FLIP ROBO

Housing Project

Submitted by:

Neha

ACKNOWLEDGEMENT

The sources that helped me out in the completion of this project are, I usually take reference from my personal notes, internet sources and some websites and all of these sources helped me out for the completion of this project.

Introduction

Business problem Framing

As we know, houses are one of the necessary needs 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.

Conceptual Background of the Problem

The concept behind the problem is, 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 company is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

For this we should know, which variables are important to predict the price of variables, how do these variables describe the price of the house.

Review of Literature

The current research on the significant attributes of house price and analyzed the data mining techniques used to predict house price. Technically, houses with a strategic location such as the accessibility to shopping mall or other facilities tend to be more expensive than houses in rural areas with limited numbers of facilities. The aim is to explain the observed changes in residential property prices over time, describe them relationship to the factors of housing demand and supply, and employ these relationships for forecasting purposes.

- **The long-term fundamentals of house prices**

The long-term fundamentals of house prices are mostly chosen on the basis of a demand equation. Many models for house prices assume that the supply of housing is relatively inelastic in the short to medium-run and that it is hence essentially the changes in demand that explain variations in house prices. When housing is treated as a consumption good, its demand is generally, a function of a number of variables such as household income, interest rates, financial wealth, or demographic and labour market factors. Studies examining the performance of credit for house price forecasting point to mixed results. Access to credit and financial conditions are additional factors influencing house price dynamics.

- **The short-term momentum of house prices**

The empirical literature suggests that house prices are generally driven by momentum, i.e., the observed tendency for rising house prices to rise further, at least in the short term. past changes in house prices are included as explanatory variables in the overwhelming majority of studies, being also highly significant. The results indicate momentum in the short-run and reversal in the longer run. This autocorrelation structure could be seen as a purely empirical necessity because fundamental variables alone are typically not enough to explain house prices, but there are several theoretical models to explain momentum: irrational exuberance and unrealistic expectations of future price appreciation (Shiller 2005, 2009), risk-shifting behavior by banks related to agency problems and their expectations of continued credit growth (Allen and Gale, 2000), a procyclical behavior of housing sales (Wheaton, 1990), or down payment constraints in sellers' reservation prices (Stein, 1995). At the same time, the autocorrelation structure is typically found to be market specific and to differ across countries.

Conclusions

In general, the choice of a house price model and its empirical estimation is very much influenced by the quality and availability of data. While for most developed economies, sufficiently long time series of house prices and relevant fundamental variables allow error-correction or vector error-correction models to be estimated, this does not hold for a large number of Central and Eastern European Countries (CEE). Short time series, insufficient market coverage, lack of quality adjustment or distinction between old and new dwellings make forecasting much more challenging. The country size and the stage of economic development are identified as factors determining house price elasticities, with smaller countries and catching-up economies having higher responses to similar sized fundamental changes than larger and more well-developed economies

Motivation for the Problem Undertaken

My only objective behind this project is to use data Analytics and predict the house prices and help the companies to increase their overall revenue, profits, by focusing on changing trends in house sales and purchases. Thus, we can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

