Final Report: Prediction of Austin House Prices

Problem Statement

Austin is ranked as the 7th best real estate market in the U.S. by Wallethub, based on key investment criteria such as activity in the housing market and strength of the economy. Austin real estate is the second-most expensive real estate market in Texas for those seeking to purchase a large home for their growing needs. Austin real estate prices are considered 10% higher than what would be a fair price on today's market considering long term trending prices, price to income and price to rent ratios of the areas. Given this, the objective of this project was to build a predictive model to determine house sale prices.

By using the Zillow Dataset, I developed a model as reference tool for helping the apartment management companies and apartment developers. Different exploratory data analysis methods and supervised learning methods were utilized to develop a model for this.

After cleaning my data and reducing my features from 47 to 30, my tuned decision tree model was able to predict the house prices with the maximum score of 0.69 and the least mean absolute error of 19% , compared to the other models.

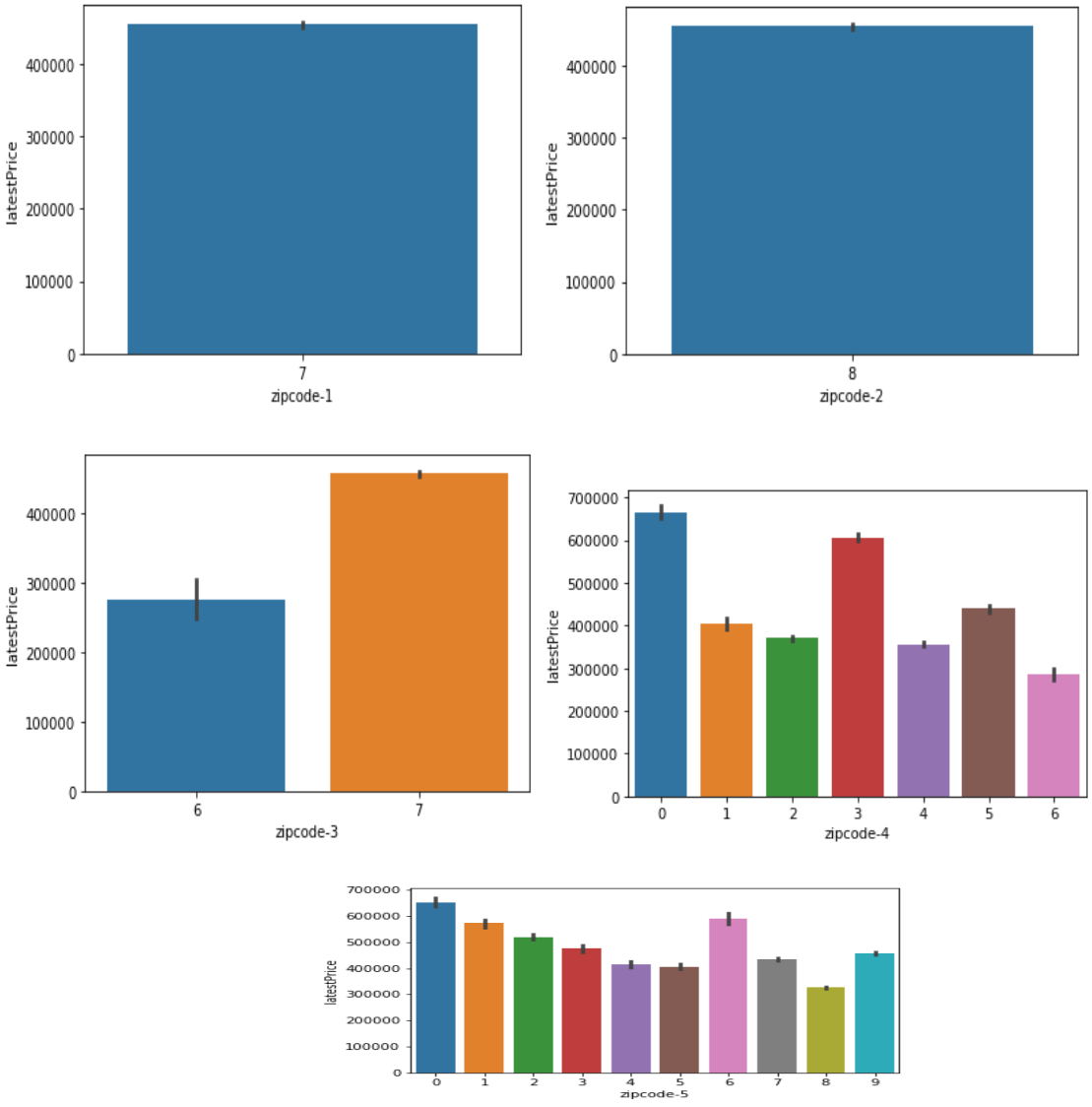
Data Wrangling

The initial dataset consisted of 15171 rows and 47 columns. In the initial data set, single family constituted more than 90 percent of our dataset. Hence, all the homes other than single family homes were removed from the dataset. After including only single family types, now there are 14241 rows and 47 columns in the dataset. The dataset includes latest home price, year built, location of homes, amenities and facilities available at each home along with the features and rating of schools near to each home. There were no missing and duplicate values found in the dataset. Outliers were examined separately for each data point and examined if they were entered incorrectly or if the measurement was legitimate. In case of incorrect entries, the outliers were removed and for legitimate measurement, they were kept as such. All the non numerical variables were converted into numerical variables by grouping them in different categories. The final data set consisted of 13703 rows and 47 columns after the data cleaning process.

