Predicting Flat Housing Prices in Budapest and Other Cities

Public Summary

By Syed Shah Ali Atraf Gardezi

In partial fulfilment of the requirement for the degree of Master of Business Analytics

Submitted to
Central European University
Department of Economics and Business

Supervisor: Gergely Daróczi

Budapest, Hungary

June 2022

Table of Contents

Executive Summary

With the rising uncertainty in property market after COVID and unfortunate socio-political situations such as Ukraine crisis ability to accurate predict the future property market trends such as price has become increasingly important. The aim of this project is to build an accurate prediction model to predict the advertised prices of the *flat* housing in the Budapest and other cities of Hungary on online property portal's website. The data for this analysis came from the internal database of company. During the prediction analysis we will see how different characteristics of a given property impacts the overall advertised prices. This research work employs simple modeling techniques and further leverages more sophisticated Machine Learning approaches to find the best fit model. The efficiency of the model is discussed based on the validated results and recommendations are made.

Problem Definition and Why am I trying to solve it?

Being one of the Hungary's leading online property portal, the company hosts a wealth of data. Thanks to digitalization and ease of use all the new buyers, whether they are investors or looking for new shelter began their search for property by searching online. Predictive analytics has therefore become pivotal to incorporate within processes to make better decisions and to stay ahead of the competition while providing value added services to their clients and reap better revenues.

Currently the company uses less sophisticated models to predict future prices and to make recommendations to the business teams. However, with the breath of information available, more often those models, because of their limitations, does not accurately predict the trends. This is where more sophisticated Machine Learning algorithms using their advanced predictive ability to cater more variables play a better role. This project aimed to employ simple linear regression model and several machine learning models such as LASSO (Least Absolute Shrinkage and Selection Operator), Classification And Regression Trees (CART), Random Forest, Gradient Boosting Machine (GBM) on the sample dataset shared by the company and estimate the results based on the best performing model. With help of this research project, the company will get the framework to apply on their larger data and understand ways in which it can fully leverage the tools available to make better predictions and hence serve their audience better.

From the business stand point this project has valuable advantages. Since we are predicting flat prices based on several characteristics of the property, the company can use this information and provide value-added services to the serious clients such as making recommendations on smarter valuation of their properties. For example, since our model considers the condition, area size and total room count of the flats, the company can advise their clients (property sellers) of the ideal pricing at which they should position their property in order to attract the buyers as they understand how much are the people will to pay for a flat in particular location with particular features.

Data & Cleaning

The data for this research was provided by the company. The data was the cross-sectional snapshot with several characteristics of the property which provided a good starting point for the analysis. The data was split to two based on cities as our goal was to make two separate prediction models, one for Budapest and the other one for Dunaujvaros and Szekesfehervar called the Other Cities. However, this data is collected from the advertisements on company's website that has been entered by people thus making it highly unorganized and raw. There were many incorrect entries for price variable, a lot of missing values and many duplicated rows. Much of time and effort went into cleaning of each variable. We had to rely on greatly on our domain knowledge as well as seek help of the industry experts from the company the during cleaning process. The missing values were either imputed or in cases where the missing values represented less than 5% of the observations, were simply removed. The duplicated entries of the same property were removed and in order to remove the extreme values, statistical decision was taken such as defining the interval of meter squared price and dropping those that fall outside of the interval range.

Analysis & Results

In order to ensure the validity of our prediction and avoid over fitting our results we divided out data into two groups, the training set and hold out set. This was doing using 80%-20% split where 80% of the data was used to train the models and 20% was reserved for the hold-out to test our prediction. We then created models of different complexity from variables in the data using our domain knowledge and observing the correlation between variables. The models

contained these variables and interaction terms. We first ran an OLS regression on these models using the K-fold cross validation and selected our best model based on the values of the lowest cross validated Root Mean Square Error (RMSE). After this ran series of prediction modelling algorithms to see which of the model gave us the best prediction in terms of lowest RMSE. The results were tabulated and compared and it was found that our best performing model was Random Forest for both prediction models. We chose the Random Forest with auto tuning in our case due to the negligible difference between auto tuning and basic tuning and trying ruling out the possibility of overfitting through our entered tuning parameters. The results of horse race of models are shown below.

BUDAPEST		
MODELS	CV RMSE	R-squared
OLS	12.19025	0.7688663
LASSO	12.20111	0.7684319
CART	13.93452	0.6979634
RANDOM FOREST: BASIC TUNING	11.50463	0.7942325
RANDOM FOREST: AUTO TUNING	11.55721	0.7925135
GBM	11.75244	0.7851295

OTHER CITIES		
MODELS	CV RMSE	R-squared
OLS	5.43913	0.7985033
LASSO	5.355468	0.8049217
CART	5.855333	0.7671033
RANDOM FOREST: BASIC TUNING	4.80155	0.8438052
RANDOM FOREST: AUTO TUNING	4.864602	0.8393978
GBM	4.910857	0.8361471

Since we selected Random Forest which is a *Black-Box* model, we performed a series of diagnostics on the hold-out data to understand how different predictors contributed to our prediction. These diagnostics tool included Variable Importance plots, Partial Dependence plot and Subsample Performance. Our model gave a better prediction for smaller area size flats, with a smaller number of rooms. This was because the number of those observations were higher in our data.

Limitations

1. The data received is highly uncleaned due to human inputs especially in the y variable which is price variable where the values are entered incorrectly with wrong number of significant figures. Cleaning those were a problem as we had to rely on domain knowledge or compare with other price variables in the data to get an idea.
2. The values entered for the variables of the same property even had different inputs making it hard to clean. It was still possible in the case of numerical variables as we could use statistical formulas calculation of mean to impute or however in case of categorical variables it was not so straightforward.
3. Data available is on county/district level due to which external data sources such as GDP, Inflation, literacy rate, purchasing power parity could not be added as they not available on county/district level. This poses a problem of the external validity of our predictive models on the live data.

4. The data is unbalanced, the number of observations of some type of feature is more than the others thus prediction of high observation groups are better than the others.

Lesson Learnt

1. Having right domain knowledge is very essential for the start of any project. One should get acquaint with the field in which one wishes to the conduct the research analysis in. This involves getting to know about the industry as well.
2. One should look of similar research that has been done in the field or the similar field. Look for the key findings and relationships.
3. If one wishes to use machine learning models, one should have right understanding about models their usage and their limitations tackle them where needed.
4. Defining the correct scope of the research project is very important. One cannot incorporate everything in one research. The narrower the scope, the clearer is the way forward. Consult your stakeholders about what exactly are they looking for and how will you be leveraging the resource to deliver.
5. Reproducibility of the work is very important. Leave comments on your code, document all steps, create a shared folder or version control for others to review or take your research forward.
6. The quality of data and how is it is collected is very important. If the data is untidy it and not collected correctly there will be errors some of which will be hard to mitigate and will thus affect the quality of your analysis results.
7. Data cleaning is most essential and takes most time if the data is untidy
8. Good communications with stakeholders are very essential for the success of project