Analytical Approach to Marketing Decisions
Project Report

Ames Housing Data - Price Prediction Model

Instructor: Sir Hassaan Khalid

Group Members

Aafia Khan (17263)

Rija Alam (16901)

Sarah Syed Naqvi (17179)

Zarmeen Lakhani (17398)

Zehra Mubashir (16834)

Table of Contents

1) Introduction **2**

2) Data Description **2**

3) Problem Statement and Motivation **2**

4) Detailed Methodology **3**

Data Segregation 3

Outlier Detection 3

Without Outliers 3

With Outliers 4

Feature Engineering 4

Working on Data 5

Stepwise Regression (y-variable: sales price data) 5

Stepwise Regression (y-variable: ln of sales price data) 5

Decision Tree Analysis 6

5) Model Results **6**

Stepwise Regression 6

Decision Tree Analysis 7

6) MAPE Calculations **8**

Stepwise Regression 8

Decision Tree 8

7) Limitations and Future Scope of Work **9**

8) Conclusion **10**

Appendix **11**

R Codes 11

For Decision Tree: 11

For Stepwise Regression: 11

Contribution Sheet 12

1) Introduction

The business problem that our group has focused on is about predicting the sales prices of Ames Housing. The coefficients generated from the data will help property owners understand what factors make a positive differential impact on the value of property. After applying advanced analytics techniques on the data downloaded from Kaggle, the predicted sales price will be used to calculate accuracy of the regression model used.

2) Data Description

Our group has chosen a dataset which contains details of houses sold in Ames, IA, during the period 2006-2010. The data set has 38 columns of data pertaining to various attributes related to the houses sold in Ames including plot size, number of rooms, floor wise area, month and year sold, and details of all other facilities available.

Since the chosen data set was part of competition on Kaggle, therefore, the actual sales prices for testing data file has not been provided. Therefore, only the file of training data will be used, and the data provided, consisting of 1953 columns will be segregated in the testing and training data set by ratio of 1:1.

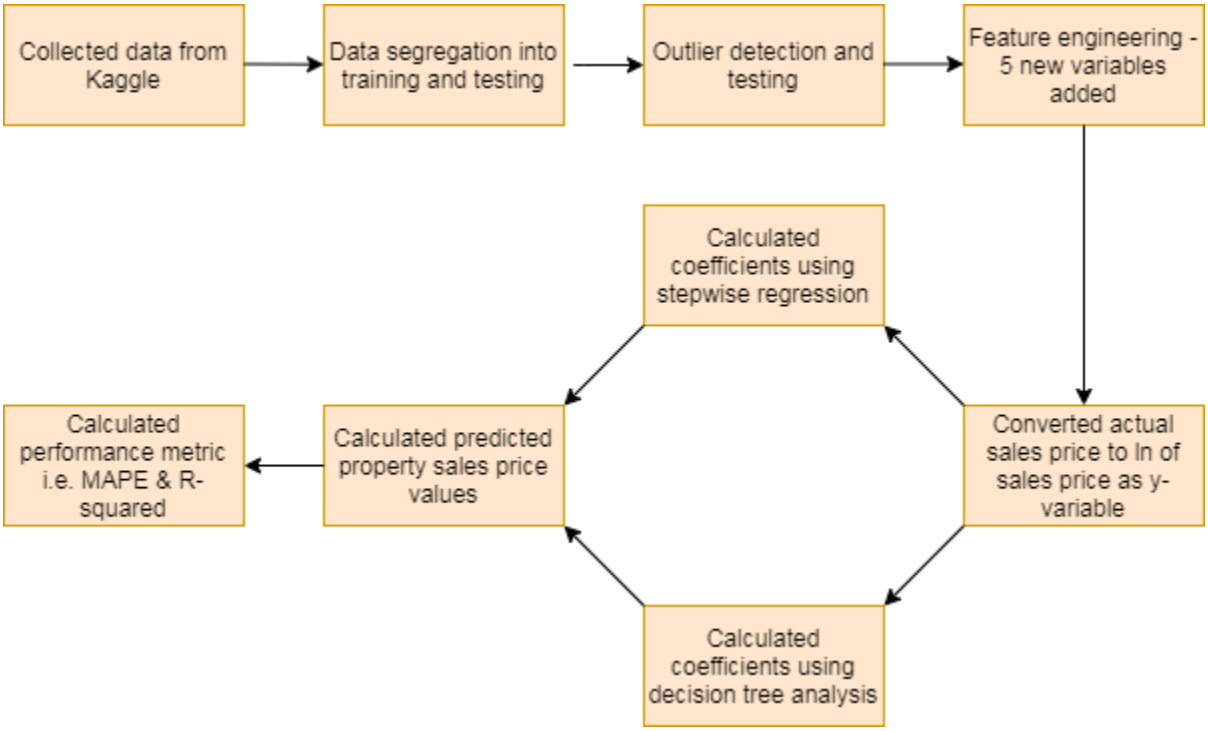
The raw data set had 41 variables and after feature reduction through the step function in R-studio, 23 variables have been found to be beneficial after running a stepwise regression model for the raw data specifically.

