**Project 2: Executive Report**

ALY6020 – Predictive Analytics, Northeastern University

Professor Justin Grosz

7/25/21

**Introduction**

The goal of this project was to assist a car manufacturer identify what features are vital to having an energy efficient car. To achieve this goal, we created a linear regression model to help identify what features cars with high miles per gallon typically possess. The first step in our process was to clean our dataset and ensure that our data was the highest quality possible. After cleaning our data, we then ran a baseline linear regression model which included all variables and included all outliers. These results were used as a baseline to determine if our optimization techniques were effective. After running the baseline model, we ran 3 additional models that stepwise selection and data with outliers removed. These optimized models were then compared to each other to identify which model was the most accurate. The resulting model was then used to inform decisions around which features of a car were the most important to maximizing MPG.

**Data Cleanup**

The first step in our modeling process was the investigation into our dataset and verify that our data was sound and of good quality. The first step in our process was to calculate summary statistics for our dataset. Through this process we were able to identify that our dataset contained six numerical values and two categorical values. We were also able to identify that we had some entries that contained question marks in the horsepower column. To address this, we went field by field and attempted to determine if any information in the dataset would be useful in determining what best to replace the irregular data with. Since there were no clues in the dataset as to what this value should represent, we replaced these missing values with the mean of the column. Through this process, we identified that the horsepower column was incorrectly classified as a character data type and that the factors model year and US.Made were also incorrectly classified as numerical values. These columns were correctly classified and summary statistics were recalculated.

In addition to identifying issues with data types, the calculation of summary statistics also provided us an understanding of the distribution of our numerical data. Upon visual review of the summary statistics, we found that their appeared to be no abnormal or impossible values in our dataset. To verify this statistically, we plotted each of the numerical values to observe the distribution of each variable. In models one and three we utilized the original distribution of data. For models two and four we utilized data that was cleaned using the following process. First, we calculated the value representing two standard deviations above and below the mean for each numerical column. Using this value as a cut off for each column, we replaced all values beyond two standard deviations from the mean in either direction with the value representing two standard deviations from the mean. Once complete, we plotted our variables individually to see how the distribution had shifted.

Chart, bar chart

Description automatically generatedOnce this was complete the last step was the creation of a correlation plot for our data to understand how all the variables relate to each other. The resulting plot can be found in figure 1. Here we can see that there are several variables that are strongly correlated with each other. We can also see that MPG has the strongest negative correlation with weight, followed by displacement, cylinders and horsepower. This implies that we may need to minimize these variables to create the most fuel-efficient car possible. Full code all steps outlined above can be found in Appendix A.

Figure : Correlation plot of all numerical variables

**Baseline Model Results**

We first started by creating a testing and a training dataset to help us validate our model. For the entirety of this project, we utilized a training dataset that was 80% of our total data. The rest was used as testing data. Our first baseline model was run using the data that still included outliers. This model yielded strong results with a RMSE of 2.92 MPG, an AIC value of 1645.47 and an R squared value of .8543. All told, this model would be useful for use in the real world as the average error was 2.92 MPG. This margin of error is good enough to give us a direction estimate of whether the car the company is looking at manufacturing will be fuel efficient. In this model, we found that the most statistically significant actionable variables were *Weight, US.Made, Displacement,* and *Horsepower*. In addition to these variables there were a couple of *model year* variables that were significant. However, since these are variables that can’t be influenced, they were not considered actionable and thus were not considered. Full model results can be found in Appendix B.

In this initial baseline model, we found that to maximize MPG the car manufacturer should focus on making sure that the car is not made in the US. According to our model, cars made in the US have 2.59 MPG less on average. It should be noted that this could be due to bias in the data. If most fuel-efficient car manufacturers are located outside of the US, then this could seriously skew the data. Logically, it does not make sense that manufacturing location alone could influence how fuel efficient a car is. We recommend that additional research be done to understand why manufacturing location plays such a big part in this model. In addition to this research, the company should focus on designing a car that has minimal cylinders and weight while aiming for an engine with a high amount of displacement. Each pound of weight equated to a loss of .006 MPG in our model. If we were able to reduce weight by just 100 pounds, we may see an increase of .6 MPG.

