Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

# Smart Real Estate Value Estimation

Bachelor Thesis


Linus Handschin

hlinus@student.ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich


**Supervisors:**
Pascal Bissig, Philipp Brandes
Prof. Dr. Roger Wattenhofer

September 7, 2016

# Acknowledgements

I would like to thank Linus Schenk from the startup *Avisum*[1] for the proposal of this thesis, the provided data and all the advice given in our meetings.

[1] Avisum: www.avisum.ch

# Abstract

In this thesis, we examine affects of close to a thousand factors on rental fees of apartments in Switzerland and develop two models for price estimation. The factors include features of the apartment, building, jobs, demographic structure, point of interests and traffic noise level in the neighborhood. We show that the neighborhood of this data has to be chosen large enough in order that some factors become influential. With the exception of the number of rooms, the most relevant feature for price estimation are location based. Jobs in finance and insurance services, corporate leadership and information services show a positive while workplaces in the primary sector a negative impact on rental fees. Among the building features solar hot water panels and among the point of interest the density of banks and public transport stations increase the attractiveness of a location. It also shows that it is difficult measure the effect of such a large number of factors because of correlation. Especially population factors such as population density showed a high correlation with many factors.
From the chosen features we develop a regression and a artificial neural network based model for price estimation. The regression model explains more than 80% of variance in rental fees in Switzerland with around 14% prediction accuracy while the error of the neural network based model is 12%.

# Contents

Chapter 1

# Introduction

Everyone who ever looked for an apartment knows it can be very time consuming. There is a plethora of internet platforms on which you can spend hours browsing through thousands of offers. Once you found a suitable apartment offer you have to compare rental fees of similar offers. But you can not only simply compare the living space, number of rooms etc, but also have to take into account the neighborhood. Location factors can have a big influence on the attractiveness. One apartment may has a supermarket close by but the other has better transport connections. Is it a quite residential area with many older people or very busy area and young folks around?
The core of this thesis is the examination of more than 800 location factors and their influence on rental fees. Those include the types of buildings, workplaces, the demographic structure and traffic noise in the neighborhood of an apartment. Using this information and the housing offers available on internet portals we can build an accurate model to estimate rental fees in Switzerland.

## 1.1 Related Work

This is the third project of real estate price estimation in Switzerland at ETH Zurich. The first iteration was in 2014 [1] and took only into account apartment features given in the apartment offers. The square meter price estimation was based on averaging offers in the same area. In 2015 [2] a second iteration created multiple linear least squares regression models for apartment rental fee and house prices estimation. It studied around 25 features about the apartment, population and some point of interests. Among the most influential location based features was the population density, percentage of foreigners and per capita income on municipal basis. The distance to the city center, lake and university also showed high significance while most of the apartment features did not.
Hedonic price models have also been developed for inner city of Stockholm [3] and Uppsala [4], Prague [5] and Abuja Satellite Towns (Nigeria) [6]. Those models considered mainly apartment features and a few location factors such as road access, crime rate, distance to the city center, etc. Also compared to this work

they only model the price for a specific city. Recently a hedonic price index for houses has been developed for Turkey [7] but does not consider location influences.
Further there also has been developed artificial neural network (ANN) based models for real estate price estimation. One has been developed for houses in Taranto (Italy) [8] and takes into account environmental factors. Another study has been done for houses offers collected over several year in Wake County (USA) [9]. The authors show in this paper that ANN based models are a practical alternative to conventional least squares forms and can efficiently model complex nonlinearities. They also argue that those models are better suited to hedonic models because ANN do not depend on the rank of the regressor matrix. Categorical variables such as location, construction material, etc, do no belong in the regression function except as dummy variables. They show that this increases the likelihood of rank failures and can make an ordinary least squares estimate impossible.

Chapter 2

# Micro and Macro Factors

Before our model can be built, the features from different data sources have to be prepared, analyzed and merged to a single dataset. In addition to the apartment features, the factors from the surrounding area such as population, buildings, jobs, point of interests and traffic noise are considered by our analysis.

## 2.1 Apartment Features

To be able to measure the performance of our model we need a ground truth to compare our predictions. This is taken from ads shown on Immoscout24.[1] This website offers a large amount of current data with a few hundred new entries every day. The crawling is done using a Python framework called Scrapy.[2]
To geocode the address of the apartments we use the Google Maps Geocoding API.[3] The geocoding is accuracy and performs well even without requiring the address having a certain format. The standard usage a is limited to 2'500 free requests per IP per day which is enough for our purpose.
The apartment features taken from the Immoscout24 ads are:

- living space
- number of rooms
- Net and gross rent
- floor
- number of bathrooms
- build and renovation year
- wheelchair accessible*
- pets allowed*
- view*
- chimney*
- attic*
- cellar*
- Minergie standard*
- dishwasher*

[1] Immoscout24: http://www.immoscout24.ch/
[2] Scrapy: http://scrapy.org/
[3] Google Maps Geocoding API:https://developers.google.com/maps/documentation/geocoding/start

- elevator available*
- balcony or terrace*
- parking*
- garage*
- old building*
- apartment in residential area*

Features indicated with a star are binary, i.e. indicating whether the feature is available.

## 2.2 Geodata

We call the first set of statistics Geodata and it includes information about the populations, jobs and buildings. The data was collected by the *Bundesamt für Statistik* (bfs) on behalf of the Swiss government. The data is available as an incomplete *100 m* by *100 m* grid across the country.
To use Geodata in our price estimation we map each apartment to the closest STATENT, GWS and STATPOP grid point. Since the grid point is identified by an unique ID we can easily switch between different $\mathbb{D}$-neighborhoods.

### 2.2.1 Population Data (STATPOP)

The population and household statistics (2014)[4] has 341′890 data points with 77 features about:

- Population and structure (age, sex, marital status, nationality, etc.)
- Structure of the household (area, composition, etc)
- Development of the population (births, movements, etc.)
- Spatial distribution of the populations and households

### 2.2.2 Jobs Data (STATENT)

The job statistics (2013)[5] has 208′358 data points with 619 features about:

- The structure of the Swiss economy (number of business, workplaces and full time equivalents (FTEs) for different job categories)

[4] STATPOP: http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/statpop/01.html
[5] STATENT: http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/statent/01.html

### 2.2.3 Buildings and Apartment Data (GWS)

The building and living statistics (2014)[6] has 385′799 data points with 155 features about:

- Building information and category (age, heating technology and energy source, number of apartments, number of floors and rooms, etc.)

## 2.3 Point of Interests

A major role for the attractiveness of a location may also be point of interests close by. *Avisum* provided a large data set of those and their location. The different kinds and counts are:

- public transport stations [28'040]
- pharmacies [1'427]
- banks [2'845]
- supermarkets [3'424]
- gyms [1'132]
- convenience shops [1'064]
- restaurants [11'616]

For each kind of point of interest we consider their count within the distances 100 m, 200 m, 500 m, 1000 m, 2'000 m and 10'000 m of the apartment as a feature.