3) Problem Statement and Motivation

There is a global problem of real estate agents quoting inflated house prices to potential buyers and clients when they are approached by prospective clients, many frauds and illegal practices are common in the real estate sector. Buyers are conned into over paying as there is asymmetry of knowledge.

To address the asymmetry of knowledge in evaluating prices of the Housing industry, we have modeled our data to return closest prices. We have taken in consideration that prices in the Housing Industry are not set on a stand alone basis rather are a cumulation of different economic factors and various attributes of the house. As real estate houses are long term investments they have very elastic demand so it is crucial for both the real estate sellers and purchasers to map out realistic prices of real estate after taking certain variables into consideration.

4) Detailed Methodology



Data Segregation

To start with, the raw data file containing 1953 rows was divided into two parts to segregate training and testing data following a ratio of 1:1 leading to 976 data rows for training and 977 data rows for testing.

Outlier Detection

In order to find the impact of the presence of outliers in the raw dataset, stepwise regression was run on both types of data – with outliers and without outliers, followed by the calculation of their performance metrics.

Without Outliers

In both training and testing data, the top 5% and bottom 5% of actual sales price values (y-variable) were identified by applying a function of rank and percentile and then sales price for identified rows (outliers) were being replaced with median of sales price data. Stepwise regression was then being run through R and resultant coefficients led to mean absolute percentage error (MAPE) of 15.78% and R-square of 54.27% for training data and mean absolute percentage error (MAPE) of 17.9% and R-square of 57.47% of testing data.

With Outliers

As the objective was to discern the effect of outliers, the stepwise regression was also run through on raw data with outliers in data still intact. Resultant coefficients gave mean absolute percentage error (MAPE) of 12.21% and R-square of 85.94% for training data and mean absolute percentage error (MAPE) of 13.70% and R-square of 71.3% for testing data.

With Raw Data

Train - Raw Data	
Mean APE	12.21%
Median APE	8.77%
Total	6.4322E+12
Residual	9.0436E+11
Regression	5.52784E+12
R^2	85.94%

Test - Raw Data	
Mean APE	13.70%
Median APE	9.17%
Total	5.90609E+12
Residual	1.69487E+12
Regression	4.21123E+12
R^2	71.30%

with Outliers Removed

Train - Data Outliers Removed	
Mean APE	15.79%
Median.APE	12.53%
Residual	1.22399E+12
Regression	1.45271E+12
Total	2.67669E+12
R^2	54.27%

Test - Data Outliers Removed	
MAPE	17.90%
Median.APE	12.58%
Residual	2.50871E+12
Regression	3.38948E+12
Total	5.89819E+12
R^2	57.47%

The resultant MAPE and R-square led us to leave the outliers in the data as it will lead to better stepwise coefficients, and hence, a better model.

