5.2 Assignment

PREDICTIVE ANALYTICS Case Study

Author : Venkat Jagadeesh Jampani

Date : Jan 15th 2023.

Introduction

What was the problem being solved?

The COVID pandemic has increased housing prices exponentially in US. Texas is one of the states where the house prices have sky rocketed. Among various cities in Texas, author selected the housing sale data available in and around Austin (capital city of state Texas) for this analysis. Author has explored several factors such as location, square foot, Zip code, property tax rate, school ratings etc., of the house affecting the sale price.

Objective of this project is to build a model that accurately predicts house prices in Austin.

Why was this problem important to solve?

Predicting home prices will help people who plan to buy a house so they can know the price projection in the future and can plan their finance well. It also helps to identify the features affecting the house price either increase or decrease. The main stakeholders for this project are sales/marketing team of real estate companies and realtors and people looking for house in general. In pitching the problem to these stake holders to gain buy-in, the focus is to understand the reason for increase in home price. As we are aware, pandemic has caused home prices to

skyrocket in almost all the cities across USA. This project could help the team to identify the features affecting the home price and focusing on those factors while buying and selling homes would eventually improve the revenue and net income of the company, and to the individuals looking to buy/sell home.

How was the data acquired?

In this project, the author worked with a dataset on the Austin housing data. The dataset is available on Kaggle at the below link:

https://www.kaggle.com/datasets/ericpierce/austinhousingprices.

The dataset contains 47 attributes and ~15000 records. The dataset contains house sale price for 3 years (2018-2021) in and around Austin, Texas area. Basically, this dataset is extracted from Zillow which tells the features of the house and sold price. "latestPrice" is the price that home has been sold and this will be used as target for our model. Some of continuous features available in the dataset as follow.

- Year of built
- Lot size square ft.
- Living area square ft.
- Average school distance
- Average School rating
- Number of bathrooms
- Number of bedrooms
- Number of Garages
- Price per square foot
- Number of price changes
- Number of appliances

Below are some of the categorical variables present in the dataset.

- Has Association
- Has Cooling
- Has Heating
- Has Garage
- Has View

- Patio porch
- Security
- Number of parking features
- Community
- Number of Primary schools
- Number elementary schools
- Number of middle and high schools

Methods and Results

What steps were taken to prepare the data?

Author has performed the following Exploratory Data Analysis and Cleansing exercise to prepare the data for modelling.

- Scaling target values for Time series
- Duplicates
- Outlier Detection
- Missing Data
- Binary Data
- Studying for our Target Variables

In addition, author has performed Natural Language Processing (NLP) on "Description Column" to understand the patterns of house and determine which features are important. Author used NLP to extract relevant information from the listing text to boost effectiveness of the model. Then author applied standardization on continuous variables and made sure that it has been applied only on training dataset.

How was this problem solved?

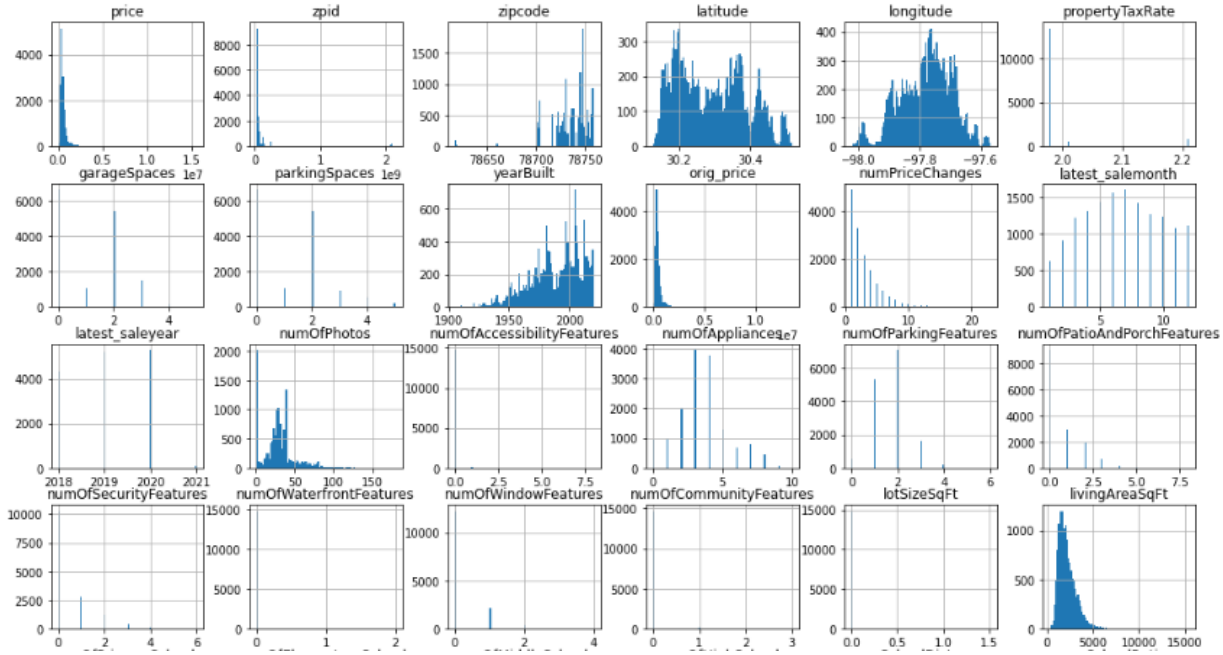
Initially, author has brought the home sale price to the same time scale. It was easy to ignore that these homes were sold over the space of many years, but a year is a long time for real estate. Author also appreciated the sale prices into most recent time series, using months.

