**AN ANALYSIS ON FACTORS AFFECTING PRICE PREDICTION IN THE REAL ESTATE MARKET​**

**Principles of Business Data Mining**

**Team 4**

**05/05/2022**

**Spring 2022**

1. **Executive summary**

Predicting sales prices is a major aspect of working in the real estate industry. As a real estate agent it’s crucial to make an accurate prediction to give the right valuation for our customers. Our business objective is to make the most efficient data driven decision using statistical modeling to make price prediction. For this analysis, we will be looking at an industry renowned company “**Zillow**” as a reference point to understand current price prediction trends. In the past, after Zillow's economic research team preprocessed and refined data, they perform time series analysis for price predictions. Conventionally managers at Zillow utilize non architectural features such as crime rate, number of hospital and school districts to predict prices.

1. **Business question**

What is the predictive power of specific housing features in estimating sales price?

1. **Data description**

For our analysis, we will be using the Ames housing data available on Kaggle. This data set is a unique resource because it includes various features that describe a wide range of characteristics that are not limited to school district and number of hospitals. The data has 79 variables and 1460 observations. 42 categorial variables and 37 continuous variables. The target variable is **Sales Price**. Approximately 5% are binary and 95% are categorical. The Average Sales price is 180921.196.

1. **Preprocessing using visualization**

As preprocessing tools, a heatmap and scatter plots were the chosen tools for our dataset analysis. Looking at our scatter plot some of the main variable relations we can observe demonstrate that a higher overall quality has higher sales price (Figure ), a property with ground level area of 3000 square feet sells for a maximum amount (Figure 2) and bigger lot area yields high prices (Figure 3).

|  |  |  |
| --- | --- | --- |
| A screenshot of a computer  Description automatically generated with low confidence  Figure 2 SALES PRICE V OVERALL QLTY | Chart  Description automatically generated with medium confidence  Figure 3 SALES PRICE V GRLVG AREA | Figure 4 SALES PRICE V LOT AREA |

Looking at the heat map (Figure 5), the only variables that have high correlation are some non-architectural features, for example there seems to be high negative correlation between Total Rooms Above Ground and the Year Built variables. In this case, for our data, data reduction does not seem necessary even though it is a big data set. Therefore, Principal Component Analysis was not necessarily due to lower correlation between variables. We utilized all 79 variables in the data set to analyze the predictive power of specific housing features in estimating our target variable **sales price**.

Chart, treemap chart

Description automatically generated

Figure 5 Heat Map

1. **Prediction models and Findings**

Our chosen software for this project was SAS Enterprise Miner (SAS). The data Mining models we used are Linear Regression, Stepwise Regression and Classification Tree. SAS was used to perform Regression and Classification Tree to understand our most significant variables which in turn help us answer our business question.

We took a four-step approach into utilizing our data sets and predictive models. STEP 1: We partitioned the data a couple of ways to find the optimum R2 and adj R2 values. We found the split of 60(training) - 20(validation/optimization) - 20(testing) to be the most optimal split.

STEP 2: We ran linear regression and stepwise regression just to confirm and cross-check the output. STEP 3: We run a classification tree to determine the accuracy of our regression models and we finally ran all our models through a model comparison to understand the best performing model.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). What that meant for our data is we will accurately be able to predict our sales prices (our target variable) based on many input variables.

Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration.

When we tried splitting our data a couple of different ways, we found various R2 and adj R2 values ranging up to 0.78. Considering the vastness of the data set, we were happy with our split to be 60-20-20, which gave an adjusted R2 value of 0.78 as you can see on this zoomed in output screen. Adj R2 is more unbiased outcome and thus 0.78 was a good figure to go ahead with.

When we compared the two regression models, linear regression was selected. But stepwise regression was important to understand the most significant variables as the linear regression only gave us the top 3 significant variables. The stepwise regression on the other hand gave 17 most significant variables that affect the sales price. We thought it was logical to cross-check our results using a classification tree. We wanted to understand how the model would split the tree. What will be the first split based on? Will it be one of the variables that we found in our regression?

Graphical user interface

Description automatically generated with medium confidence

Figure 6 Linear Regression Output

We ran a classification tree for our continuous output variable. We pruned the tree to reduce the complexity of the final classifier which helped increase the predictive accuracy by reduction in overfitting. We also reduced the depth of the tree to make it look precise. The classification tree was split on the Overall Quality variable, which was not surprising, because in our project proposal, we had predicted that that would be our most significant variable, which is only logical if thought about conventionally.

Graphical user interface, application

Description automatically generated

Figure 7 Classification Tree Output

After running all our three models in the model comparison, Classification tree was chosen. This is simply because it has a lower misclassification rate and gives a better accuracy(overall). The classification tree accurately predicts the sales price based on a couple of variables and we deemed that desirable. Unlike Zillow, the company we’re comparing with, our misclassification rate (in a way the mispricing rate) is extremely low.

Graphical user interface

Description automatically generated

Figure 8 Model Comparison Output

1. **Managerial or Policy implications and implementations**

Conventionally managers at Zillow utilize non architectural features such as crime rate, number of hospital and school districts to predict prices. In addition, they use time series analysis. However, after running our analysis, we can conclude that a classification tree is a better choice.

For Zillow, they use a time series analysis which requires a time-series data. In some cases, they reported that their predictions gave over-priced estimates. In our case, since we use a classification tree, we can make a more accurate price prediction which does not require time-series data. Based on our project’s results and managerial insights

The classification tree using the overall quality as a key factor narrows down a property selection for managers trying to sell a house. For instance, if a client is looking at a specific price range, managers can save time and resources by looking at houses on one side of the tree. By using this implementation our classification tree can be highly beneficial.

Diagram

Description automatically generated

Figure 9 Pruned Classification Tree

1. **Conclusion**

The purpose of this analysis was to find the most optimal price prediction model for the real estate market. We based our models around our business question” What is the predictive power of specific housing features in estimating sales price?” Thanks to our unique data set we were able to make a thorough analysis of the majority of features. We were able to determine architectural features as driving forces to how a house is priced. Additionally, by implementing our model comparisons we were able to conclude a classification tree is the best model for price prediction.

1. **References**

Kaggle.com. 2022. *House Prices - Advanced Regression Techniques | Kaggle*. [online] Available at: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data> [Accessed 8 May 2022].

Medium. 2022. House Price Forecasting using Zillow Economics dataset. [online] Available at: <https://towardsdatascience.com/house-price-forecasting-using-zillow-economics-dataset-2b58e6cd1c03> [Accessed 8 May 2022].

Wikipedia. 2022. *Wikipedia, the free encyclopedia*. [online] Available at: <https://en.wikipedia.org/wiki/Main\_Page> [Accessed 8 May 2022].