Feature Engineering

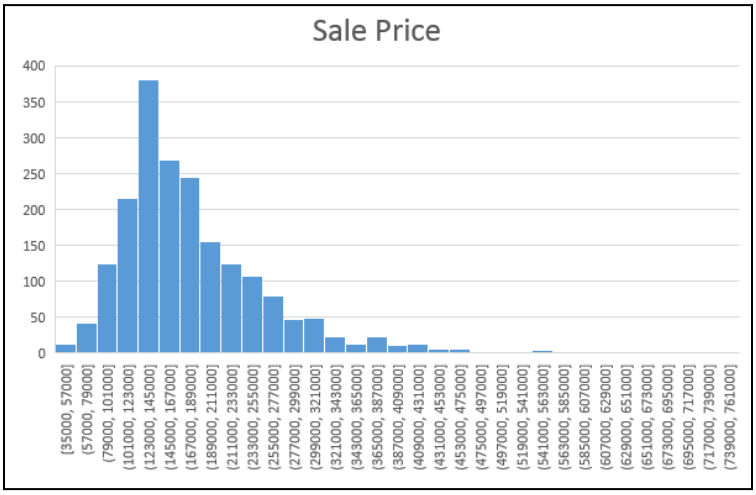
In an effort to find better and relevant variables' coefficients, 5 new variables (unemployment rate, mortgage rate, Ames US crime rate in proportion to the United States crime rate, Ames US property crime rate and inflation rate for year from 2006 to 2010) were made to be part of raw data file whose data was acquired from secondary sources i.e. state's official website. Unemployment rate, inflation rate and mortgage rate are taken as an average for each month

while Ames US crime rate and Ames US property crime rate are taken as an average for each year.

Working on Data

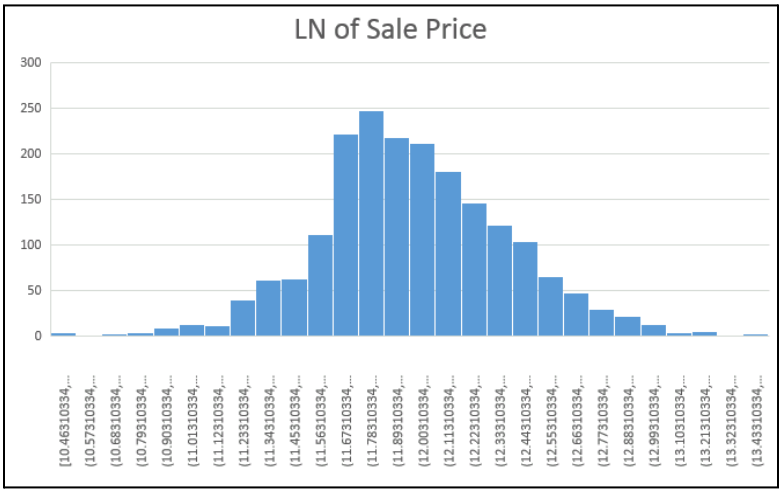
Stepwise Regression (y-variable: sales price data)

The feature engineered data (with outliers) of training data was made to run through RStudio for stepwise regression with sales price as being the y-variable and resultant coefficients were then used to calculate predicted sales price values for both training and testing data leading to MAPE of 12.23% and R-square of 85.98% for training data and MAPE of 13.76% and R-square of 71.12% for testing data.



Stepwise Regression (y-variable: ln of sales price data)

As a matter of fact, the histogram of ln (natural log) of sales price data resembled more to normal distribution leading our group to also run stepwise regression taking ln of sales price as y-variable to help us decide whether sales price in isolation is a better y-variable or ln of sales price.



