TU DORTMUND
Fakultät Statistik
Prof. Dr. Jörg Rahnenführer
JProf. Dr. Antonia Arsova
Dr. Birte Hellwig
M.Sc. Julia Duda
M.Sc. Sven Pappert

Winter semester 2020/2021

Project III
Regression modelling

Immobilienscout24 (immoscout24.de) is one of the three biggest real estate web-portals in Germany. On the website you may find listings of rental properties and homes for sale. The given data set (`ImmoDataNRW.csv`) contains 12118 rental offers for properties located in the province of North Rhine-Westphalia as of February 20, 2020. The full data are available on kaggle.com for educational/research purposes only.

The data set contains the following 16 variables:

- `ID` - Listing identification number
- `newlyConst` - whether the property is newly constructed (in 2019 or in 2020)
- `balcony` - whether the property has a balcony
- `totalRent` - total rent (usually a sum of base rent, service charges and heating costs)
- `yearConstructed` - construction year
- `noParkSpaces` - number of parking spaces provided with the property
- `hasKitchen` - whether the property features a fitted kitchen or not
- `livingSpace` - property size in square meters
- `lift` - whether the property has a lift
- `typeOfFlat` - type of the flat
- `floor` - the floor the property is in
- `garden` - whether the property has a garden
- `regio2` - city/municipality where the property is located
- `condition` - condition of the property
- `lastRefurbished` - year of last renovation
- `EnergyEfficiencyClass` - energy efficiency class of the building

Task 1: Data preparation

1. Compute the rental price per square meter (`sqmPrice`) of each property. This will be the dependent variable to be modelled in Task 2: Linear regression.
2. Group the values of the variable `noParkSpaces` into categories "0 or no information" (no information or no parking spaces provided) and "1+" (one or more parking spaces provided) and use the categorized variable for further analysis.
3. Group the values of the variable `typeOfFlat` into categories "apartment", "luxurious_artistic_other" (comprising the values "loft", "maisonette", "penthouse", "terraced_flat" and "other"), "r_ground_floor" (comprising the values "ground_floor" and "raised_ground_floor") and "roof_halfBasement" (comprising the values "roof_storey" and "half_basement"). Treat the missing values in a meaningful way to avoid data loss. Justify this grouping in view of Tasks 2 and 3. Use the newly categorized variable for further analyses.
4. Derive a suitable transformation of the variable `yearConstructed` and use that one for further analyses.
5. Derive a new variable `CityType` with three levels describing whether the property is located in one of the 5 biggest cities in North Rhine-Westphalia, in the 6-10 biggest cities or whether it is located elsewhere.
   Note: You may want to consult the Wikipedia entry for North Rhine-Westphalia.

Task 2: Linear Regression

1. Estimate a linear model for the rental price per square meter (`sqmPrice`) using the remaining variables as predictors.
2. Estimate a "best" possible model for the rental price per square meter (`sqmPrice`) employing backward stepwise variable selection and the Akaike Information Criterion (AIC) as a model selection criterion.
   Interpret the coefficients of the resulting model and their statistical significance. Evaluate the goodness of fit as well.

Task 3: Logistic Regression

1. Model whether the property is newly constructed or not (dependent variable: `newlyConst`) by a logistic regression.
   Note: You may not use information on how old the property is as a predictor, i.e. the variables `yearConstructed`, `lastRefurbish`, or their transformations.
2. Perform stepwise (backward) variable selection using the AIC as a criterion and interpret the coefficients and their significance.
   Evaluate the discriminatory power of the model by interpreting the confusion matrix.

Submission

Submission of the report and the corresponding (executable and commented) program code until Friday, Feb 5th, 2021, 08:30 am, in Moodle.