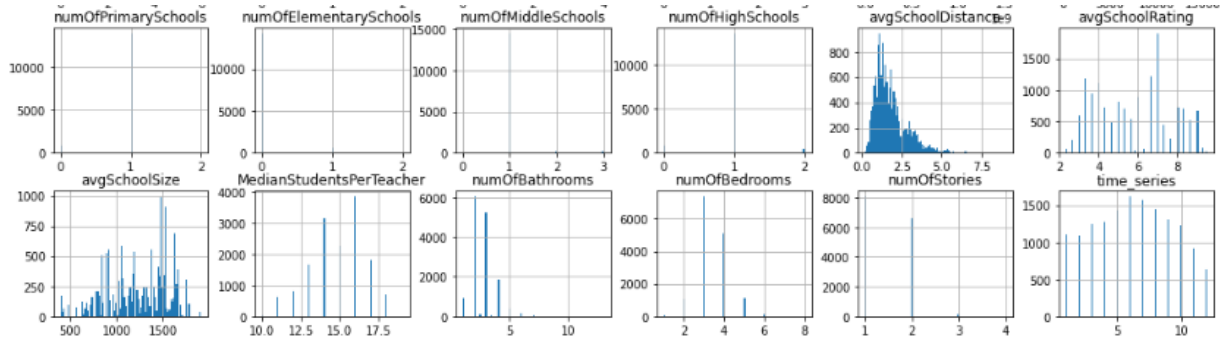
Then, author used normalization technique to choose the home types which make up most of the dataset. Since Single Family, Condo and Townhouse making up most of the data, author has removed all other multi-family type listings. Duplicate and missing values checks have also been performed and found neither duplicates nor missing values.

Dataset contains numerous ordinal features like numOfAccessibilityFeatures, numOfAppliances, and numOfParkingFeatures with values as True or False. So, all these values have been converted to binary values of 0 and 1.

Outlier has been detecting by plotting scatter plot for latitude and longitude. There doesn't seem to be any zones that are well outside of the Austin area, except for just a few down in the lower SE area. So removed the data having latitude just above 30.1.

Histogram has been plotted for all the features to remove the outliers from the dataset. For the square footage variables, large houses and lots are so seriously under-represented in the dataset that we won't be able to reliably predict on them anyway and they are better left off. IQR method has been followed to remove outliers in livingareasqft and lotsizesqft. The outliers present in some of the other features like zipcode, have been removed manually by looking at the data.

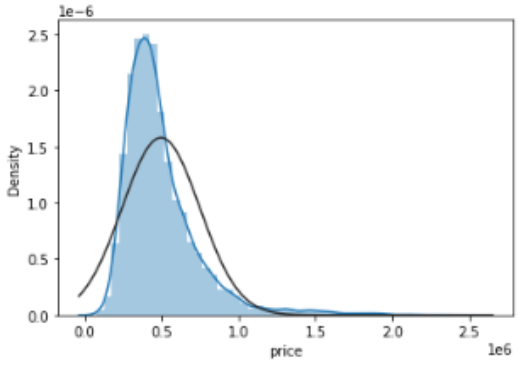




Target variable "Sale Price" has been analyzed by plotting Histogram and found positively skewed and has a high peak. QQ-plot shows that target has heavy tails with right skew. Author calculated kurtosis and the value indicates lots in the tail.