## 2.4 Road Traffic Noise

Another important factor we consider is traffic noise during the day. An apartment located directly at a busy and noisy road is unpleasant and should therefore be cheaper. The traffic database *sonBASE*[7] was created in 2009 as a one-time project by the *Bundesamt für Umwelt (bafu)* and is freely available for download. The model basis are traffic counts and model calculation taking into account the type of vehicle, noise barriers, velocity, etc. of over 72'000km of roads. The noise level in decibel is computed across the country as a 30 m by 30 m grid and can easily be mapped to the apartments.

[6] GWS: http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/gws/01.html

[7] sonBASE: http://www.bafu.admin.ch/laerm/10312/10340/index.html?lang=de

Figure 2.1: Road traffic noise model by day taking into account 72'000km of roads. Data resolution is a 30 m by 30 m and the data is given in decibel (red - noisy, light - quiet).

## 2.5 Data Cleaning

All data except for the apartment offers collected from Immoscout24 are highly accurate and do not need any cleaning. One of the problem encountered are ads for shared apartments with the price given for one room but the description for the whole apartment. We therefore set the following validity constraints to exclude most of single room offers:

- Gross rent must be larger than 500 CHF
- living space more than 15 $m^2$

The other main problem were typos, missing and impossible values causing strange prediction behavior. We therefore set the following constraints:

- size per rooms must not be smaller than 10 $m^2$ per room
- number of rooms must be given and larger 0
- the given floor must be between 0 and 100
- the build year must be between 1700 and 2018

- price per area must be larger 15 CHF/$m^2$

Further some apartment offers are limited to a certain time frame since the building will be renovated or demolished in the near future and therefore the rental fee is lower. We tried to exclude those offers manually by searching for certain keywords.
Additionally, we require that the found location by the geocoder must be in Switzerland. As the Geodata grids is incomplete we also have to guarantee a grid point close by. Figure 2.2 shows that most distances between the closest grid point and the apartment is less than a 100 m. Since the data showed high similarity between neighboring grid points we allowed a distance up to 400 meter to the closest grid point for STATENT, STATPOP and GWS.

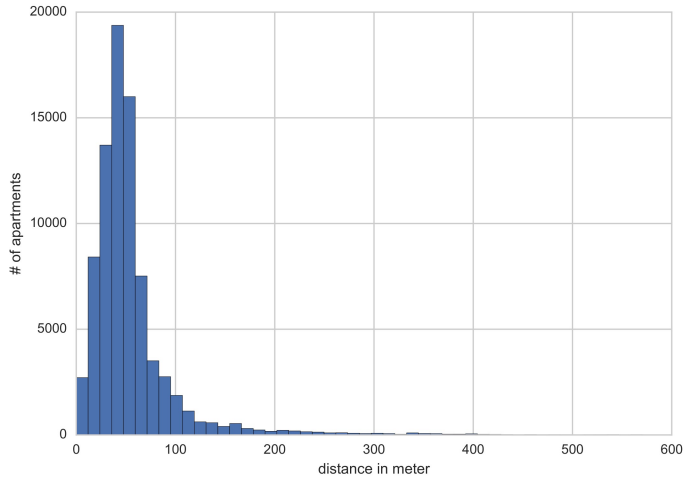


Figure 2.2: Distance between all apartments and the closest grid point of STATENT data set. Even though the grid is incomplete most of the apartments are closer than a hundred meters from the closest grid point available. The distances for STATPOP and GWS are comparable.

Out of 90'000 collected apartment offers we have to mark close to 50'000 offers invalid due to the above restrictions.

Chapter 3

# Methods

In this chapter we describe the methods used for finding the most influential features and describe the models used. In a feature preprocessing step dummy variables have to be created for some features. We show a method to measure the goodness of fit of a model and perform feature selection. We present and explain a baseline and two other models for price estimation.

## 3.1 Dummy Variables

For some very sparse and non-linear feature we have to create dummy variables. A dummy variable has value 1 if the apartment has this feature and 0 if not. We decide to apply this method to the features: rooms, floor, build and renovation year based on the their distribution as shown in Figure 3.1. We decide to create dummy variable for the values 1,2,3,4,5,6 and more than 6 for rooms, 0,1,2,3,4,5,6 and larger 6 for floors. For build and renovation year we divide the years in ranges. We divide build year into the ranges [1600-1800), [1800-1900), [1900-1950), [1950-1970), [1970-1990), [1990-2000), [2000,2005), [2005,2010), [2010,2012), [2012,2014), [2014,2016). For simplicity we use the same ranges for the renovation year.

## 3.2 Feature Selection

We perform feature selection on our input data which helps solving the problem in the following ways:

1. Ignoring features with low significance makes the model smaller and therefor easier to understand.

2. Removing features with low significance may help against overfitting and reduces multicollinearity.

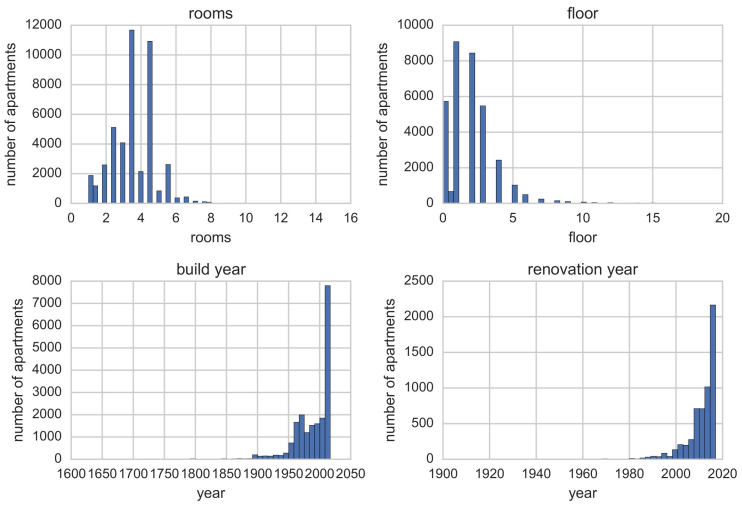


Figure 3.1: Distribution of feature for which dummies are used.

3. The low significance features induce greater computational cost. Since we mainly focus on regressions the impact is small but not negligible.

Since we are dealing with close to a thousand features we decided to perform greedy forward selection which can be computed fast but may not choose an optimal solution. We start with an empty feature set and repeatedly adding the feature which gives us the best estimator model with respect to our input data. For this we first have to define how to measure the performance of an estimator in mathematical terms. Let $y = [y_1, y_2, \ldots, y_n]$ be the prices of $n$ apartments as a vector and $\hat{y} = [\hat{y_1}, \hat{y_2}, \ldots, \hat{y_n}]$ the corresponding estimates of our model. The mean $\bar{y}$ of our data is given as:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The total sum of squares is defined as:

$$SS_{tot} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})$$

The sum of squares of residuals is defined as:

$$SS_{res} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

The coefficient of determination $R^2$ is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

It is easy to see that the $R^2$ is the percentage of the $y$ variation that is explained by our model. Therefore our best estimator has the highest $R^2$ and in our greedy forward selection algorithm, in each step we select the feature increasing the $R^2$ value most.

## 3.3 Baseline

We propose a very naive approach to compare our models to. As a baseline model we estimate the rental fee $\hat{y_i}$ by multiplying the corresponding living space with the average square meter price of all apartments in our dataset:

$$\hat{y_i} = s_i \cdot p_{avg}$$

where $s_i$ is the living space of apartment $i$ and $p_{avg}$ the average square meter price of all apartments in the data set.