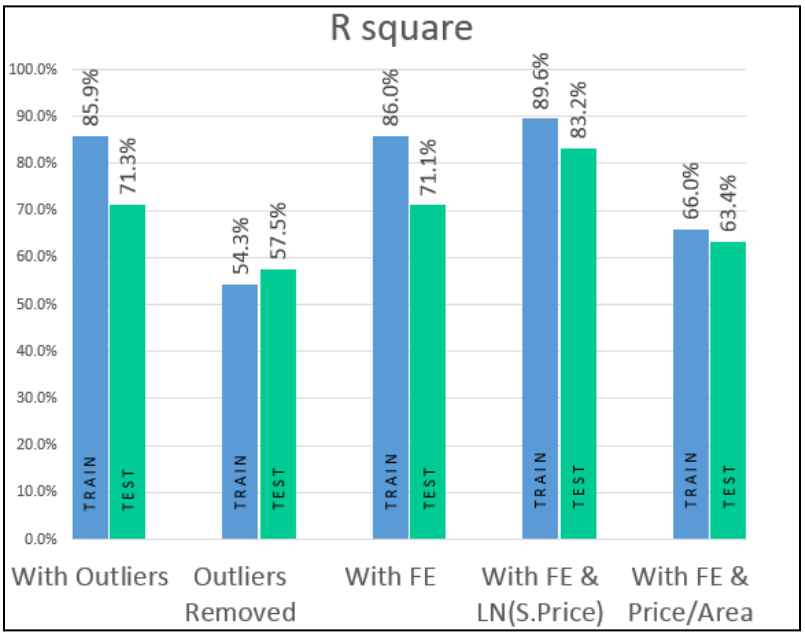
The resultant coefficients were then used to calculate predicted ln of sales price and using function of exponent on each of them led to predicted sales price values. For training data, the MAPE came out to be 9.13% and R-square of 89.57%. For testing data, the MAPE turned out to be 11.41% and R-square was 83.24%.

Decision Tree Analysis

A random forest tree regression was done on the raw Real Estate Data of Ames by importing libraries such as dplyr, rpart and rpart.plot for generating decision trees and calculating predicted values. As the data with outlier removed captured less R^2 so the logarithm function of sales prices were treated as y variable with 28 independent variables.

5) Model Results

Stepwise Regression

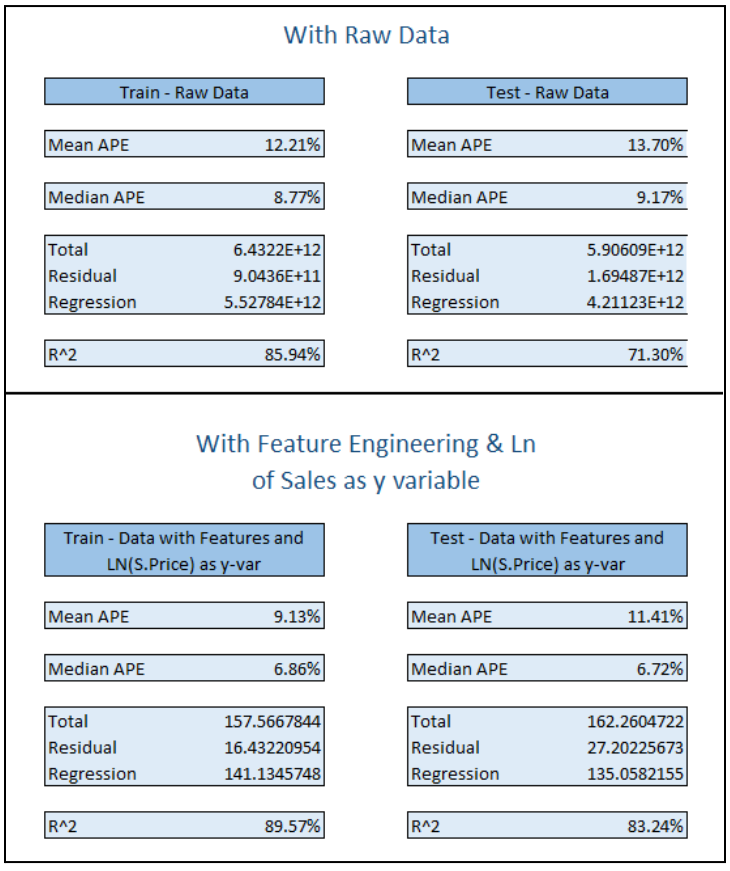


The first point of concern was to see whether replacement of outliers will give better coefficients and better model or the outliers should be kept as they are in the data. The calculation of MAPE and R-square for both training and testing data confirmed that it was more accurate to keep the outliers as they are instead of replacing them as the MAPE was lower by 4.2% and R-square higher by 13.86% for testing data compared when outliers were not removed. It was also inferred from the results and the nature of the case that if the outliers, sales prices too low and high, had been replaced with a median value of all the prices, then the model would yield inaccurate and inconclusive results. For instance, a property covering a larger lot area, pool, and other fixtures

must have a higher sales price. Had its value been replaced by a median value, or had the entire row been replaced, then the model would have been confused, unable to identify property with exceptionally high or low value.