What modeling techniques were used?

Housing Sales is the target variable in the dataset. Linear regression model has been chosen since sale price is continuous variable. Different variations of linear regression model have been evaluated on the dataset. Below are some of the models implemented by author.



- Basic LR with Top Features One-Hot Encoded
- Basic LR with Top Features Target Encoded
- LR with ALL model features
- Linear Regression with various Feature Selection Methods
    - Permutation Importance
    - Forward-Backward Selector
    - RFECV
- K-Nearest Neighbors
- Support Vector Regression
- XGBoost Models
    - XGBoost - One Hot Encoded
    - XGBoost - Target Encoded

What metrics were used to evaluate the results? Why was this metric chosen?

The following metrics are used to evaluate the result. These metrics are primarily used to evaluate regression model results.

- R^2 - The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model.
- MAE - The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.
- RMSE - Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

The model utilizes a combination of continuous variables and one-hot-encoded categorical to produce a support vector machine regression with R^2 of 79.2%, a mean absolute error of 64k, and a root mean squared error of 105k. Several zip code transformations including polynomial features, mean target encoding, lower-granularity binning, and median rank as a continuous, and ALL of these efforts resulted in a lower R^2 and higher mean absolute error, leading to a final decision to one-hot encode all zip codes individually.

Conclusion

How were the results or model implemented?

Initially, high value for MAE and R^2 as 71.3% have been seen by running the model. So, author decided to standardize the features using StandardScalar method. Several of the predictors chosen for linear regression has p-value over 0.05, which indicates that there is more than a 5% chance that the changes attributed to that feature were actually by random chance. Then, feature selector has been used to choose the best features from the dataset. Author has tried to improve the result by running various models and following is the output for those models.

	r2	mae	rmse
Models			
SVR	7.916000e+01	63971.02	104935.12
XGB - Encoded	8.133000e+01	66776.46	105886.40
XGB - One-hot	7.984000e+01	68528.25	108696.28
Ridge	7.747000e+01	71298.55	111843.75
LR w/RFECV	7.746000e+01	71510.21	112359.40
LR All - One Hot	7.741000e+01	71513.72	112360.81
LR w/Forward-Backward Selector	7.716000e+01	71545.24	111757.54
LR w/Permutation Importance	7.679000e+01	73547.43	115514.83
Lasso	7.604000e+01	74124.95	116235.12
LR All - Encoded	-2.981080e+20	81398.01	123291.72
Basic LR - Top Features Only, One-Hot	7.132000e+01	81867.41	126034.88
KNN	6.931000e+01	87579.03	139224.82
Basic LR - Top Features Only, Target Encoded	6.199000e+01	100869.56	148486.63

Plotting MAE calculated by various modelling techniques

Among various features present in the dataset, square footage is, unsurprisingly, a key player in house pricing. And as they say, location is everything, and it is the primary influencing factor for a home price in the Austin metro area. Number of bathrooms, school rating, and lot size all contributed a large amount as well. These five features alone explain 71% of the price variance.

What were the actionable consequences of the case study?

After Covid-19 pandemic, the house price across USA particularly in Texas have sky rocketed. This case study on house price prediction can be used to predict the price of the house. At the same time, this modelling demonstrates the features affecting the house price. This modelling cannot be used as-is. Rather sourcing better data, or easy-to-use features that an average realtor is capable of evaluating/acquiring should be added to the model to improve its predictive quality.

What did the team learn from the case study?

The model, while explaining about 80% of the price variance with our features, was nonetheless far from accurate in absolute terms. A mean average error of 64k in either direction is a huge variance to a home price - one that is so large that it renders the prediction much less meaningful and useful. The model done by the author can be a best baseline starting point.

How should or would the team approach the problem differently in the future?

As mentioned by the author, this model can be used as baseline starting point for house prediction. The data set to be missing some key features that have high influence in other housing sets, foremost among them reasonable metrics of home condition, home quality, and neighborhood quality. Author tried to pick up some quality and condition metrics from NLP but it is not complete. In addition, number of external factors mentioned below that I considered important are missed out while considering the data. So, the model can be improved by choosing the dataset having these additional features.

- Interest rate
- Environment factors like pandemic
- Work from home making people to migrate to low cost of living area.
- Property tax of the house, including Mud and Pid tax

Reference

https://www.kaggle.com/code/threnjen/austin-housing-eda-nlp-models-visualizations#Model-Explorations