## 3.4 Multiple Linear Least Squares Regression (MLLSR)

Since previous work [2] has shown that good results can be achieved using a linear model and because it allows us see the contribution of each feature we also decided to use a Multiple Linear Least Squares Regression (MLLSR) model. MLLSR has two major assumptions:

1. **Linearity** - The response variable, in our case the apartment price is a linear combination of the regression coefficients
2. **Lack of multicollinearity** - The absence of features which are highly linearly related

Since we are dealing with a large number of features and multicollinearity is unavoidable we have try to confine it by only selecting a subset of features.Since the living space is highly non-linear we decide to estimate the price by square meter (CHF/$m^2$) by solving a MLLSR problem. The rental fee estimate is then obtained by multiplying the square meter estimate with the living space of the corresponding apartment.
More formally, given an apartment $i$ with living space $s_i$, input variable vector $X_i$ the corresponding rental fee $y_i$ can then be estimated as:

$$\hat{y_i} = s_i \cdot (\sum_{j=1}^{M} a_j X_{ij}) + b$$

where $a_j$ are the regression coefficient of feature $j$, $b$ the intercept and $M$ the number of features.

## 3.5 Artificial Neural Network Regression (ANNR)

Our third model is based on a Artificial Neural Network. They can learn complex (non-linear) relationships directly from the observed data. Hedonic regression models rely on categorical values or counts. Dummy variables such as the binarized number of rooms or build year and strongly correlated features such as the number of people for different age categories can lead to rank failure and make an ordinary least squares regression impossible [9]. Neural Networks are in this sense more robust. As a downside Neural Networks results are harder to interpret compared to hedonic regression models in which each feature contribution is isolated. We therefore only use this method to compare the prediction performance of linear model against the ANN based model.

Chapter 4

# Results

For the following results we use a data set of 40'000 apartments fulfilling the constraints defined in Section 2.5. We split the data randomly in a training set of approximately 30'000 and a test set of around 10'000 apartments.
The prediction accuracy is measured by the mean absolute percentage error (mape) which is defined by the following formula:

$$E = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

where $y_i$ is the actual value and $\hat{y}_i$ is the estimation value of apartment $i$ and $N$ the dataset size.
Since we are working with different models we perform a Min-Max feature scaling. We rescale the features to the range in [0, 1] by the following transformation:

$$x' = \frac{(x - min(x))}{(max(x) - min(x))}$$

where $x$ is the original value and $x'$ the normalized value.

## 4.1 Optimal Neighborhood Size

Inspection of our data showed that the 100 m by 100 m grid data points only show a very small diversity even between very distinct locations. For example Figure 4.1 shows comparable population density of the most populated cities Zurich and Geneva and the rest of Switzerland. To overcome this problem we aggregated adjacent grid points to a larger grid. We can define the computation of a $\mathbb{D}$ by $\mathbb{D}$ grid as follows:

*Definition* 1. We define a $\mathbb{D}$-neighborhood the aggregation of Geodata grid points to a larger grid where $\mathbb{D}$ is a distance in meter and $\mathbb{D} \geq 100m$ as:

$$\forall \hat{p} \in \text{grid} : p = \sum_{D_{Chebyshev}(p,\hat{p}) \leq (\mathbb{D}-100m)/2} \hat{p}$$

$$\text{where } D_{Chebyshev}(p, \hat{p}) := \max(|p_x - \hat{p}_x|, |p_y - \hat{p}_y|)$$

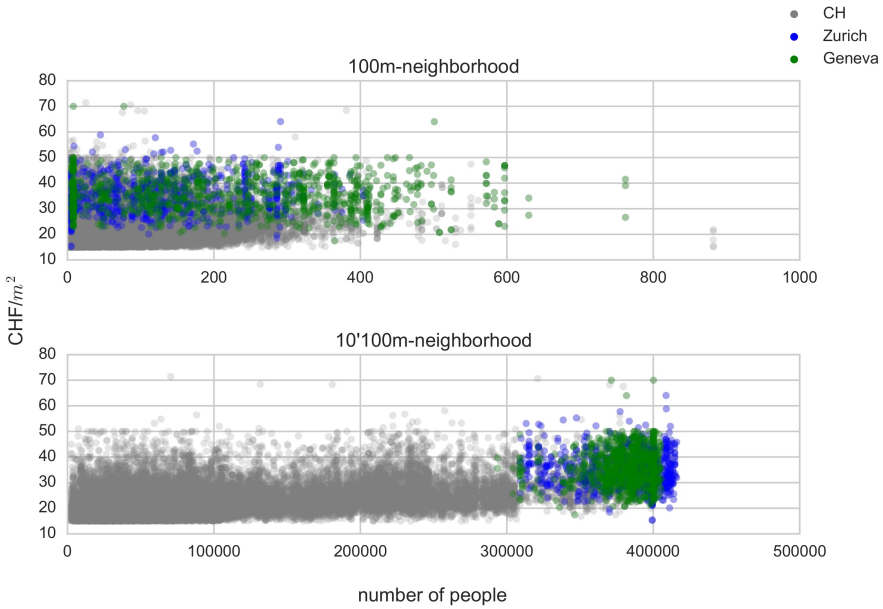


Figure 4.1: Population density in the neighborhood of apartment offers. Only after aggregating the grid points to a larger size, the largest cities of Switzerland show a significant larger population density.

To find the optimal neighborhood size we perform a greedy forward selection as described in Section 3.2 using the apartment features and the Geodata features of different neighborhood sizes. We compare the $R^2$ value with increasing number of features selected for the neighborhood sizes 100 m, 700 m, 2'100 m, 6'100 m, 10'100 m and 12'100 m. Figure 4.2 shows an improvement for growing neighborhood sizes until 10'100 meters. The $R^2$ values are slightly worse for a 12'100 m-neighborhood, no matter how many features are selected. We therefore do not test any larger neighborhood sizes and choose 10'100 m as the optimal size.

## 4.2 Feature selection

Feature selection is performed on the training set using all features and 10'100 m as neighborhood size. The algorithm stops after selecting a hundred features since the $R^2$ value no longer grows significantly and the model is getting too big. In Figure 4.3 we observe the $R^2$ curve flattening after around 20 selected features. The same applies for the prediction error on the train and test set (see Figure 4.4). Even the first few features selected by the algorithm are highly correlated. Some features have negative regression coefficients even though their impact is positive. Therefore it makes no sense to interpret some of the selected

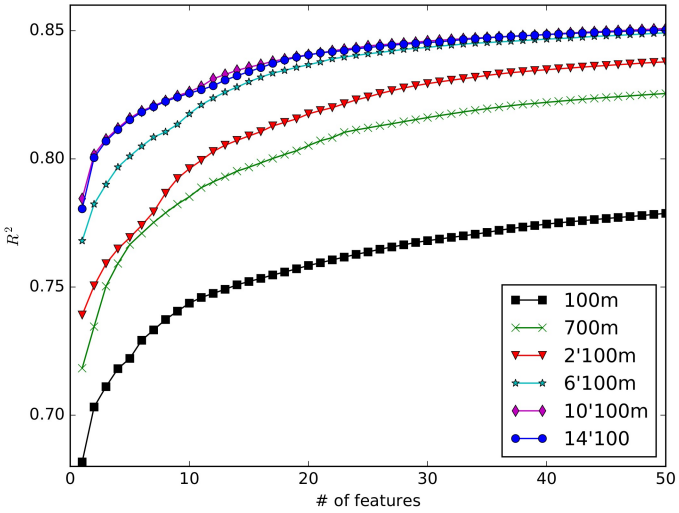


Figure 4.2: $R^2$ value during greedy forward selection for different neighborhood sizes

