Abstract

German Rental Market and Offers

Principles and Elementary Models

Group 7

Joaquin Arias
Lou De Gaetano Nerino De Visconti
Sicheng Huang
Himanshu Midha
Matilde Monti

Introduction

The German rental market, with its complexity and diversity, offers a valuable opportunity for data-driven analysis and finding insights that can provide key information to policymakers, tenants, and property owners. In this project, we explore a dataset containing rental property listings in Germany and try to predict the Total Rent for each listing, as well as identify trends, relationships, and actionable insights.

The importance of this project is huge for the three stakeholders involved in this market:

- **Tenants** will be able to use the models to predict fair rental prices based on the property features and avoid overpriced listings or be able to negotiate better and more informed based on the market.
- **Owners** can also benefit from the market information and our models by being able to price and market their rentals effectively. They will be able to set competitive yet profitable rents. They will also know which features can maximize their rent-price property and invest properly.
- **Policy Makers** will be able to understand the market and address housing inequality and affordability issues. This will make them make informed decisions and policies.

Methods

The data used in this analysis was retrieved from Kaggle[1] and it contains rental property listings data scraped from Immoscout24, the biggest real estate platform in Germany. The dataset contains offers from the dates 22/09/2018, 10/05/2019, and 08/10/2019 with information spanning from regional characteristics to energetic data.

After having identified many issues in the information contained, the dataset went through a series of steps to clean and prepare it for modeling:

- Remove columns (duplicates and unnecessary variables)
- Remove rows with +15 null values
- Outlier detection and treatment for imputation errors and unrealistic values
- Imputation of missing values (based on logical assumptions and/or statistical methods)

With the dataset ready to be worked on, we applied Machine Learning approaches and chose the most appropriate ones to obtain meaningful outcomes. We first chose Supervised Learning techniques to predict rental prices: linear, ridge and lasso regression; tree based methods such as decision trees and random forest; as well as gradient boosting methods like XGBoost, LightGBM and CarBoost models.

---

[1] Link: https://www.kaggle.com/datasets/corrieaar/apartment-rental-offers-in-germany/data

To improve the model accuracy further, we also added feature engineering to gain additional information from the columns.

To see whether other variables could be inferred from the data, we also used classification algorithms to try and predict different features of each listing like if the property had a balcony and garden.

Afterward, we performed an unsupervised learning analysis to try to understand additional information from the data. PCA and clustering were the two main strategies adopted.

Results

Most models perform fairly when predicting total rent, with the best-performing ones being RandomForest and CatBoost (both with an R2 of 0.77 on the initial dataset). Hypertuning and cross-validation further improved their precision, raising the R2 score to 0.79. Adding features such as regional differences and property conditions raised the accuracy further, increasing the R2 to 0.80.

Key predictors of rental prices included "livingSpace" and "noRooms", as well as regional differences. Larger properties in urban areas and better technological appliances required higher rents, as well as additional features such as garage, balcony and energy efficiency.

Classification algorithms also performed quite well, with high Accuracy and F1-scores for both the presence of a Garden (accuracy and F1-scores of 0.97 with RandomForestClassifier) and the presence of a balcony (0.73 accuracy with RandomForestClassifier).

Discussion and Implications

Having obtained models that can accurately predict total rent given the characteristics of a property, these algorithms will be useful to owners or agencies that need to price their listing. These predictions will provide a comprehensive understanding of the whole market of rental offers in Germany, as well as offering a benchmark to other offers.

Additionally, the classification models prove to be useful when the information is incomplete and details need to be imputed. For example, the models will be able to accurately predict if a property has a garden or a balcony with enough precision to be published.

Our findings offer a full view of the German rental market, highlighting the importance of information such as property features, their condition and regional attributes that shape their value in the market. This lesson proves to be valuable for owners, tenants and people in between them, as they can now understand what features to focus on to make their offers more interesting for clients and clients can now negotiate with more information on how the market behaves.