**Project Report**

**Business Analytics**

**Professor: Asadi Shahla**

**Akshaya Mamidipalli**

**Problem Description:**

Zillow's Zestimate home valuation has shaken up the U.S. real estate industry since first released 11 years ago. A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first-time consumers had access to this type of home value information at no cost. "Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning. This project is a very simplified version of the Zillow Prize competition. Zillow Prize was a competition with a one-million-dollar grand prize with the objective to help push the accuracy of the Zestimate even further. Winning algorithms stand to impact the home values of 110M homes across the U.S.

You test and compare the following three models.  
A) Build a regression [module 5] and decision tree [module 7] model that can accurately predict the price of a house based on several predictors (you select appropriate features).  
B) Use classification [module 6] to model OverallQual (rating 7 and above is considered as class 1, otherwise class zero).

1. **Project Goal:**

The main goal of the project is to develop Regression, Decision tree, and Classification models and implement the model on the test dataset to find its accuracy.

1. **Overview of the data:**

There are 3 kinds of variables which are:

1. Area related.
2. Date related
3. House facility related.

Under,

**Area related:**

LotArea: Lot size in square feet

GarageArea: Size of garage in square feet

BsmtFinSF1: Finished square feet

**Date related:**

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

YrSold: Year sold

**House facility related:**

FullBath: Full bathrooms

HalfBath: Half baths

BedroomAbvGr: Number of Bedrooms above the ground

TotRmsAbvGrd: Number of rooms above the ground

Fireplaces: Number of fireplaces

**Data Exploratory Analysis:**

**A graph of a graph showing a number of numbers and a number of numbers

Description automatically generated with medium confidence**

The image above shows a scatter plot of sales price and lot area with the overlaid regression line and coefficient of determination (R^2) of 0.734.

Displaying Positive Correlation,

The data points' rising trend implies a favorable relationship between sales price and lot area. This means that properties on larger lots tend to sell for more money.

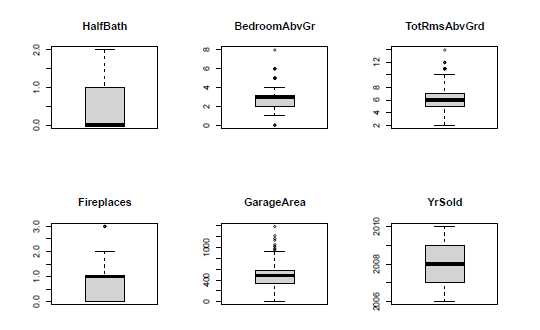
Displaying the Strength of the relationship,

The R2 score of 0.734 suggests that the two variables have a moderately strong association. However, it also suggests that the lot area does not explain 26.6% of the variation in sales price.

We have also plotted box plots for all the variables in the dataset as below:

A group of graphs with text

Description automatically generated with medium confidence



A graph with a line and a line

Description automatically generated with medium confidence

In all the box plots shown above, we see that the outliers are present in most of the variables of the data.

1. **Modeling Strategy:**

Here, we are predicting the sales price of the houses in the given dataset where sales price is the dependent variable and we have performed all the models for all the available variables based on the output of R-squared values, I choose “GarageArea, YearBuilt, FullBath, OverallQual, LotArea, BsmtFinSF1, YearRemodAdd, Fireplaces” as the best predictors for my model.