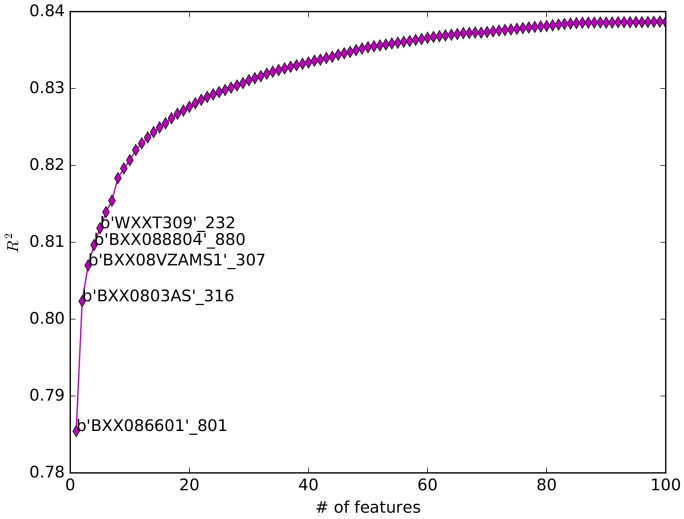


Figure 4.3: Greedy forward selection with top 5 features annotated on 10'100 m-neighborhood. The Algorithm prematurely stops after selecting 100 features since the $R^2$ does no longer improve significantly and the model is getting too big.

features. We make a finding when we take a closer look at the *Y_KOORD* feature. Since most of the apartments in the south are located in Geneva and are

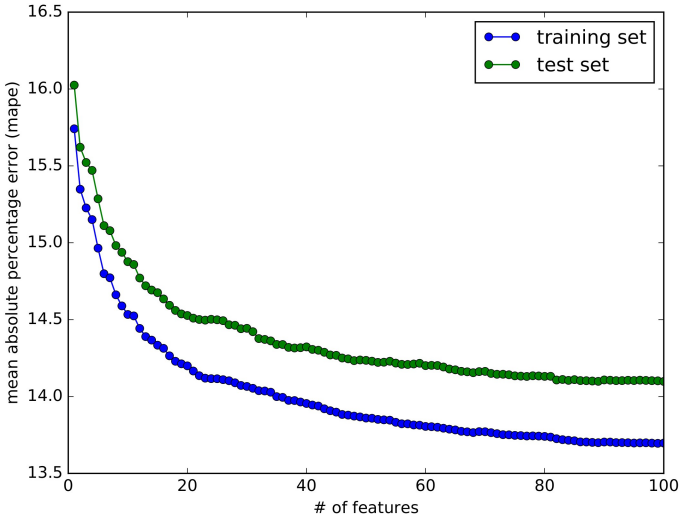


Figure 4.4: mape on the train and test set with the number of features selected

expensive, this feature explains the rising prices in the south.

Since a lot of the features selected by the greedy forward selection are correlated we decide to use a heuristic approach. We are interested in uncorrelated features with high variance in order to ensure robust identification of the primary structures in our data. If the feature is sparse it often explains a specific influence for some apartment samples and the effect can not be generalized. Therefore we perform the feature selection algorithm with manual intervention. If the regression coefficient of a selected feature contradicts the influence on its own, we skip the feature. We stop the selection when the $R^2$ value no longer grows significantly. Compared to the unrestricted forward selection, the $R^2$ value is a bit more than 1% lower after selecting 9 features (compare Figure 4.5 and Figure 4.3). Table 4.2 describes all the features selected and the regression coefficients used by the MLLR model. All features are Min-Max scaled and the regression coefficients and intercept rounded by 2 decimal places.

### 4.2.1 Feature Influence

This section covers the features described in the Table 4.2 in more detail.

#### Apartment Features

The most influential apartment feature is the number of rooms. Figure 4.6 shows that apartments with only a few rooms are substantial more expensive. This is reasonable since those apartments are often rented by singles with more money

Table 4.1: Top 20 features selected by the greedy forward selection

| rank | name | description |
|---|---|---|
| 1 | **BXX086601** | workplaces in finance and insurance services |
| 2 | **BXX0803AS** | workplaces in fishing and aqua culture |
| 3 | **BXX08VZAMS1** | FTEs in first sector (male) |
| 4 | **BXX088804** | FTEs in social services |
| 5 | **WXXT309** | living space of 3 room apartments built in 2000-2005 |
| 6 | **Y_KOORD** | Y coordinate of Geodata grid point |
| 7 | **BXX087001** | FTEs in management and corporate leadership |
| 8 | **BXX0814AS** | workplaces in clothing manufacturing |
| 9 | **BXX086002** | FTEs in radio broadcasting |
| 10 | **cnt_banks_1km** | number of banks within 1km |
| 11 | **BXX0809EMP** | mining services |
| 12 | **minergie** | apartment fulfills Minergie standard |
| 13 | **GXXW02** | number of buildings with 2 apartments |
| 14 | **BXX0811VZA** | FTEs in beverage manufacturing |
| 15 | **BXX0811EMP** | worklaces in beverage manufacturing |
| 16 | **cnt_pubtrans_2km** | number of public transport stations within 2km |
| 17 | **BXX0864AS** | workplaces in finance services |
| 18 | **WXXT401** | living space of 4 rooms apartments built before 1919 |
| 19 | **BXX081804** | FTEs in manufacturing printing material |
| 20 | **BXX080301** | FTEs in fishing and aqua culture |

Table 4.2: MLLR model with heuristic based feature selection

| feature | description | coeff* |
|---|---|---|
| **rooms1** | indicating the apartment has 1 or 1.5 room(s) | 6.94 |
| **rooms2** | indicating the apartment has 2 or 2.5 rooms | 2.65 |
| **BXX086601** | workplaces in finance and insurance services | 8.88 |
| **GXXEWW08** | buildings with solar hot water panels | 4.47 |
| **BXX08S1** | workplaces in the primary sector | -2.55 |
| **BXX086304** | FTEs in information services | 1.73 |
| **cnt_banks_1km** | number of banks within 1 km | 3.61 |
| **BXX087001** | FTEs in management and corporate leadership | 2.74 |
| **cnt_pubtrans_200m** | number of public transport stations within 200 m | 1.97 |
| **intercept** | | 19.55 |

than families. Large luxury apartments increase the average price of apartments with more than 6 rooms. For simplicity we added rooms1 and rooms2 as binary features to the model. They indicating the apartment has 1 or 1.5 rooms and 2 or 2.5 rooms respectively.

