Predicting Estonian apartment rent prices based on KV.ee listings

Sebastian Mais, Caroline Markov, Rainer Talvik

Project repository: https://github.com/S33bu/DS_RENT

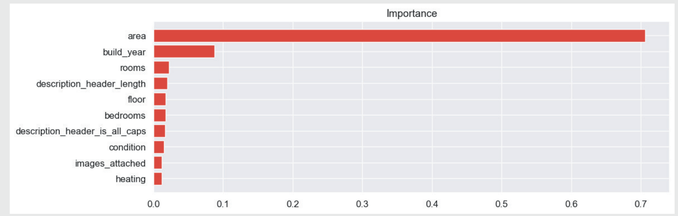

Figure 2. Most important features as chosen by GBM, 2024

Motivation & Goal

The real estate market in Estonia, particularly in cities like Tallinn and Tartu, has seen significant growth and fluctuation in recent years. With the increasing demand for rental properties, both landlords and tenants face challenges in determining fair rental prices. The project aimed to analyse and model long-term apartment rental listings from the Estonian real estate sale and rental site KV.ee [1].

> Our goal was to create a predictive model with an RSME (Root Mean Square Error) of 50€ to be within ±15% of most rental listings.

Data scraping

Our primary data source was the Estonian real estate site KV.ee, from where we scraped the data needed.
We used the Web Scraper [2] browser extension to gather data from KV.ee, Python for data analysis and different libraries such as NumPy, Pandas and Scikit-learn for data processing and modelling.

References

[1] Real estate KV.EE. (2024). Used on 16.11.2024, https://www.kv.ee/
[2] Web Scraper. (2024). Used on 16.11.2024, https://webscraper.io/

Data cleanup & preprocessing

The dataset contains 2676 listings, with approximately 1470 from Tallinn (see Figure 1) and 500 from Tartu, making it suitable for modeling rental prices in those cities. Various preprocessing steps were performed to clean the data. For example, most values in the column floor_out_of_floors were misinterpreted as dates, which were corrected to numerical values. Features like heating and energy_mark were converted to numerical form using one-hot encoding. Key information from text fields like description and summary was extracted, such as heating, house_type, and furnished, which were represented as boolean values.

Figure 1. Map of KV.ee listings in Tallinn, 7.12.2024

Modelling methodology

By the end of data processing, we were left with 31 columns of usable data. We created a correlation heatmap to understand the relationships between features. Then, we used grid search to assist with evaluating models and tuning hyperparameters.

Results & Conclusions

The features most useful for predicting rental price were area, build year, rooms, description header length, the floor of the apartment and the number of bedrooms (see Figure 2).
We managed to get an RSME of about 266€ using a Gradient Boosting model (GBM), which is more than 5x larger than our goal (see Figure 3). However, we consider this a big success. First of all, an RSME of 266€ isn't actually huge, considering that the mean rental price in our dataset is approximately 684€ and the standard deviation is about 503€ – our initial goal was just unrealistic. Secondly, this shows that with surface-level information that can be extracted from rental listing texts couldn't accurately reflect the possible value of an asset such as an apartment. Computers can't yet compete with human subjectivity. While image processing and additional feature engineering could narrow the price predictions by a considerable amount, the human element cannot be put into numbers.
This project underscored the importance of having a thorough understanding of the topic before starting, as well as the value of high-quality data.

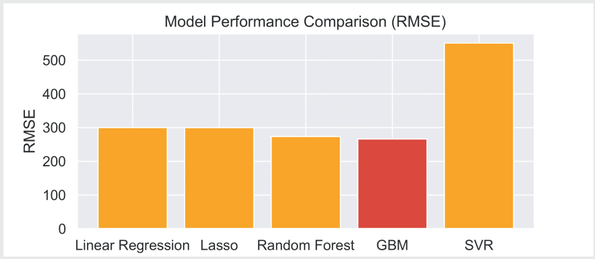

Figure 3. KV.ee price prediction model RMSE comparison, 2024