FLIP ROBO

Housing price prediction project report

Submitted by,

Nidhi Gupta

Intenship 16

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

Problem statement

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry on housing prices . In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, in it is observed that Seattle apartments close to specialty food stores such as Whole Foods experienced a higher increase in value than average. This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties in Calgary between 2017-2020 based on data from . The aim of our project was to build a predictive model for change in house prices in the year 2021 based on certain time and geography dependent variables. The main steps in our research were the following.

Goals

We required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

observation

- • Exploratory Data Analysis (EDA). By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

- • Feature Selection In order to avoid overfitting issues, we select (according to PCA) variables out of the original by using methods checking skewness , outliers , forward feature selection, backward feature selection.

- • Modeling We apply RandomForestRegressor , LinearRegression and GradientBoostingRegressor models for prediction of the percentage change of the housing prices.

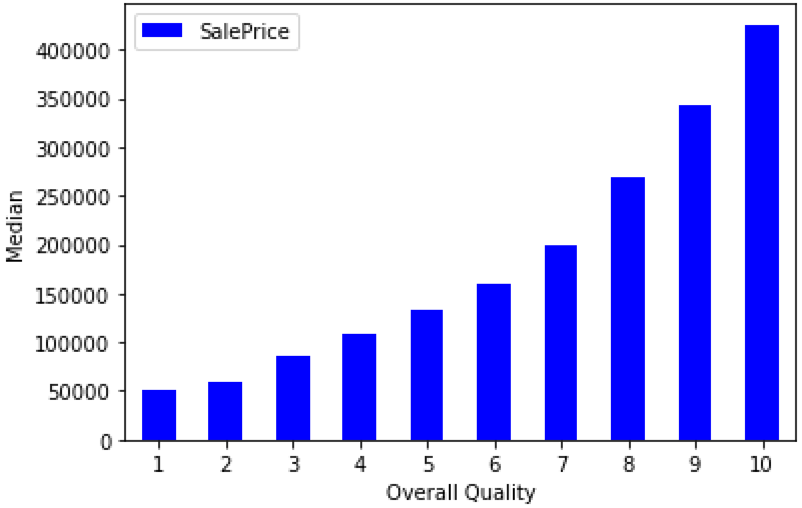
- • Exploration of reasons for misclassification in model We then go back to the original data to find out why some samples are misclassified by our model. In this report, we describe our approach to these steps and the results that we obtained.

Data Manipulation and Visualization

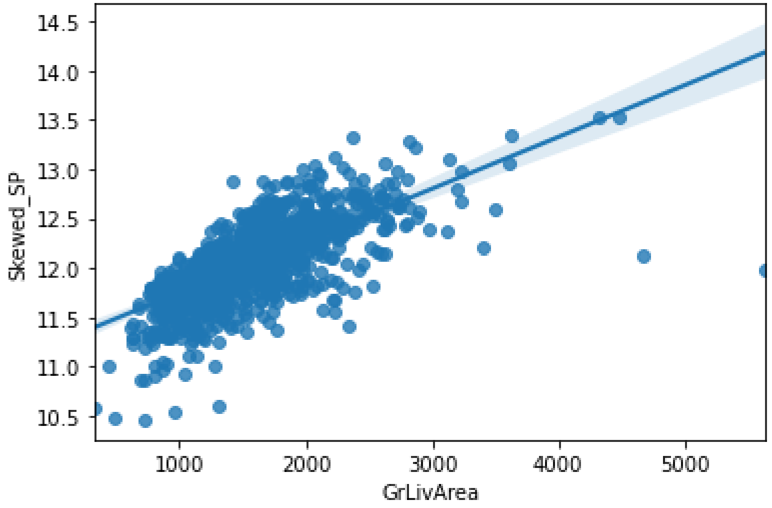
In order to understand our data, we first perform exploratory data analysis. This will provide us with insights that will be useful in building prediction models, as well as insights that may be of interest to stakeholders. As part of the Exploratory Data Analysis we aim to: Look into the relationship between each variables and annual house price percentage change, and identify any patterns. For example, between the year of construction of a house and its annual percent price change. We will also analyse relationships between the features. This may reveal that certain features are redundant and this would help the subsequent analysis.

Firstly we have checked for NaN (null) values in the given dataset.

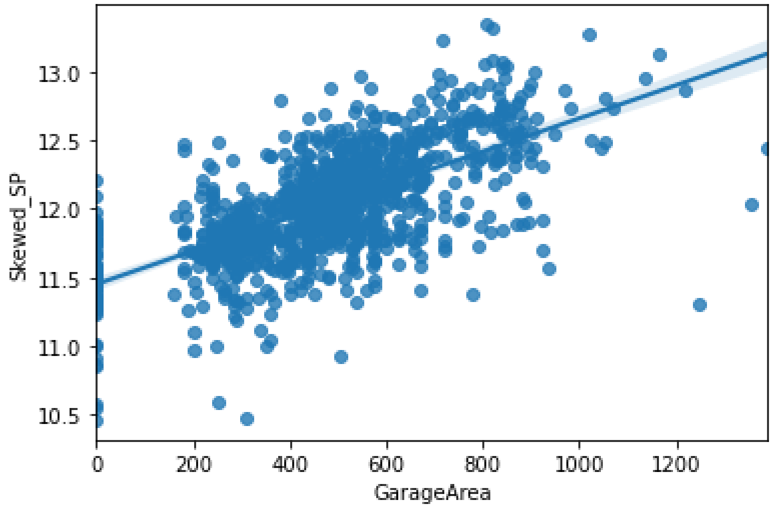
Second, we have checked for the mean, standard deviation for Sales price.



SalePrice varies directly with the Overall quality



SalePrice increases as the GrLivArea increases. We will also get rid of the outliers which severely affect the prediction



GarageArea and SalePrice are directly proportional.

conclusion

- According to observation R^2 IS ; -0.11386232534936935. and RMSE is : 0.1868001892542121
- Original predictions are : 11.94692097, 12.18112284, 11.93533502, 12.02779442,12.1123206
- Final predictions are : 154341.19374962, 195071.77317036, 152563.3227073, 167341.91904376, 182101.70050543