Analytical Problem Framing

- **Mathematical/Analytical Modeling of the Problem**

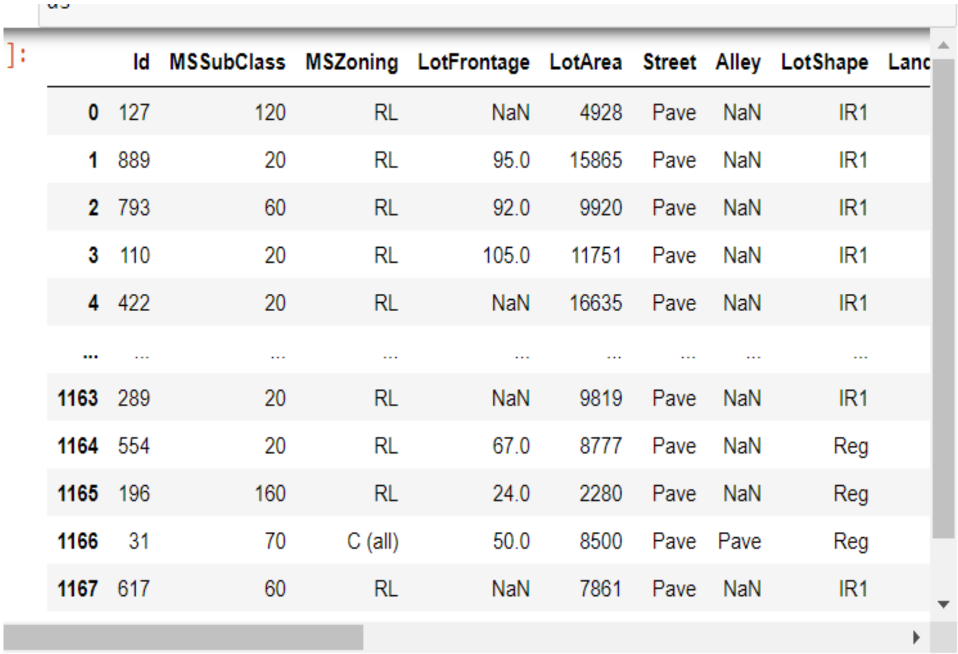
Mathematical /Analytical functions and methods are used such as df.shape () to check out the number of rows and columns in the dataset. df.dtypes() is used to check out the data type of the columns we have in the dataset. df.isnull() is used to check whether null values are present or not, we have also checked it by using Heatmap from seaborn. And the other one is checking out the statistical summary of the dataset by using df.describe() method. We got lots of fresh information from statistical summary is like:

1. We can see that in the count function many columns have different values, which clearly means many null values are present in the columns.
2. The difference between mean and median is not similar.

3. There is a small difference in 75% and max column in many columns like overallQual, OverallCond, YearBlt etc. which shows that there are few outliers present in the columns.
4. There is a large difference in 75% and max column in many columns like LotArea, MasVnArea, BsmtFinSF1, MiscVal etc. which shows that there are few outliers present in the columns.

Data Sources and their Formats

The data source is private, the datafile is provided in xlsx. format, We have imported the dataset in the file, the dataset has 1168 rows and 81 columns. We can see the snapshot of the dataset below:



]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	Lanc
0	127	120	RL	NaN	4928	Pave	NaN	IR1	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	
...	...	...	...	...	...	...	...	...	
1163	289	20	RL	NaN	9819	Pave	NaN	IR1	
1164	554	20	RL	67.0	8777	Pave	NaN	Reg	
1165	196	160	RL	24.0	2280	Pave	NaN	Reg	
1166	31	70	C (all)	50.0	8500	Pave	Pave	Reg	
1167	617	60	RL	NaN	7861	Pave	NaN	IR1	

Data Preprocessing Done

Following are the steps followed for data cleaning:

Missing Values if we have missing values then treat them by using suitable method, but we have no missing data in our dataset.

Label Encoding, we have used label encoding on the columns which were in string format.

Removing the outliers, we have few outliers present in the dataset, we can remove these outliers by using zscore or IQR, but as we have few outliers, we can ignore them.

Removing the highly correlated columns, we have from the correlation matrix that few columns are highly negatively correlated, thus we have dropped those columns.

Removing Skewness, we have observed that skewness is present in many columns, Soo we have removed the skewness by using power transform method.

Data Inputs-Logic–Output Relationships

The input columns of the data frame express the features provided in the house, so that by the help of these features we can predict the Saleprice of the house . The price of the house complrtrly depends in the features.

Hardware and Software Requirements and Tools used

The hardware device used is only the pc. In software devices so many applications and libraries are used to for the completion of the project. I have used Jupyter software for using jupyter notebook. The libraries used are pandas, NumPy, matplotlib, seaborn these libraries are used for analysis purpose. And from sklearn power transform is used for removing skewness.

Conclusions

- The key point is that we have a finalised trained machine learning model, that will help us to predictsaleprice of the houses.
- The prediction of house saleprice will help the companies to understand how exactly the price vary.
- The model will be a good way for the management to understand the pricing dynamics of a new market.