN model to improve its accuracy only feature with a feature importance of >0 from the RF model were used
- For the model K=7 was used since it gave a good accuracy (89%), smaller Ks could lead to overfitting of the model and larger Ks are problematic due to a small dataset size (29) with unequal number of samples per class



RESULTS: DIFFERENT CLASSIFICATION WITH RF AND KNN

- KNN: 3 neighbourhoods were classified as ‘favourite’
- RF: All neighbourhoods in Zürich were classified as ‘boring’
- Non of the neighbourhoods was classified as ‘too busy’

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



RESULTS: CLASS PROBABILITY

KNN ‘favourite’ classified neighbourhoods show comparably more distributed RF class probability compared to the other neighbourhoods
→ Less certainty for belonging to class ‘boring’

RF class probability

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

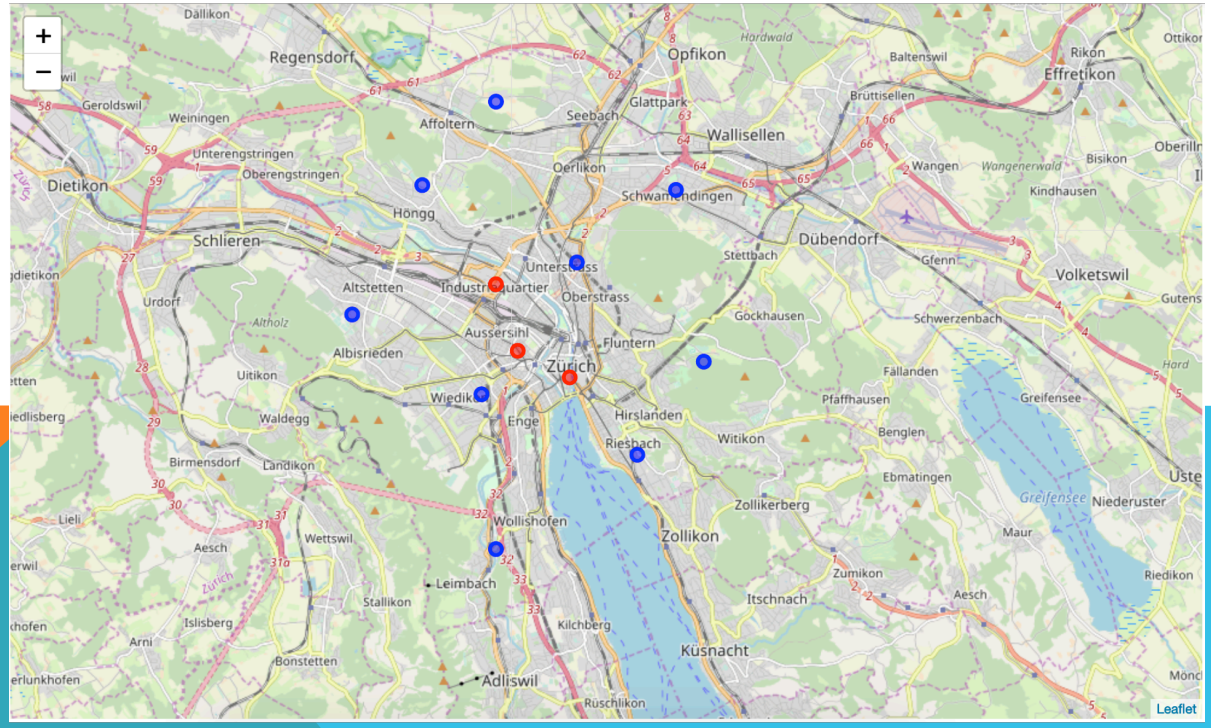
KNN class probability

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

CONCLUSION & SUMMARY

- The Zürich neighbourhoods Kreis 1, 4 and 5 are according to KNN model neighbourhoods where one should look for a flat ('favourite')
- These neighbourhoods were classified by RF model as boring, however the probability values revealed that the model was least certain about that decision

→ Look for a flat in Kreis 1, 4 and 5



OUTLOOK

The Models can be further improved by:

- Adding more training data, especially for ‘too busy’ class
- Adding features that might be more specific for certain classes like ‘distance to central station’ since these features showed a high importance
- Using dimensionality reduction (e.g. PCA) of feature space to extract most descriptive/important features
- Further optimize hyperparamters of the models

APPENDIX

Sources Images pages 2-3:

https://de.wikipedia.org/wiki/Stadtteile_der_Stadt_Z%C3%BCrich#/media/Datei:Karte_Stadtquartiere_Stadt_Z%C3%BCrich.png

https://de.wikipedia.org/wiki/Bonn#/media/Datei:Bonn_Subdivisions.svg

https://de.wikipedia.org/wiki/Frankfurt_am_Main#/media/Datei:Frankfurt_Subdivisions_boroughs.svg

https://de.wikipedia.org/wiki/Datei:Berlin,_administrative_divisions_(%2Bdistricts_-boroughs_-pop)_-_de_-_colored.svg