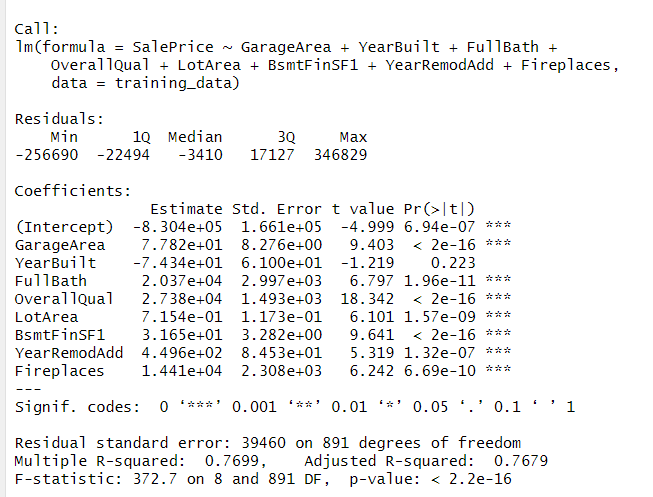
I trained the dataset with all the variables on trial and error basis until I got the optimal R-squared values and selected those variables as my predictors for all the models i.e., Regression, Decision-tree and Classification.

1. **Estimation of the model’s performance:**

Let’s establish the models starting with.

**Regression:**

I have used the lm function to perform the regression model and as shown in the snip below, I have used “GarageArea, YearBuilt, FullBath, OverallQual, LotArea, BsmtFinSF1, YearRemodAdd, Fireplaces” as best predictors.

****

**Summary of the above output of regression**:

Model Fit:

* R-squared: 0.7699 indicates a strong fit, explaining 77% of the variance in house sale prices.
* Adjusted R-squared: 0.7679 slightly lower than R-squared, suggesting the fit is good and not inflated by the number of features.
* F-statistic: 372.7 with p-value < 2.2e-16 indicates the model is statistically significant and not due to chance.

Coefficients:

* Intercept: -830,400 represents the predicted sale price when all features are set to 0 (unlikely scenario).
* GarageArea: 77.82 increase in sale price per additional square foot of garage space.
* YearBuilt: -743.4 decrease in sale price per year built (older houses sell for less).
* FullBath: 20,370 increase in sale price per additional full bath.
* OverallQual: 27,380 increase in sale price per unit increase in overall quality score.
* LotArea: 7.15 increase in sale price per additional square foot of lot area.
* BsmtFinSF1: 31.65 increase in sale price per additional square foot of finished basement space.
* YearRemodAdd: 44,960 increase in sale price per year of remodel/addition.
* Fireplaces: 14,410 increase in sale price per additional fireplace.

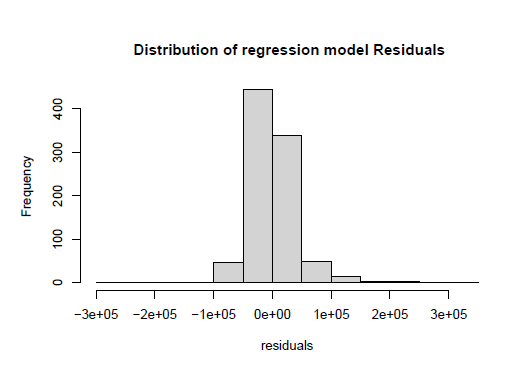
Significance:

* All coefficients are statistically significant (p-value < 0.05) except YearBuilt, suggesting it might not have a strong impact on sale price or could be correlated with other features.

Overall:

This model seems to be a good fit for predicting house sale prices based on the provided features. Most features have statistically significant impacts on the price, with garage area, full baths, overall quality, and lot area having the strongest positive influences.

**The histogram on residual of regression:**

****

**Summary of the above histogram:**

The primary plot depicts the distribution of residuals, which are the disparities between real house sale prices and the values predicted by the model. The x-axis displays the residual values, and the y-axis displays the density of residuals at each value. The distribution's general form appears to be nearly symmetrical, which is a favorable indicator. There are also no significant outliers, indicating that the model is not producing big errors in individual predictions.

After implying the model on the testing dataset “Predict”, here is the output of the 90 houses in the test dataset.

A group of people in a row

Description automatically generated with medium confidence

Summarizing the output of the predicted dataset, we can say that:

* The predicted prices range from approximately $38,967 to $331,960, highlighting a large variation in housing costs within the tested set.