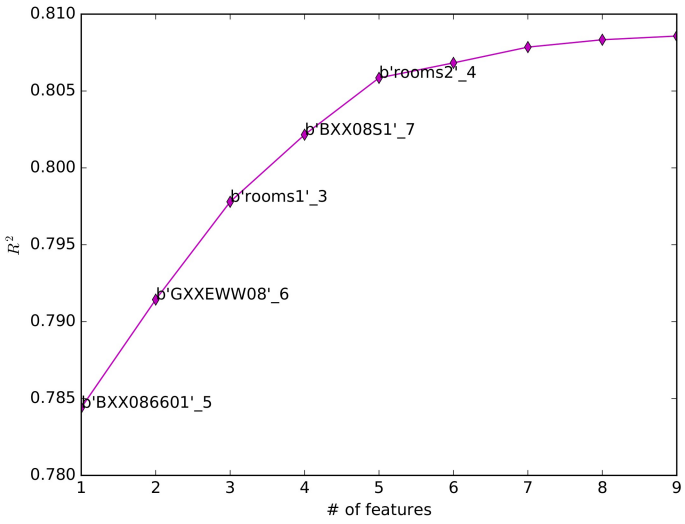


Figure 4.5: Greedy forward selection with manual intervention. We find 9 features with low correlation explaining more than 80% of the variance in rental fees.

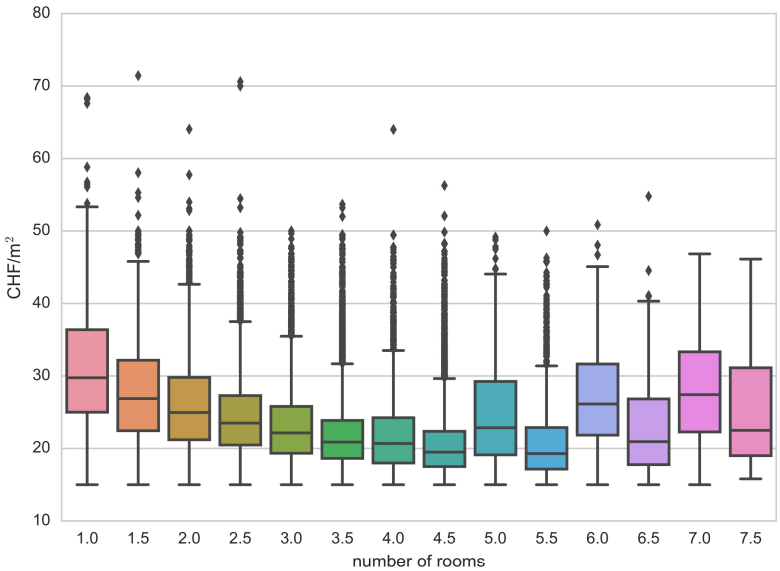


Figure 4.6: Relation between price by square meter and number of rooms of apartment offers with more than 100 samples.

### Population

Figure 4.7 shows a clear relation between the population density and the price per square meter. The population density explains around 75% percent of the variance in rental fees. Since we found features explaining a higher variance and many other feature correlate with the population features, we did not include any those in our model.

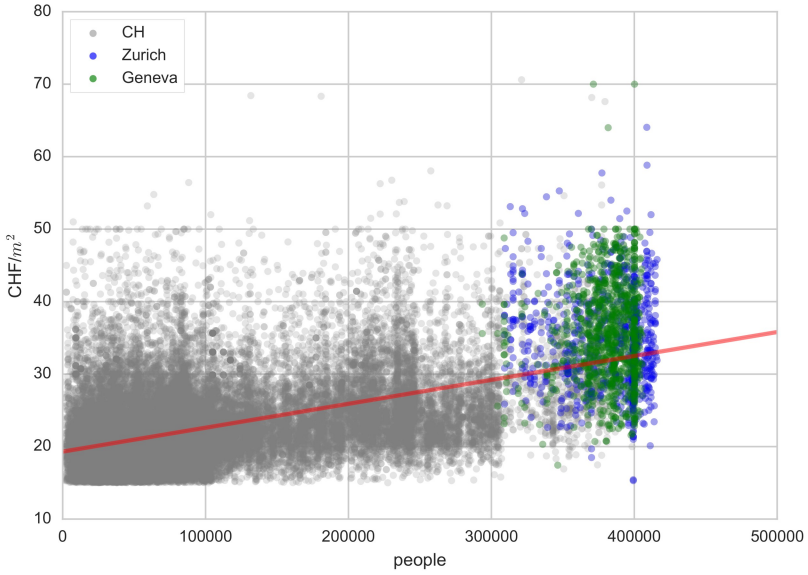


Figure 4.7: Relation between price by square meter and number of people living in a 10'100m.

### Jobs

Figure 4.8 shows a very clear relation between the price per square meter and the number of workplaces in small (less than 10 full time equivalents) finance and insurances service companies. The largest number of workplaces are found in the biggest cities of Switzerland. Figure 4.9 shows the negative impact of workplaces in the primary sector on the price per square meter of an apartment. The example of Zurich and Geneva show that those are mainly located outside of city areas. Figure 4.10 shows the number of FTEs in corporate management and leadership in small companies (less than 10 workplaces) increases the rental fees significantly. It also shows Zurich has the largest amount of FTEs in this area.

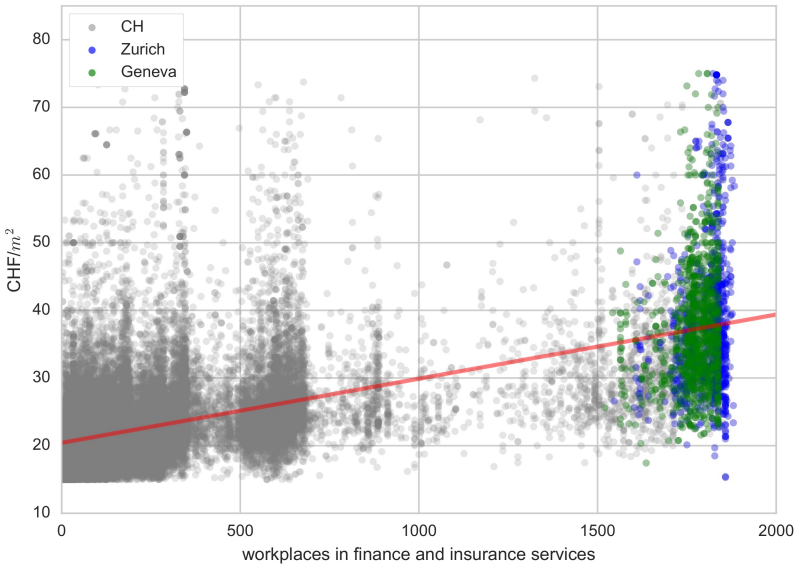


Figure 4.8: Relation between price per square meter and number of workplaces in finance and insurance services in a 10'100m-neighborhood.

