Apartment Rental Prediction System

Dr. Ivan S. Zapreev

2020-01-06

Introduction

Dataset overview

As stated on the webpage of the 'Apartment rental offers in Germany' dataset, it contains `198,379` rental offers scraped from Germany biggest real estate online platform called ImmobilienScout24.

The data set consists of a single CSV file: immo_data.csv which contains offers for rental properties only. The data features important rental properties, such as living area size, the rent (both base rent as well as total rent), the location, type of energy, and etc. The `date` column present in the data set defines the time of scraping, done on: *2018-09-22*, *2019-05-10* and *2019-10-08*.

The complete list of data set columns is as follows:

```
##  [1] "regio1"                   "serviceCharge"
##  [3] "heatingType"              "telekomTvOffer"
##  [5] "telekomHybridUploadSpeed" "newlyConst"
##  [7] "balcony"                  "electricityBasePrice"
##  [9] "picturecount"             "pricetrend"
## [11] "telekomUploadSpeed"       "totalRent"
## [13] "yearConstructed"          "electricityKwhPrice"
## [15] "scoutId"                  "noParkSpaces"
## [17] "firingTypes"              "hasKitchen"
## [19] "geo_bln"                  "cellar"
## [21] "yearConstructedRange"     "baseRent"
## [23] "houseNumber"              "livingSpace"
## [25] "geo_krs"                  "condition"
## [27] "interiorQual"             "petsAllowed"
## [29] "streetPlain"              "lift"
## [31] "baseRentRange"            "typeOfFlat"
## [33] "geo_plz"                  "noRooms"
## [35] "thermalChar"              "floor"
## [37] "numberOfFloors"           "noRoomsRange"
## [39] "garden"                   "livingSpaceRange"
## [41] "regio2"                   "regio3"
## [43] "description"              "facilities"
## [45] "heatingCosts"             "energyEfficiencyClass"
## [47] "lastRefurbish"            "date"
```

but not all of them will be used in our study. We shall consider the following sub-selection thereof:

1. `hasKitchen` – has a kitchen
2. `balcony` – does the object have a balcony

3. `cellar` – has a cellar
4. `lift` – is elevator available
5. `floor` – which floor is the flat on
6. `garden` – has a garden
7. `noParkSpaces` – number of parking spaces
8. `livingSpace` – living space in sqm
9. `condition` – condition of the flat
10. `interiorQual` – interior quality
11. `regio1` – Bundesland
12. `regio2` - District or Kreis, same as geo krs
13. `regio3` – City/town
14. `noRooms` – number of rooms
15. `numberOfFloors` – number of floors in the building
16. `typeOfFlat` – type of flat
17. `yearConstructed` – construction year
18. `newlyConst` – is the building newly constructed
19. `heatingType` – Type of heating
20. `energyEfficiencyClass` – energy efficiency class
21. `heatingCosts` – monthly heating costs in €
22. `serviceCharge` – auxiliary costs such as electricity or Internet in €
23. `electricityBasePrice` – monthly base price for electricity in €
24. `baseRent` – base rent without electricity and heating
25. `totalRent` – total rent (usually a sum of base rent, service charge and heating cost)
26. `date` – time of scraping

Which reduces the number of dataset columns from 48 to 26 and is motivated by the personal preferences of the author of this report and thus has no scientific motivation. Moreover, this column selection shall be seen as a part of problem statement. In other words, the task is to build an accurate rental price prediction model based on the predictors from this set of columns. The further data pre-processing steps are described and explained in the *"Data wrangling"* section of this document.

Project goal

Execution plan

Data wrangling

```
nrow(arog_data$selected_data)
```

```
## [1] 198332
```

```
## # A tibble: 26 x 3
##    `Column name`          `N/A count` `N/A percent`
##    <chr>                        <int>         <dbl>
##  1 electricityBasePrice       151158         76.2
##  2 energyEfficiencyClass      143315         72.3
##  3 heatingCosts               135154         68.2
##  4 noParkSpaces               130405         65.8
##  5 interiorQual                83001         41.8
##  6 numberOfFloors              71792         36.2
##  7 condition                   50317         25.4
##  8 yearConstructed             42293         21.3
##  9 floor                       37612         19.0
## 10 heatingType                 32605         16.4
## 11 totalRent                   29762         15.0
## 12 typeOfFlat                  27571         13.9
## 13 serviceCharge                5110          2.58
## 14 hasKitchen                      1          0
## 15 lift                            1          0
## 16 garden                          1          0
## 17 cellar                          1          0
## 18 livingSpace                     1          0
## 19 noRooms                         1          0
## 20 baseRent                        1          0
## 21 balcony                         0          0
## 22 newlyConst                      0          0
## 23 regio1                          0          0
## 24 regio2                          0          0
## 25 regio3                          0          0
## 26 date                            0          0
```

```
nrow(arog_data$cleaned_data)
```

```
## [1] 168543
```

```
## # A tibble: 24 x 3
##    `Column name`          `N/A count` `N/A percent`
##    <chr>                        <int>         <dbl>
##  1 hasKitchen                       0             0
##  2 heatingType                      0             0
##  3 balcony                          0             0
##  4 lift                             0             0
##  5 garden                           0             0
##  6 cellar                           0             0
##  7 noParkSpaces                     0             0
##  8 livingSpace                      0             0
##  9 typeOfFlat                       0             0
## 10 noRooms                          0             0
## 11 floor                            0             0
## 12 numberOfFloors                   0             0
## 13 condition                        0             0
## 14 newlyConst                       0             0
## 15 interiorQual                     0             0
## 16 energyEfficiencyClass            0             0
## 17 regio1                           0             0
## 18 regio2                           0             0
```

```
## 19 regio3                  0            0
## 20 baseRent                0            0
## 21 heatingCosts            0            0
## 22 serviceCharge           0            0
## 23 totalRent               0            0
## 24 date                    0            0
```

$168562 \approx 168543$

Data analysis

Modeling approach

Results

Conclusions