**Decision Tree:**

Summary of the output of the Decision tree model:

**Structure of a Tree:**

The output depicts the decision tree's hierarchical structure, with each node indicating a decision and branching dependent on the feature value selected.

The initial split is based on "OverallQual", then "YearBuilt", "GarageArea", and so on, iteratively separating the data into smaller groups depending on their characteristics.

Each terminal node (leaf) provides the expected mean sale price as well as the Mean Squared Error (MSE) for the dwellings within that node.

**Importance Variable:**

The "Variable importance" section illustrates how important each attribute is in separating the data and creating predictions.

"OverallQual" is the most important, followed by "GarageArea", "YearBuilt", and so on, showing that they have a significant influence on the expected sale price.

**Interpretation:**

This model forecasts house selling prices using a set of decision rules. Houses with higher "OverallQual" scores, for example, are expected to sell for more money.

The offered output allows you to study the model's specific splits and rules, providing insights into its decision-making process.

A diagram of a graph

Description automatically generated with medium confidence

This snip depicts the plot of the decision-tree model.

After implying the decision tree model on the predicted dataset, we receive an output of 900 records of house prices which includes multiple repeating values such as 351838.2 and 474931.1.  
The predicted values appear to be quite diverse, indicating that the decision tree model can capture a wide range of patterns in the data.

**Classification Model:**

I converted the “OverallQual” variable from factor to numeric which later is classed into 1 and 0 (Binary) based on the rating 7 and above as 1 and rest as 0.

CA screenshot of a computer

Description automatically generated

Here is the output of the classification model.

Coefficients:

For a one-unit rise in the relevant attribute, each coefficient indicates the projected change in the log chances of "OverallQual" being 1.

favorable coefficients indicate a favorable relationship with "OverallQual" of 1, whereas negative coefficients indicate a negative relationship.

The significance levels (Pr(>|z|)) represent the likelihood of detecting such a coefficient by chance. Lower values (0.05) indicate statistically significant relationships.

Coefficient Interpretation:

GarageArea: Each unit increase in GarageArea corresponds to a 5.697e-3 rise in the log probabilities of "OverallQual" being 1. This implies a favorable, albeit little, relationship between larger garages and higher overall quality.

YearBuilt: Older houses (those with a greater YearBuilt) have a somewhat better likelihood of being rated as "OverallQual" 1. This could imply that older houses have qualities or attributes that are valuable in this setting.

FullBath: Adding another full bath raises the log probabilities of "OverallQual" becoming 1. This emphasizes the significance of bathrooms in terms of overall quality perception.

LotArea: The LotArea coefficient is negative but not statistically significant. This shows that the relationship between lot size and "OverallQual" is insignificant or poor.

BsmtFinSF1: Having a greater finished basement area reduces the log probabilities of "OverallQual" being 1. This could be due to worries about dampness, basement condition, or personal preferences.

YearRemodAdd: More recent renovations (higher YearRemodAdd) enhance the log probabilities of "OverallQual" being 1 greatly. This emphasizes the importance of new updates and renovations.

Fireplaces: Having a fireplace raises the log odds of "OverallQual" being 1 substantially. This indicates that fireplaces are regarded as a desirable feature.

The model outperforms just guessing (null deviance), implying that it captures meaningful correlations between the features and "OverallQual."

The residual deviation is still relatively high, indicating that the model's fit might be improved.

The AIC is reasonable, indicating that the model adequately balances complexity and fit.

After implementing the model on the predict dataset, we get the following accuracy:

A black text on a white background

Description automatically generated

We can summarize that the accuracy of this model is 82.2% approx.

1. **Insights and Conclusions:**

Analyzing all the model outputs and their implementation on the testing data set, we can conclude that:

Regression has an overall accuracy of 76% which is derived from adjusted R-squared value from its output.

Whereas decision tree model has 20 nodes which represent a complex structure, so I can say that it needs to be improved.

The classification model with an overall accuracy of 82.2% has higher accuracy than the remaining models, therefore classification can be concluded as the best model among all.