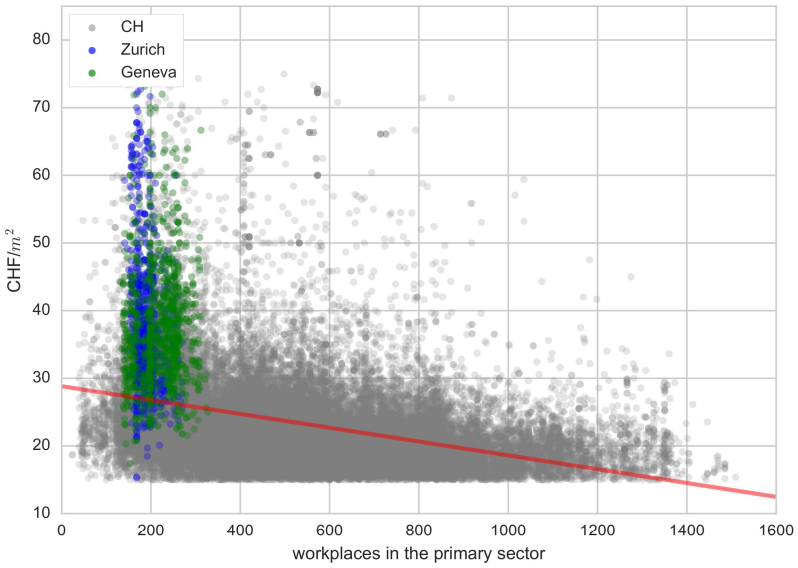


Figure 4.9: Relation between price per square meter and number of workplaces in the primary sector in a 10'100m-neighborhood.

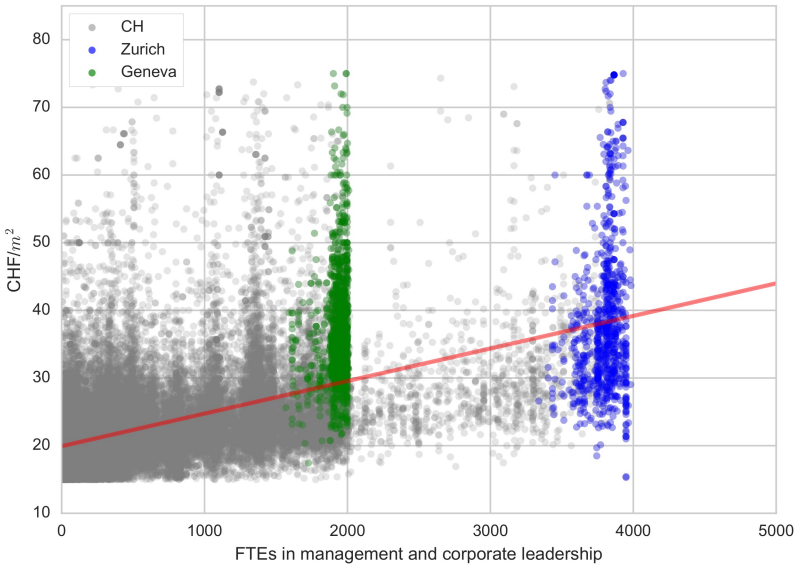


Figure 4.10: Relation between price per square meter and number of FTEs in management and corporate leadership in a 10'100m-neighborhood.

### Buildings

Even though the greedy forward selection algorithm selected some of the building features, the choice can only partly explained. A high number of buildings or a high living space in the neighborhood indicates a densely populated area and therefore higher apartment rents. But the features selected are very specific and dependent on the built year or number of apartments of the building. For example it is hard to justify the feature *WXXT401*, the living space of all 4 rooms apartments built before 1919 in the neighborhood. Those are average sized and there is no evidence to explain popularity for apartments built before 1919. We includes the number of buildings with solar hot water panels in our model (see Figure 4.11). Even though this feature shows some positive correlation on the price, it mainly explains the high prices in Geneva.

### Point of Interests

The most influential point of interests are banks and public transport stations which both make the apartments close by pricier. This is very intuitive since banks are often located in bigger cities and many public transport stations implies good public transport connections. Figure 4.12 shows that the highest density of banks can be found in Zurich and Geneva. Figure 4.13 shows that the positive

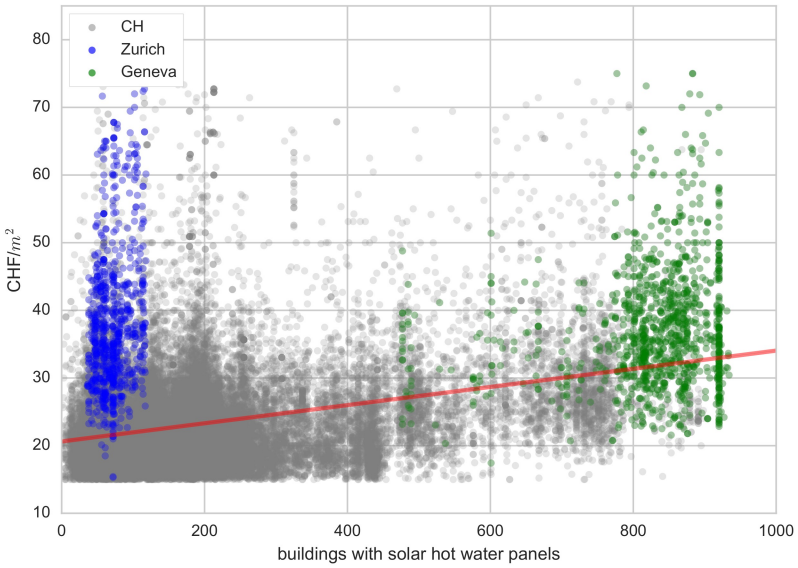


Figure 4.11: Relation between price by square meter and number of buildings with solar hot water panels in a 10'100m-neighborhood.

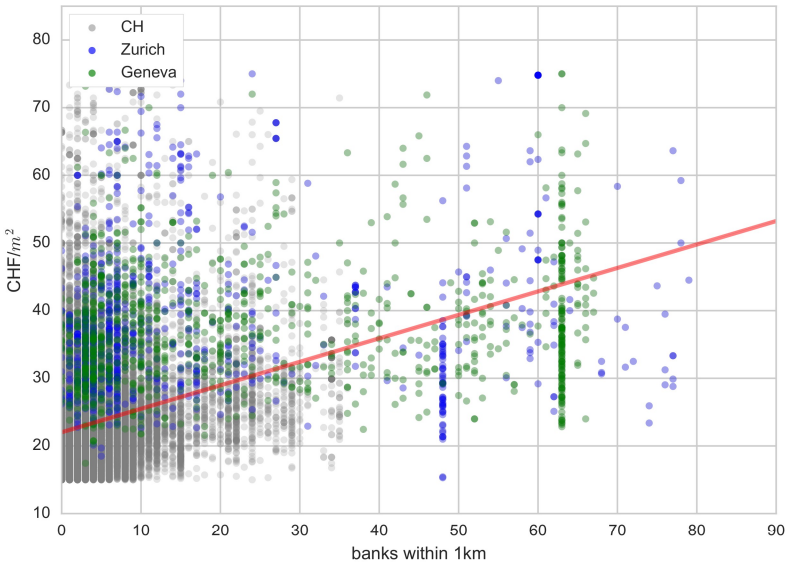


Figure 4.12: Relation between price by square meter and number of banks within 1 km of the apartment.

impact on rental fees of the number of public transport stations within 200 m. Since every platform of a trains station is considered as a station the numbers go up to 37.

