Estimation of the asking rental price in Budapest
based on advertisement history

Master's in Business Analytics
- Capstone Project Summary

by Skirpichnikova Ksenia
Budapest 2020

Contents

Goal

The goal of the project is to build a model to predict the asking rental price in Budapest based on advertisement history.

We are interested in finding answers to the following questions:

1) How current data can be used to identify the main drivers for rental listing?
2) What listings are over- or undervalued?
3) What machine learning algorithm is the optimal one for estimation of rental listing?

Background

Ingatlan.com is the company with more than 20 years of experience on real estate market. The company is interested in finding out if available data for rental prices can be used to determine any patterns between features of the advertised property and the asking rental price. Results of the project can be used to improve customer experience by changing the structure of the information required for placing advertisement: focus on important features and delete not important features, as well as help content management team screen submitted advertisements to detect incorrect/fake ones.

Analysis plan

Ingatlan.com has an advertising database, flats and rooms for rent in Budapest, data for the period 01.07.2018. - 30.06.2019 was used for the analysis. Sampling rate was not specified.

What questions do I need to answer?

I have questions related to data quality, data engineering, modelling and evaluating the results.

What risks to this analysis exist in the business, data or problem that I need to account for, especially adversarial?

There are two risks related to data quality:

- imputation errors and missing values when a customer did not fill fields for amenities or preferred to give information in the text to the advertisement, but text was not taken for analysis because it had some sensitive information;

-some features will not be included in the analysis, again due to high proportion of missing values;

Among external risks I need to mention the impact of COVID-19 pandemic on every company's data and analytics strategy due to changes in the underlying data. That is relevant for the current analysis as well: if the rental price in Budapest falls in comparison with historical data, the company will have to collect the new data again and retrained the models. I have

conducted a small research and got an information from some landlords that asking rental price for this uncertain period April – August was decreased due to pandemic, but landlords plan to return the rental price to pre-pandemic level since September 2020. This is a bias that was not taken into consideration by the model I built.

The data used for the analysis is for the 1-year period while it is treated as cross-sectional data. The assumption was made that the asking rental price was constant within the considered period.

Sample design

Sample for the analysis contains data for apartments (rooms were excluded as there were only a few observations) in the price range from 35K to 600K HUF.

This sample captures the typical apartments for rent and excludes too cheap or too luxury apartments.

I filtered out observations with more than 4 rooms and kept area size in the range 20-200m2 due to the scarcity of observations outside these ranges.

Once a variable has more than 30% of missing data it is not included in the analysis. Exception was done for comfort level having 31% of missing data.

Dealing with outliers

Outliers were checked for variables utility cost and price per m2. On the scatter plot there are a few observations with too high utility cost or too high price for their area size.

I used Z-score with threshold 3 to filter out outlier observations. Z-score finds the distribution of data where mean is 0 and standard deviation is 1 i.e. normal distribution. For calculating the Z-score data is re-scaled and centered, data points which are way too far from zero are treated as the outliers.

Loss function

I use Mean Absolute Percentage Error, also known as MAPE for summarizing and assessing the quality of a machine learning model and to compare the performance of the models. MAPE is a simple average of absolute percentage errors:

Equation 1 Mean absolute percentage error

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

where At is the actual value and Ft is the forecasted value, n is number of observations.

To compare results of different models I used MAPE on holdout dataset, the actual and the predicted prices transformed back from logs.

Models

I have specified three models for predicting rental price:

(1) Linear Regression (OLS) models (package caret): 12 hand – picked variables

(2) Gradient Boosting model (package caret, package xgboost): 100 variables (including dummies)

(3) Stacked – ensemble model (GLM, GBM, Random Forest, Deeplearning – package H20): 34 variables

Summary

Based on mean absolute error (MAPE) XGBoost model (package caret) was chosen as the best model for prediction of rental prices.

At the same time the models specified in this project have close prediction results as the difference in mean absolute errors is negligible.

While XGBoost model captures 77% of the rental price variance, the mean absolute percentage error the model has is 13.28%. Almost one third (28%) of the observations have an error less than 5%, 62% of observations have an error ±13.28% (in frames of MAPE).

The benchmark OLS model can be used for understanding of key relationship between features and the asking rental price. For example, the rental price for the same apartment but with an air conditioner can be expected to be 8% higher than without the air conditioner.

The most important feature for rental price is the area of an apartment. Importance of all other features is significantly lower. Next important features are number of rooms, if an apartment is renovated, availability of an air conditioner, if an apartment is located in district V and the value of utility cost indicated in an advertisement.

One of the challenges of the dataset is the volume of missing values among indicated features, nevertheless, as checked on paid advertisements, it has no influence on the mean absolute percentage error for XGBoost model. According to the results of robustness tests inclusions of zones instead of districts can improve prediction as mean absolute percentage error decreased from 13.28% to 11.39%. Yet prediction of the rental price per square meter instead of the price for the whole apartment does not improve results.

The chosen XGBoost model can be used on new listings to find out overvalued or undervalued listings and be an additional tool for content management team.