Furthermore, since the condition of linear relationship between y-variables and x-variables had to be fulfilled for regression, better relationship was being portrayed when y-variable was taken as the natural logarithm value of sales price value compared with directly taking sales as y-variable and resulted in decline in MAPE for testing data by 2.35% and rise of R-square by 12.1%.

Another attempt was made as an effort to construct a model with greater accuracy by transforming the y-variable, taking the values of sales price per lot area. This new y-variable although generated a more normally distributed histogram, the variables were run through R studio, and the subsequent calculations from coefficients achieved showed that MAPE turned out to be higher by 22.15% (training data) and by 18.77% (testing data) and R-square lower by 23.55% (training data) and by 19.87% (testing data). In summary, the idea of taking ln of sales price per lot area did not work out and therefore, did not become part of the final model.



With Raw Data

Train - Raw Data	
Mean APE	12.21%
Median APE	8.77%
Total	6.4322E+12
Residual	9.0436E+11
Regression	5.52784E+12
R^2	85.94%

Test - Raw Data	
Mean APE	13.70%
Median APE	9.17%
Total	5.90609E+12
Residual	1.69487E+12
Regression	4.21123E+12
R^2	71.30%

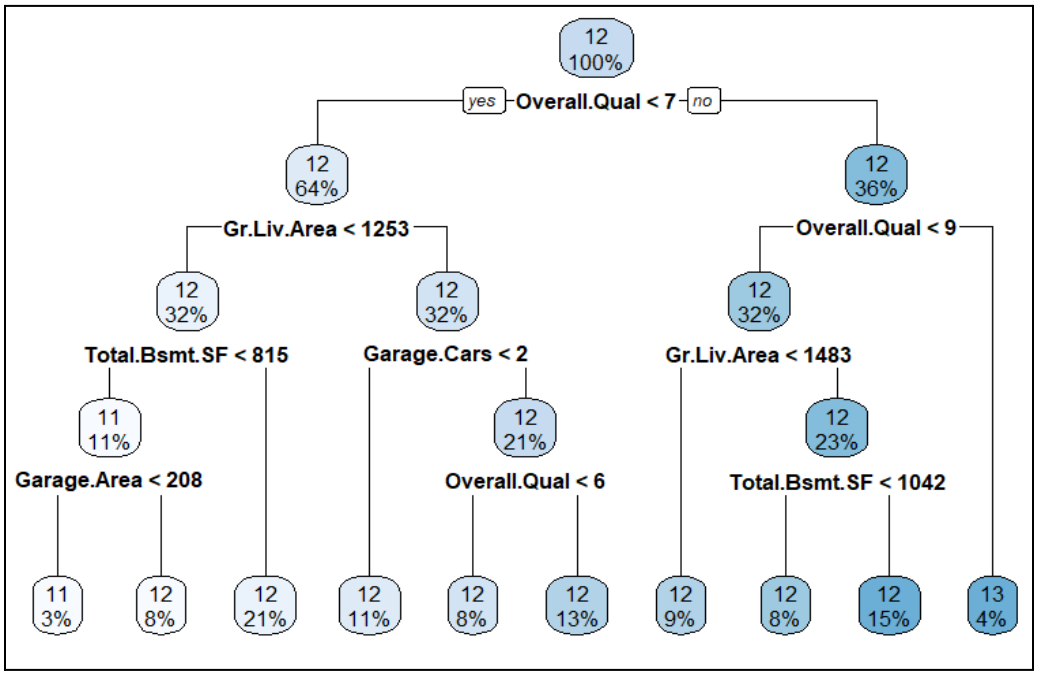
With Feature Engineering & Ln of Sales as y variable

Train - Data with Features and LN(S.Price) as y-var	
Mean APE	9.13%
Median APE	6.86%
Total	157.5667844
Residual	16.43220954
Regression	141.1345748
R^2	89.57%

Test - Data with Features and LN(S.Price) as y-var	
Mean APE	11.41%
Median APE	6.72%
Total	162.2604722
Residual	27.20225673
Regression	135.0582155
R^2	83.24%

In conclusion, it was established that the model that generates the most accurate predictions of property sales price is the model which takes the natural logarithm of sales price as the transformed dependent variable and incorporates the feature engineered variables rendered significant by the R stepwise regression code. This is the model the results of which are compared with the Decision Tree results in the following section.