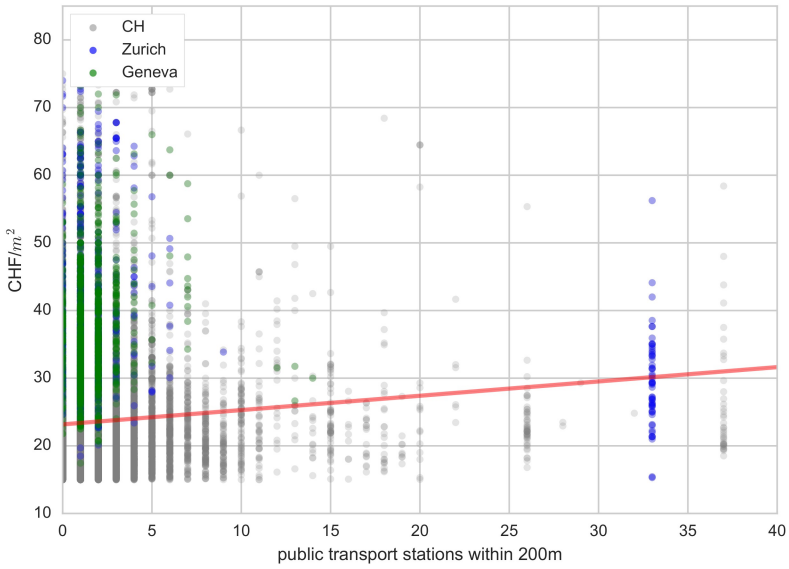


Figure 4.13: Relation between price by square meter and number of public transport stations within 200 m of the apartment. Every platform is counted as a separate station.

### Road Traffic Noise

Unfortunately, the traffic noise data available did not help explaining price differences and was not included in our model. Figure 4.14 shows a small positive impact of traffic noise on rental fees. Even though the correlation is quite low it is counterintuitive. A plausible explanation could be that the traffic noise was computed based on road data and not measured in the apartment. Since the roads are more busy and the rental fees are higher in cities this correlation seems plausible.

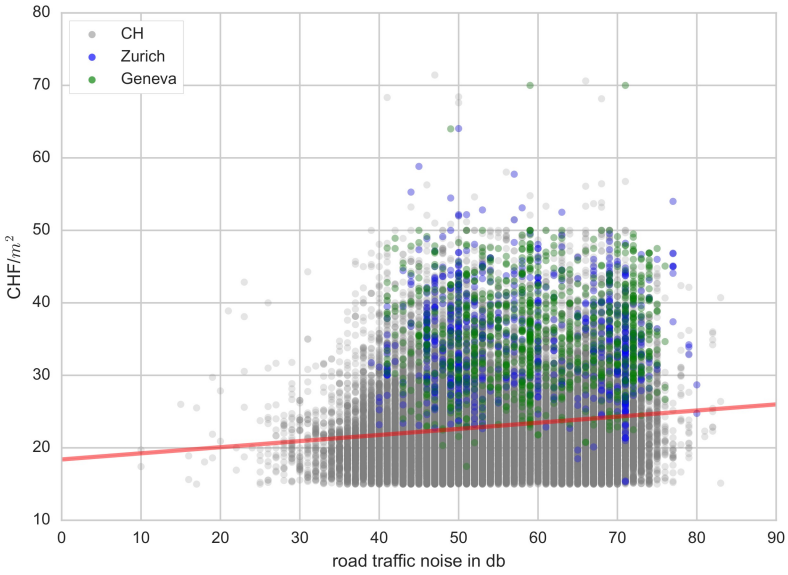


Figure 4.14: Relation between price by square meter and road traffic noise.

## 4.3 Artificial Neural Network based Model

The ANN model is composed of 5 hidden layers containing 90,78,60,42 and 24 neurons with ReLU (rectified linear unit) activation. We train our network using the features described in Table 4.2 on training set of around 30'000 apartment samples and validate the performance on a test set of around 10'000 samples. Neural Network are very powerful and can learn a very complex model to fit the data perfectly. [9] But when evaluated on unseen data, it performs very poorly. We try to prevent this by looking at the training and test set error during the training. If those errors start to depart consistently, it indicates overfitting has occurred and we should stop training. Figure 4.15 shows the error on the training and test set after each epoch, which is the number of times all of the training vectors are used to update the weights of the Neural Network. Since error curves start diverging at around 125 epochs and therefore we decide to stop training at this point.

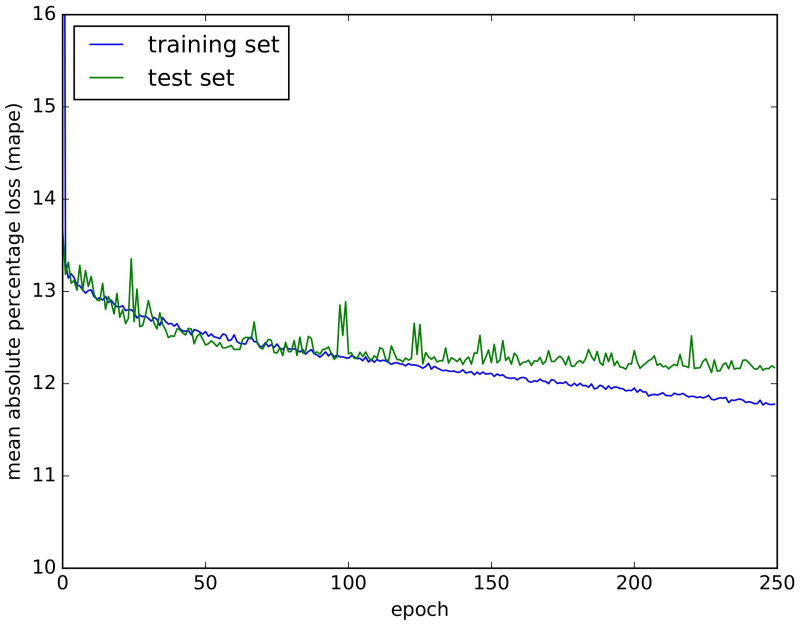


Figure 4.15: Training and test set loss during the training of ANN model. Overfitting occurs after around 125 epochs.

## 4.4 Rental Fee Prediction

For all our models we use the features described in Table 4.2. We train our models on the training set of 75% (around 30'000 apartment samples) and test the models on 25% (around 10'000 samples) of our data set.
Figure 4.16 shows the cumulative distribution function of the percentage deviation from the real rental fee for all models. The ANN curve shows the steepest ascent and therefore the smallest error. It also shows that the ANN model estimates prices too high for more than half the samples in the test set. The Baseline and the MLLR model tend to predict the prices too low. For the ANN model around 8000 samples (80% of the test set) are within -20% and 20% error deviation. The variance of estimation error is therefore less than 20%. The MLLR model performs slightly worse. The error histograms are shown in Figure 4.17. We observe that the MLLR model severely underestimates rentals fees for a few samples. Figure 4.18 shows the cumulative absolute errors of all models. The mean absolute percentage loss of the Baseline is 20%, 14% for the MLLR and 12% for the ANN model.

