**George Mason University**

**STAT-515: Applied Statistics & Visualization for Analytics**

**Course Professor’s Name- Tokunbo Fadahunsi**

**Final Report**

**Predicting House Values in Ames City**

**By**

**G01285455 - Rahul Ravikanth Bhonsle**

**Logo, company name

Description automatically generated**

**Abstract:**

Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rater than being based on valuation. The phenomenon of the falling or rising of the house price has attracted interest from the researcher as well as many other interested parties. Predicting house values in Ames city based on real factors is the main crux of this project. Aim of this project is to make evaluations based on every basic parameter that is considered while determining the value. We make use of various Data Mining techniques such as linear regression and random forests in this pathway.

**Introduction:**

Development of civilization is the foundation of increase of demand of houses day by day. Accurate prediction of house value has been always a fascination for the buyers, sellers and for the bankers also. Every Property has distinct features that varies from other properties, and no two properties are same. The intrinsic heterogeneity of land properties, as a side-effect of varying construction, financing, area, and features. There are three normal methodologies for esteeming land, the income approach, the cost approach and the market approach. The income approach values real estate based on rental projected cash flows, the cost approach depends on replacement cost for construction of property, the market approach depends on the universe of similar real estate properties that have been sold.

**DATASET:**

Dataset that is being used for this project is namely “**Ames lowa housing data**” this dataset was provided as part of **regression analysis competition hosted on Kaggle.com**. Dataset holds 1460 observation of real estate property and 81 variables. Since the dataset is only specific to Ames City in lowa the dataset is not very large and it will be difficult to build a model that will make use of all variables, so to proceed with model through analysis of datasets variables is made to determine which variables are important.

After through analysis out of 81 variables the following variables shows key features for model development: a) zoning classification, b) lot area, c) neighborhood, d) building type, e) quality, f) year built, g) year remodeled, h) central air, I) number of bathrooms, j) kitchen grade, k) number of rooms, l) functionality, m) garage area, n) grade living area, and o) sales price. Dataset has been filtered with the above-mentioned variables.

**Data Visualization:**

**Categorical Variables:** The filtered dataset consists of few categorical variables which needs to be cleaned before development of model. To identify which record needs to be deleted or treated we have plotted the graph for the same as shown in the below attached Figure 1.

Chart, pie chart

Description automatically generated

Figure :Plot of categorical variables

**Note**: Each plot has been done separately but for convenient I have made use of ggarange to hold list of plots.

I could see real estate values vary between tenant types and in the dataset, there were five classifications present: Residential High Density (RM), Residential Medium Density (RM), Residential Low Density (RL), Floating Village Residential (FV), and Commercial (C). With the help of bar plot I could see that RL zoning is majority in the dataset nearly 1200 records. Accordingly the dataset is transformed this variable to binary, 1 representing RL and 0 representing other zoning classifications. Next plot is a piechart that shows the building types : Single-family detached, Two-family converstion, Duplex, Townhouse end unit, and Townhouse inside unit. With the help of pie chart we can see the majority of building are of type single-family detached . similar to the previous step we transform the dataset with the variable 1 for single-family detached and 0 for all others.

The rightmost plot is based on neighborhood, After ploting we could see 25 levels of neighborhood, the datset has been transformed the neighborhood variables to binary where the four largest neighborhoods are represented with value 1 and the remaining neighborhood values have been set as 0.

Neighborhood values have been also ploted with interactive bubble chart using ggplotly() function of the plotly library. In this plot when you place your mouse pointer on the bubble it gives the neighborhood name along with the value as shown the figure 2. Higher the couunt larger the size of the bubble.

**Chart, scatter chart, bubble chart

Description automatically generated**

Figure : Bubble chart of neighborhood

**Response Variables**: In the dataset the response variable or the class variable is sale Price, to build an accurate model the first thing is to understand the response variable, and this can be done by visualization of the distribution of the response variable (sale price) by means of histogram which is shown in the below attached figure 3. In the leftmost plot which is histogram of sales price which clearly shows the presence of outliers, and the distribution is Right skewed. From the plot it can be seen there were multiple values above $500k, and it is reasonable that within the area there can be small number of houses which will be high priced. The histogram shows the peak concentration around price $150K, with a mean that’s slightly above $200k.

