ML2 Project

Eric Jonas, Caroline Graebel, Naim Iskandar Zahari, Kseniia Holovchenko

2024-11-30

Introduction

Understanding rent prices in Berlin is essential for anyone looking to rent an apartment in the city, given its dynamic housing market and the growing demand for affordable living spaces. To address this, our project focuses on analyzing historical rental data from Berlin in 2020, aiming to uncover insights and develop predictive models that estimate rental prices based on key property attributes and evaluate them on current apartment offers.

The following dataset, available at Kaggle, contains information about rental properties listed on ImmoScout24. The data was scraped from ImmoScout24 between February and October 2020 and includes listings from all German federal states. For our project, we will focus exclusively on Berlin and train a model to predict either the `totalRent` or the `basePrice`.

Data Description

```
## 'data.frame':    10406 obs. of  30 variables:
##  $ serviceCharge     : num  320 79 150 229 147 ...
##  $ heatingType       : chr  "central_heating" "central_heating" "floor_heating" "floor_heating" ...
##  $ newlyConst        : logi  FALSE FALSE TRUE FALSE TRUE FALSE ...
##  $ balcony           : logi  TRUE FALSE TRUE TRUE TRUE TRUE ...
##  $ picturecount      : int  10 17 15 2 9 21 8 15 3 6 ...
##  $ pricetrend        : num  4.99 7.35 6.6 8.63 7.56 4.99 6.06 8.54 6.3 7 ...
##  $ telekomUploadSpeed: num  NA 40 40 NA 40 40 40 40 40 40 ...
##  $ totalRent         : num  1140 955 1300 1429 1559 ...
##  $ yearConstructed   : int  NA 1918 2019 2017 2019 2014 1980 1870 1984 1988 ...
##  $ noParkSpaces      : int  1 NA 1 NA NA 1 NA NA NA NA ...
##  $ firingTypes       : chr  NA "gas" "district_heating" "district_heating" ...
##  $ hasKitchen        : logi  TRUE FALSE TRUE TRUE TRUE TRUE ...
##  $ cellar            : logi  FALSE FALSE TRUE TRUE TRUE TRUE ...
##  $ baseRent          : num  820 808 1150 1200 1338 ...
##  $ livingSpace       : num  77 62.6 46.4 67 73.5 ...
##  $ condition         : chr  NA "refurbished" "first_time_use" "mint_condition" ...
##  $ interiorQual      : chr  NA NA "luxury" "sophisticated" ...
##  $ petsAllowed       : chr  "negotiable" "negotiable" "no" "negotiable" ...
##  $ streetPlain       : chr  "Metropolitan_Park" "Börnestraße" "Stallschreiberstraße" "Hallesche_Strai
##  $ lift              : logi  TRUE FALSE TRUE TRUE TRUE FALSE ...
##  $ typeOfFlat        : chr  "ground_floor" "ground_floor" "apartment" "apartment" ...
##  $ geo_plz           : chr  "13591" "13086" "10179" "10963" ...
##  $ noRooms           : num  3 2 2 2.5 2 3 3 4 3 1 ...
##  $ thermalChar       : num  NA 100.4 NA NA 66.2 ...
##  $ floor             : int  0 0 3 6 0 1 16 1 4 2 ...
```

```
##  $ numberOfFloors    : int  3 3 5 7 6 2 NA 2 NA 5 ...
##  $ garden            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ regio3            : chr  "Staaken_Spandau" "Weißensee_Weißensee" "Mitte_Mitte" "Kreuzberg_Kreuzbe
##  $ heatingCosts      : num  NA 68 NA NA 73.5 ...
##  $ lastRefurbish     : int  NA NA NA NA 2019 NA NA 2003 NA 2019 ...
```

Objectives

Modeling Objectives

- **Predict Missing Values**:
  Our goal is to impute the missing values in the `totalRent` column and use this data to train a predictive model. We will then compare the performance of this model with one trained on the original, non-imputed data to assess the impact of imputation on model accuracy.
- **Model Selection and Comparison**:
  We aim to train a series of tree-based models on the dataset, starting from simple decision trees that can provide explainable predictions and gradually progressing to more complex models such as Random Forests. Additionally, we will compare the performance of these tree-based methods with Support Vector Machine (SVM) regression to determine which approach yields the best results.

Data Overview and Cleaning

Models

References

Bar, C. (2020). Apartment rental offers in germany. In *Kaggle*. https://www.kaggle.com/datasets/corrieaar/apartment-rental-offers-in-germany/data