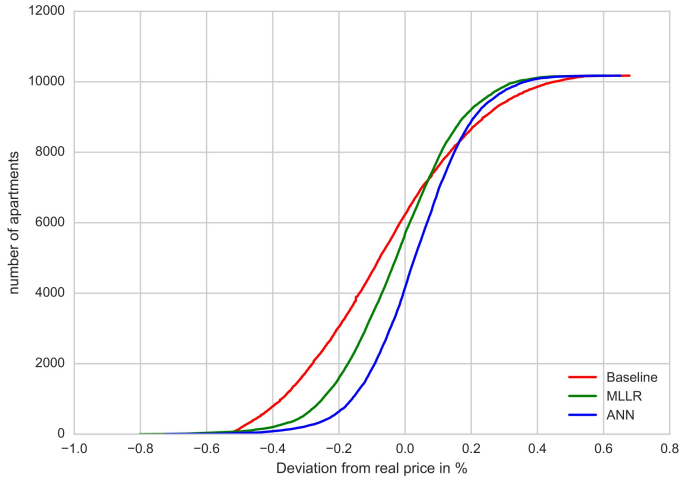


Figure 4.16: Cumulative error distribution function for all models on the test set

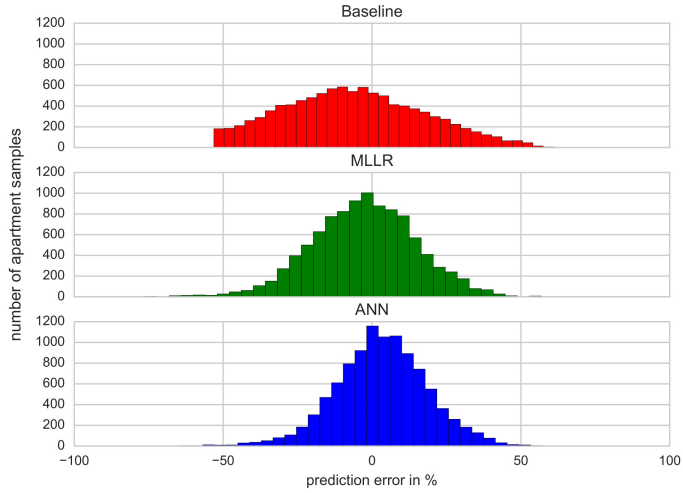


Figure 4.17: Error histogram of all models on the test set.

### 4.4.1 Webtool for price estimation

We create a simple web application to test the prediction performance of our models. The tool allows you to fetch apartment data of a Immoscout24 URL or enter the data manually. Along with the price estimation of all models the tool shows the feature influence graphically (see Figure 4.19). The length of the bars are proportional to the Min-max scaled feature weights.

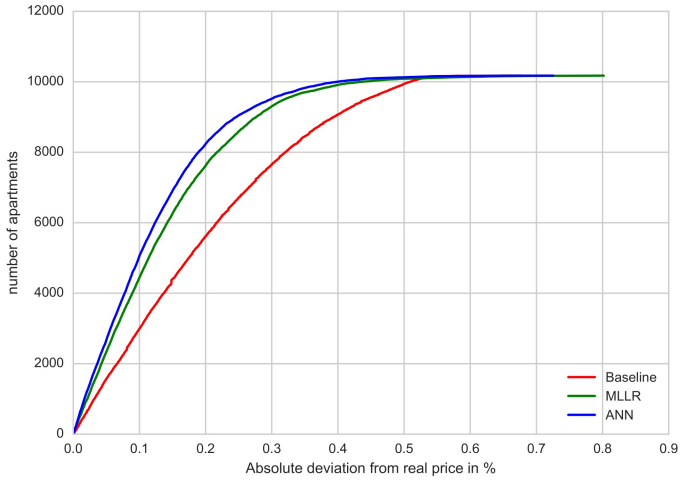


Figure 4.18: Cumulative absolute error distribution function for all models on the test set

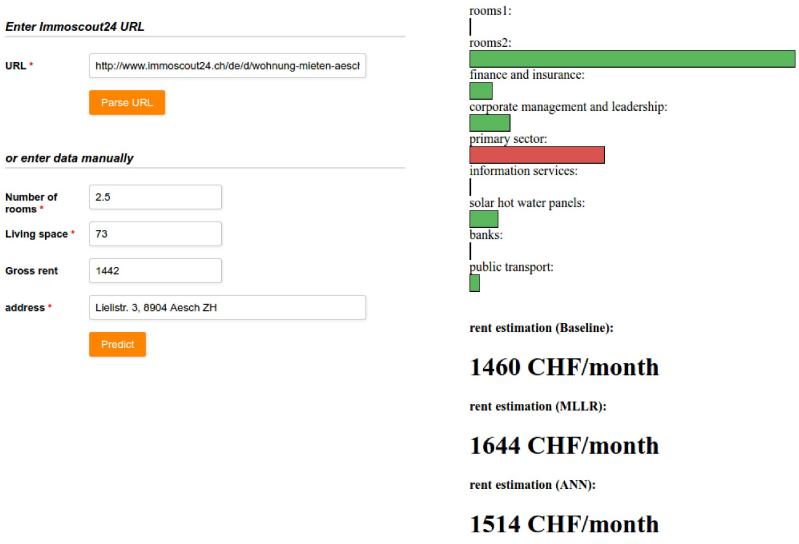


Figure 4.19: Webtool to test the prediction performance of our models.

Chapter 5

# Conclusion

Demographic structure, jobs and the building data only show influence on rental fees if the data is not too fine-grained. We found a high correlation between finance, insurance, management jobs, high density of banks and public transport and the real estate prices. Since many features are connected with the people density, demographic factors were not considered in our models. The same applies for the traffic noise feature as no association could be found. Our linear model was able to explain more than 80% of variance in rental fees and provide reasonable prediction performance. A simple neural network model performed even better.
Access to a complete and accurate data set of apartments offers could be helpful. Information about the noise level measured inside, a standardized and detailed rating for the apartment condition, sunlight and view could improve estimation. Further financial information such as income or buying power of the apartment neighborhood might also be useful. Future work could involve finding non-linear correlation by interpreting the weights of an artificial neural network regression model. For good price estimation it might also help to first cluster the data and then train a model for each cluster.

# Bibliography


[1] Strietzel, R.: What is it worth? (2014)

[2] Trauber, M.P.: Smart real estate value estimation. (2015)

[3] Hu, R., Sjögren, E.: Analysis and prediction of apartment prices in inner city stockholm. (2014)

[4] Krouthen, J.: Apartment values in uppsala: Significant factors that differentiate the selling prices. (2011)

[5] MELICHAR, J., VOJÁČEK, O., RIEGER, P., JEDLIČKA, K.: Application of hedonic price model in the prague property market

[6] Oduwole, H., Eze, H.: A hedonic pricing model on factors that influence residential apartment rent in abuja satellite towns. Mathematical Theory and Modeling **3**(12) (2013) 65–73

[7] Hülagü, T., Kızılkaya, E., Ozbekler, A.G., Tunar, P.: A hedonic house price index for turkey. (2016)

[8] Chiarazzo, V., Caggiani, L., Marinelli, M., Ottomanelli, M.: A neural network based model for real estate price estimation considering environmental quality of property location. Transportation Research Procedia **3** (2014) 810–817

[9] Peterson, S., Flanagan, A.: Neural network hedonic pricing models in mass real estate appraisal. Journal of Real Estate Research (2009)