Capstone Project Summary

Transaction Price on Property Market

Ghazal Ayobi

June 2022

Contents

1 Introduction

The client of my capstone project is real-estate company whose mission is to provide its customers with the best solutions for their investment plans. Thus, for this project, I aim to predict property transaction prices for Budapest and other major cities. Many factors may influence the real estate prices when selling/buying a property. Such as; macroeconomic factors, environmental characteristics, and property information. To summarize, the goal of this project is to build a model which predicts the transaction price of a given property based on detailed property characteristics using regressions and algorithms.

Based on the Hungary's Central Bank report stated that the number of property transactions declined by 58-70% for both local and foreign customers in April 2020. In 2021, the Global Property Guide publication published that Hungary's house price increases are accelerating. Thus, we can say several factors may influence transaction price in the property market and the number of property transactions in the real estate market. As mentioned housing prices tend to fluctuate over time thus, predicting the transaction prices help the company to set price for their new apartments which are not yet in the market and estimate the selling price of a property and assist customers to make informed decisions for their investment.

To answer the research question, this paper explores various regressions and algorithms to cross compare multiple prediction models. Root mean squared error is used evaluate model performance. Root mean squared error highlights the average distance between the predicted values and the actual values in the dataset. RMSE can be used in the original data and live data sets to evaluate predicted values.

2 Data Engineering

The raw data for this project comes from the client property advertisements website. During the data engineering procedure the main aim was to automate data cleaning at the most possible rate. Thus, I initially performed data cleaning by setting the encoding environment to read the data in its natural form, remove decimals encoding errors, created factor variables, added new meaningful predictors, created pooled variables to group sub values, replaced words with the same meaning for the property condition.

The main goal of the project is to predict prices of flats in capital region and other major cities. The key variable is transaction price, which contained extreme value and missing observations, thus I dropped the missing observation and only focused on the properties which are common and disregard very cheap and expensive properties. As a result of data cleaning and transformation, the file data work file is divided into two datasets; which are capital region districts and major cities.

3 Data Analysiss

The best model gives the best prediction in the live data. Before turning to the modeling part of the project, it is worth mentioning that in order to avoid over fitting, the original data is split into two random parts by 20% to 80% ratio. Holdout set contains the 20% and the rest is work data set.

For the purpose of data analysis and feature engineering, first I created several groups of the predictors based on the variable importance, which as following: Basic variables, Basic addition, Polynomial level, and interactions. Afterwards, I applied various linear regressions on the grouped variables, creating five OLS models with different difficulty levels. As a result the best performing regression is chosen for the further analysis in the each group of data sets. Based on the 5-fold cross-validation RMSE best models were chosen for the capital region and cities. These models has the lowest RMSE in the test sets.

After selecting the best OLS model, I evaluated the data using machine learning models. A machine learning algorithm is a methodology through which an AI system undertakes a task and predict the value based on the given dataset. It is important to run and evaluate different models and methodologies for a given data set, as this assist to understand the dataset in broader perspective. In order to predict flats transaction prices, I used CART, Random Forest and GBM algorithms. Based on the work and holdout data performance and RMSE, ordinary least square method provided the best performance across different algorithms.

The best selected model, OLS, where we can see the pattern of association that drive prediction. It is important to perform post prediction diagnostics tools which can be used to uncover information about the patterns of association which drives prediction. I performed the following diagnostics: plotting variable importance plots, Partial Dependence Plot, performance across sub samples and displaying actual versus predicted prices

Variable importance plot indicated that the important variables for property price prediction is previous price, area size, balcony size, location and property condition. These variables are response for predicting more that 80% of transaction prices. Partial dependence plot illustrated that based on the available data the maximum number of rooms in cities is four. The mentioned partial dependence plot also demonstrated that there is large difference between the minimum and maximum number of rooms in capital region.

Sub sample performance showed that even though if a model performs well, there are instances where the model does not provide better performance in some sub samples. For example the best selected model, OLS, provided better result predicting small rooms than large.

On the other hand, actual versus predicted price plots assisted to deeper understand the result of partial dependent plots, PDP. Actual versus predicted prices showed that where model performs well in PDP, on average it has more observations and a narrower price range.

4 Findings

- It is important to narrow down the scope of research while predicting the price, because model may predictions well.
- This project demonstrated that to model transaction price we require a large data set.
- Based on the sub sample performance location is more important in the cities than in the capital region.
- We can see that Balcony size has higher importance predicting transaction price in the capital region than that of Cities.
- The project demonstrated that there is a significant price difference between capital region and cities. In the capital region prices difference is 10 million HUF. On the contrary price difference in the country side between small and large apartments is one mullion HUF.
- The project demonstrated that properties advertised by office on average are two Million HUF expensive
- Property sub type is one of the least important variables for transaction price prediction.

5 Recommendations for further Steps

- This section aims to provide recommendation or suggestions on the improvement of data collection. As mentioned, during the project it is important to collect more data for transaction prices.
- Data encoding should follow the provided formats, as there has been instances where various decimals formats have been used.
- All the steps from cleaning to reporting for the report is automated, it is important to collect more data on houses to build price prediction models.

6 Summary

The aim of this project was to evaluate transaction price of properties to build a prediction model. For this purpose the data for the project is provided by the client. Initially I cleaned the data file based on the major business decisions. First I filtered the data to keep observation with the transaction price, second I created factorized predictors from categorical variables, moreover, I created new meaningful variables to enrich the data to create prediction models. In order to evaluate models performance RMSE is used.

The cleaned data set was divided into two data work files; one for Budapest and the next one for the cities. Initially OLS, ordinary least squared regressions were used to evaluate model performance and select best model for the given data sets. Consequently, Model 3 performed the best for Budapest data set and Model 1 performed well for cities data set. Afterwards, Machine learning algorithms, such as CART, Random Forest and GBM, were used to cross compare regressions and algorithms prediction performance. As a result, OLS provided the best prediction result for both data set given the lowest RMSE.

Subsequently, to understand about patterns of association which drives prediction in the model this paper assessed sub sample performance. Sub sample performance demonstrated interesting finding about the model performance such as location is more important in case of cities than Budapest. It was a great learning experience, as I was encountered encoding errors and limited number of observations which helped me to work with real world data. I also learned as the nature of data differs across metropolitan area and cities, it is important to evaluate them using different models.