Feature Engineering and Exploratory Data Analysis

The relationship of different factors that can possibly affect the home prices with the latest home price in Austin was examined using visualizations and correlations. One main trend that was seen was the effect of location on the home price. To exploit maximum information from the feature zip code it was regrouped into different features based on the different digits in each zip code. For example if the zip code is 78660, the first digit is categorized as one feature, second as other, and so on. Thus there are five features representing each digit in the zip code after regrouping zip code. The first and second digits were not found to have any relation with the home price while the next three digits were found to have relation with the prices. Figures below illustrate these.

Figure 1:Zipcode and Price



Hence in the final model, the third, fourth and fifth digits of the zip code were only included and as three separate features. Some other features were regrouped into categories and new features were created to exploit maximum information. For example, the feature "year built" was regrouped into 0 and 1 for all those with year built greater than 2010 as 1 and year built less than or equal to 2010 as 0. These new features were included in the final model. The preliminary analysis based on scatter plots, box plots and correlations shows that along with the third, fourth and fifth digits of the zip code, the features that were found to have a relationship with the home prices include number of photos, living area and lot size, whether its new construction or not, cooling and spa facilities, number of schools like primary, elementary, middle and high schools, number of appliances, number of parking features, number of Patio And Porch Features, number of Security Features, number of Waterfront Features, number of Window Features, number of bathrooms and bedrooms, average school distance and average school rating, garage and parking spaces, number of accessibility features ,property tax rate, has View and number of Stories. Thus in total, there were 30 features in the final model and the dependent variable is latest home price. The remaining 17 features were excluded in the final model based on preliminary trends.

Modeling

Three models were tested for predicting house prices: Linear Regression, Support vector regression and Decision trees. The metrics that were focused include Coefficient of Determination R^2, regression scores and the error metrics namely RMSE, MAE and MAPE. The data was divided into train and test data sets with a ratio of 80: 20. The models were fitted on the train data and the predictions were validated on the test data. For linear regression, the coefficient of determination for the training data is 90.5 percent while that on test data, the regression score is 52.3 percent. While using support vector regression, the regression score on train and test data were obtained as 0.5 percent and -0.3 percent respectively. The regression scores on train and test data for decision trees were 99.9 and 49.3 percent respectively.

Table 1 shows the error metrics for the three models

Table1 : Error Metrics for the initial models

Model	RMSE	MAE	MAPE	
Linear regression	150522.962	107335.370	26.4%	
SVR	218215.1636	147588.97	33.23%	
Decision trees	155058.2910660444	102040.733	23.9%	

Based on the scores and the error metrics, decision trees was selected as the best model for the analysis but it is over fitting since test score is very less compared to the high training data score. Hence, the hyper parameters of decision tree model were tuned using randomized search cv and the tuned model shows a score of 77.1 percent on train data and 68.8 percent on test data, thus removing the problem of over fitting. The error metrics based on the final tuned model is given below.

Table2 : Error Metrics for the final model

Model	RMSE	MAE	MAPE
Hypertuned Decision trees	121516.9039	81612.507	19.69%

Thus the performance of the hyper tuned model has improved much compared to the original model, as seen from the test scores and the error metrics. Hence this model is chosen as the final model.

Conclusion

The objective of this model was to build a predictive model for Austin house prices. The final model reached RMSE of150522.962, MAE of 81612.5 and MAPE of 19.7% much lower than the baseline linear regression model(RMSE=121516.90,MAE=107335.370 and MAPE=26.4%).The final model show the main factors affecting the house price include living area in square feet, location of the house represented by zip code and the lot size in square feet. The least important factors affecting house price are found to be the view of the house, number of water front features, heating facilities, number of primary schools in the location and number of middle schools in the location. A comparison between the actual and predicted values for the test data set shows where the model performs well and where the model performs poorly. This comparison shows that the model performs very well in areas where actual price is 325000. The zip codes from 78702 to 78759 have the actual price of 325000(predicted=324999.5).

There is only a difference of 0.5 in these areas of Austin between the actual and predicted house price. In many other areas, the difference is in the range of 100 to 10,000.The model performs poorly in the zip codes that have actual price 1300000 and 325000.

Takeaways/Recommendations

Decision trees were found to be outperforming all the other models in terms of scores and evaluation metrics compared to the baseline linear regression model and the support vector regression model. While hyper tuning the parameters, the performance of decision trees improved much better. Many of the features were not found important in affecting the home price and they were removed from the final model.

The house searchers might identify those apartments in Austin area that are good deals or bad deals based on the model. For example, in the areas with zip codes like 78744,7813 etc the actual price is only $ 325000 while the predicted price can be up to $ 1080545.This can be considered good deal for the house searchers. From a seller's perspective, this model can help them to make price deals that are competitive with respect to the market competition. For example, for the zip codes 78701, 78704, 78733 and 78759, the actual price is $1049000 while predicted price is only $ 513046.7. Thus, these are overpriced deals that the sellers need to be aware of.

Future Research

The analysis can be further extended by including boosting models like XG boosting regression models and bagging models like random forest regression models and then examine how those models perform in comparison with the baseline model. Moreover, the features can be reduced further by making use of principal component analysis so that the model performance is improved.