The Right most plot is logarithmic distribution of sales price which is in-line with normal distribution using which a better model accuracy can be obtained. Not doing so will violate the assumption of linear regression model.

Chart, histogram

Description automatically generated

Figure : Distribution of sale price

**Continuous Variables**:

Chart, box and whisker chart

Description automatically generated

Figure :Visualization of Continuous Variables

In the dataset we could find some continuous variables such as Total Rooms, Number of full baths, number of half baths, Above Grade living area and Garage area. We have performed descriptive plots for these five variables as show in above figure 4. In the upper chart we can see the relationship between total rooms vs sales price. The plot shows the linear relationship between 0-10 rooms. This indicates there were negligible changes in the log of sale price after 10 rooms.

Further we can see scatterplots for above grade living area and garage area indicates the linear relationship and which can be justified by the points being closely inclined towards the line, and thus did not warrant the transformation of those variables. The below to graphs with respect to number of full baths and half baths against sales price did not seem to heavily indicate the existence of relationship that may not be explained by normal distribution.

**Correlation Matrix:** Correlation Matrix shows the correlation coefficients between variables. So, we will be using correlation matrix to know the correlation between continuous variables within dataset. We make use of function corrplot to visualize a correlation matrix.

Chart, scatter chart

Description automatically generated

Figure :correlation matrix

From the above matrix we can see the correlation coefficients within the dataset. This provides the information on which variables needs to be selected to build our model.

**Data Model:**

Initial we split our dataset into training and testing data through random selection, In which 70 % of the data is allocated to a training set and the remaining 30 % is allocated to testing set. Which will be further used in our regression analysis.

**Linear Regression:**

First step is to select variables that play major role in model accuracy to do so we ran a best subsets regression on training set, using backward stepwise method. The best subsets regression model indicated significant improvement to R-Squared up until 11 variables, where incremental benefit was negligible (see figure 5). We notice from figure 5 that R-Squared starts with an RSQ of ~62.5% and tops at ~87.5% at 11 variables. With that, we determined that it was best to fit a linear regression model with 11 variables and, as such, filtered the training set to include only the best combination of 11 variables.

Chart, line chart

Description automatically generated

Figure : Variable selection

After running the model, we get the results as follows in the below attached figure 7.

A picture containing text, battery, receipt

Description automatically generated

Figure : Linear Regression summary and VIF Table

Note: Plots for Linear Regression Diagnostic with outliers and without outliers is also performed but not mentioned in the report but can be found in the R code provided.

**Random Forests:**

Next model that we are making use of is Random forests or random decision forests. Random Forest is a powerful machine learning algorithm that can be used for a variety of tasks, including regression and classification. It is a decision tree in a semantic ensemble method that arbitrary forest models consist of a small number, each producing its own prediction. Here we are making use of mtry 5 which indicates the number of variables that are randomly sampled at each split and arrived at an R-Squared equivalent of 87.31%.

Text

Description automatically generated

Figure : Random Forest Result

Six Variables that played major role were Above Grade Living Area, Overall Quality, Year Built, Garage Area, and Year Remodeled, as shown in the below figure 9.

Table

Description automatically generated

Figure : Important variables

**Conclusion:**

In this project I have performed data selection through the help of visualization and built the model for regression analysis using two machine learning concepts such as liner regression and Random Forest and after training the data and executing the model it comes to be random forest giving the better and optimal results with an R-Squared equivalent of 87.31%. Real estate poses challenges that are not fully captured by the data. There can be human psychological factors that affect the house prices.

**References:**

Dataset: Ames lowa housing data: <https://kaggle.com/marcopale/housing>

The R Graph Gallery: <https://www.r-graph-gallery.com/index.html>

R- Linear Regression: <https://www.tutorialspoint.com/r/r_linear_regression.htm>

House Price Prediction Using Regression Techniques: A Comparative Study: <https://ieeexplore.ieee.org/document/8882834>

R-Random Forest: <https://www.tutorialspoint.com/r/r_random_forest.htm>

Log Transformation: <https://calcworkshop.com/linear-regression/log-transformation/>

Lecture notes and examples (STAT-515): <https://mymasonportal.gmu.edu/ultra/courses/_440479_1/cl/outline>

ggarrange: <https://rpkgs.datanovia.com/ggpubr/reference/ggarrange.html>

Model Selection: <http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/>