Decision Tree Analysis



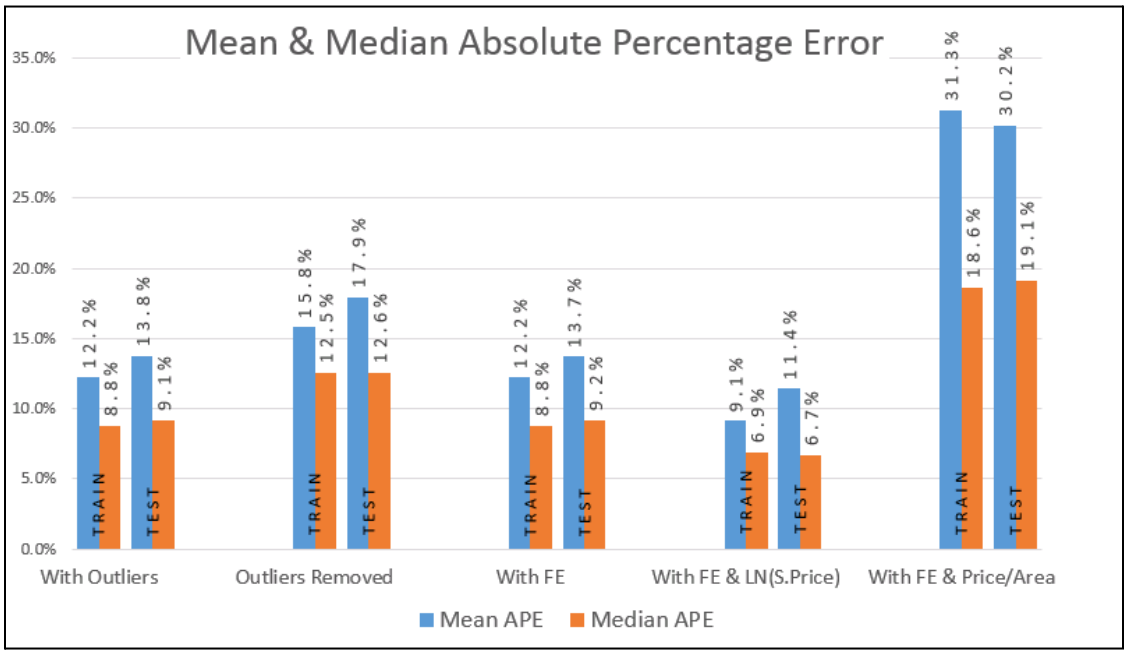
The Random Forest regression model turns the linear data into two nodes by setting a threshold level, the value plugged in the data were the log of the sales as an independent variable with 28 other variables including feature engineering independent variables. The threshold value of Overall Quality being greater than 7 explains 64% of the data points. The tree further breaks into nodes which explain that if a certain threshold level of a variable is satisfied then the data will be explaining relevant percentages of the data explained by the model and we are able to calculate significance of variables through the model.

As compared to the performance metrics yielded by the final linear regression model i.e. ln(sales price) and feature engineering, the decision tree model generated results indicative of low variance explained away and greater deviation from actual sales price. Hence based on the MAPE and R-square for both training and testing data achieved from stepwise regression and random forest regression, it can safely be concluded that a better model was built through stepwise regression. The following section discusses the results of the models' performance metrics.

6) MAPE Calculations

Stepwise Regression

The chart below is a graphical representation of the variations witnessed in the Mean and Median Absolute Percentage Error as different transformations or variables were added to the data set. The graph also depicts the significant improvement in the model once the ln(sales price) transformation was applied and new variables introduced to the dataset, resulting in a more robust and accurate model.



