German Apartment Rentals

Magnus Paal and Sander Nemvalts

The data

- A recently published Kaggle dataset
- Rental apartments in Germany from ImmobilienScout24
- 268 000 rows/listings
- 49 columns of varying amounts of data.
- Latitude and longitude added with postprocessing

name
regio1
serviceCharge
heatingType
telekomTvOffer
telekomHybridUploadSpeed
newlyConst
balcony
picturecount
pricetrend
telekomUploadSpeed
totalRent
yearConstructed
scoutId
noParkSpaces
firingTypes
hasKitchen
geo_bln
cellar
yearConstructedRange
baseRent
houseNumber
livingSpace
geo_krs
condition
interiorQual
petsAllowed
street
streetPlain
lift
baseRentRange
typeOfFlat
geo_plz
noRooms
thermalChar
floor
numberOfFloors
garden
livingSpaceRange
regio2
regio3
description
facilities
heatingCosts
energyEfficiencyClass
lastRefurbished
electricityBasePrice
electricityKwhPrice
date

What problem needed to be solved?

- Can we predict the rental price?
  - Help both parties know if the price is fair for a given property.

- Can we predict the heating costs?
  - Help the renter understand the not-so-obvious costs of renting.
  - More advanced: Heating type, heating costs -> CO_2 intensity of apartment heating.

Predicting rent

- Regression (predict total rent)
  - Explored various models
    - Random forests, neural networks, linear regression, more general gradient boosting, XGBoost
  - Best results from XGBoost regressor
    - Mean rent of dataset (y) - ~802
    - Standard deviation of y - ~437
    - Mean absolute error on test set - ~87
    - R^2 of ~0.9
    - Exceeded expectations
- Classification (predict price class/quantile)
  - 4 equally sized classes based on total rent
  - Best results from random forest classifier
    - Accuracy of ~77%, random baseline is 25%
    - Results did not match expectations

Predicting heating and CO_2 output from heating

- Also regression (predicting monthly heating costs).
- Was confident in XGBoost, stuck only with it.
- Best results:
  - XGBRegressor
  - Dataset mean heating costs was 69, standard deviation 22
  - Mean absolute error of ~11 on test set, pretty good.
  - R^2 of ~0.5 on test set. Would've liked better results.
- CO_2 prediction
  - Very simple logic, without ML
    - Calculate CO_2 emitted from heating type, €/kWh and $kgCO_2$/kWh.
  - Calculations seem close to reality (1-3 tons of CO_2 per property per year)
  - Also created a model that takes in same data as heating model and predicts CO_2 emitted
    - R^2 of ~0.9

What we learned

- Preprocessing the data turned out to be the most important step
  - We were not getting the results we wanted
    - Discovered that we had made a mistake preprocessing the data
    - Didn't approach cleaning with enough methodology, there were some considerable outliers left when we started training the models.
  - Cleaned data even more
  - Fixed mistakes made in cleaning
  - Results got much better

GitHub Repository

https://github.com/magnuspaal/germany-rental-ml