Decision Tree

After capturing coefficients from csv file after running random forest regression when y-variable was taken as logarithmic function of sales values, error was calculated for each entry from predicted which then led to calculation of mean absolute percentage error. For training data, MAPE turned out to be 15.5% and R-square 75.26% while for testing data MAPE was 16.9% and R-square was 71.6%.

7) Limitations and Future Scope of Work

Our study consisted of stepwise linear regression and random forest analysis, of which the former gave more accurate results. However, it must be taken into account that several other algorithms – including support vector machine (SVM) and gradient boosting machine (GBM) –

could be utilized in further research, on the basis of whether processing time, accuracy, or some other factor is given greater priority

Five additional variables were included in our study; however, different variables were identified as significant in different settings (with or without outliers, sales price or natural log of sales price, or random forest analysis). This indicates that most of these variables may have an impact on property pricing and may have been significant under a different approach. Another limitation identified was that of overfitting in decision trees using random forest analysis. This occurs when the tree tries to accommodate the training data entirely, making it less suitable for the testing (and other) dataset(s).

Further research on property pricing could result in accurate models that can be used for similar upcoming housing projects and policy making at state level. This could take into account the difference in pricing in different areas as well, which may include geographic and climatic risks with regards to natural disaster management.

8) Conclusion

We conducted an exploratory research on Ames Housing's 5 year dataset to predict property prices using stepwise regression and random forest analysis. As the MAPE was higher and R-square was lower after replacing the outliers with the median, the outliers were kept in the dataset, as opposed to usual data cleaning standards. Similar results were observed when the model took sales price, sales price per lot area, or the natural logarithm function of sales price per lot area as the y-variable, thereby, natural log of sales price was used for further calculations, comparisons and modeling.

Our key determinants of a good model in this study were the Coefficient of Determination (R-square), the Mean Absolute Percentage Error (MAPE), and the Median Absolute Percentage Error. While comparing the results of these two machine learning algorithms, stepwise linear regression proved to be more accurate with a R-square value of 71.30% and MAPE of 13.70% for the training dataset with outliers, and 85.94% and 12.21% respectively for the testing data.

Along with the variables given in the dataset, the five additional variables proved to be useful in predicting the sales price in the Ames Housing Dataset. However, as the scope of this study only covered stepwise linear regression and random forest analysis, it is probable that other algorithms may have included other variables and given a more accurate pricing model. Further developments in this field could prove to be extremely helpful in planning and strategizing in the construction sector and for rural/urban development.

Appendix

R Codes

For Decision Tree:

```r
data <- read.csv("Ames_Housing_Raw_WithLN.csv", header = T)
colnames(data) #checking column names
str(data) #checking data structure
data <- data[, -1]
library(dplyr)
colnames(data)
train <- data[1:976,]
test <- data[977:1954,]
library(rpart)
?rpart
set.seed(123)
mymodel <- rpart(formula = LN.SalePrice ~ .,
data = train)
mymodel
mymodel$variable.importance
summary(mymodel)
library(rpart.plot)
rpart.plot(mymodel)
mypredictions_train <- predict(mymodel, train)
mypredictions_test <- predict(mymodel, test)
allpredictions <- c(mypredictions_train,
mypredictions_test)
write.csv(allpredictions, file = "LN+FE-DT.csv")
```

For Stepwise Regression:

```r

data <- read.csv("Ames_Housing_Raw.csv", header = TRUE)

data <- data[-1]

str(data)

train <- data[1:976, ]
test <- data[977:1953, ]

model <- lm(SalePrice~., train)

model_new <- step(model)

coeff <- coefficients(model_new)

allpredictions <- fitted(model_new)

write.csv(coeff, file = "coeff.csv")
```

Contribution Sheet

Group member	Contribution
Aafia Khan 17263 AAMD 50715	
Rija Alam 16901 AAMD 50716	
Sarah Syed Naqvi 17179 AAMD 50716	
Zarmeen Lakhani 17398 AAMD 50715	
Zehra Mubashir 16834 AAMD 50715	