**Optimized Model Results**

As seen in Table 1 below, four total models were utilized on our dataset. The first model in our table represents the baseline results obtained using the unmodified data. Model two represents a linear regression model run with outliers replaced in our dataset. Models three and four were stepwise regression models which attempted to select variables that further optimized our modeling.

Table 1: Full Model Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AIC** | **R-Squared** | **RMSE** | **Data Used** | **Technique Used** |
| *1645.47* | *.8446* | *2.94* | *Unmodified* | *Linear Regression* |
| *1649.89* | *.8436* | *2.83* | *Outliers Replaced* | *Linear Regression* |
| *1646.63* | *.8443* | *2.82* | *Outliers Replaced* | *Stepwise Selection* |
| *1641.79* | *.8455* | *2.94* | *Unmodified* | *Stepwise Selection* |

From these results we can see that the model that minimized RMSE was model three. While the AIC is slightly higher than most models, the gain in accuracy was significant enough to opt for this model. The AIC informs us that our model may be over predicting. All told, this model is suitable for our business case and should be utilized to help solve our business problem.

Upon further investigation into this model, we found that the most significant and actionable variables were again *US.Made, Weight, Horsepower,* and *Displacement*. Our advice from our original model remains unchanged regarding these variables. However, it should be noted that in this version of our model, cylinders are no longer included. Feature selection determined that this variable was not a part of the optimal model. As a result, we recommend that the car manufacturer focus on increasing displacement and limiting horsepower and weight. It should be noted that displacement and weight are positively correlated, which could mean that increasing displacement could cause an increase in weight. This is also the case with horsepower and displacement. The car manufacturer would need to evaluate each change wholistically to understand the overall impact to the model since it appears these actionable and significant variables are highly corelated.

**Conclusion**

The goal of our analysis was to assist a car manufacturer identify what features are vital to having an energy efficient car. To achieve this goal, we utilized linear regression and stepwise regression to create the most accurate predictive model that would allow the company to identify what features to focus on. Through this process we can conclude that stepwise regression using our standardized dataset was the most accurate model. Additionally, we recommend that the company focus on designing a car that has high displacement while limiting the weight and horsepower of the car. It should be noted that each of these variables have a strong positive correlation with each other. Every design change should be evaluated wholistically to see the overall impact to MPG. Finally, our model suggested at a statistically significant level that cars that were manufactured in locations outside the US had nearly 2.94 MPG more on average. We recommend further analysis be done to determine if this is due to bias in the data. Logically it seems impossible that the physical location of manufacturing has that significant of an impact on fuel efficiency.

**Appendix A: Model Code**

library(tidyverse)

library(corrplot)

library(ggplot2)

library(stats)

#load data

data <- read.csv(file.choose(), sep=",", header = TRUE )

# Summary Stats

summary(data)

# Set Correct Data Types

cols<-c(7,8)

data[cols] <- lapply(data[cols], factor)

data$Horsepower <- as.numeric(data$Horsepower)

# Check for N/A's

# Find NAs

new\_DF <- data[rowSums(is.na(data)) > 0,]

# Replace ?'S in Horsepower

# 6 HP values with ?'s in the column

# 6 out of 398 = 1.5% of entries- replace with mean

data$Horsepower[is.na(data$Horsepower) == TRUE] <- mean(data$Horsepower, na.rm = TRUE)

# Correlation

cordata <- data[,c(1:6)] #correlation plot (numerical only)

res <- cor(cordata)

round(res, 2)

corrplot(res, method="color")

# Create Test and Train Datasets

set.seed(123)

allrows <- 1:nrow(data)

trainrows <- sample(allrows, replace = F, size = 0.8\*length(allrows))

train\_df <- data[trainrows,c(1:8)]

test\_df <- data[-trainrows, c(1:8)]

# Modeling Initial

model = lm(MPG~., data=train\_df)

summary(model)

alias(model)

model$coefficients

plot(residuals(model))

## Accuracy Measures

# Adjusted R^2 = 0.8446

aic <- AIC(model) # aic = 1645

mse <- mean(model$residuals^2) # 9.12 MSE

pred1 <- predict(model, newdata = test\_df)

rmse <- sqrt(sum((pred1 - test\_df$MPG)^2)/length(test\_df$MPG))

c(RMSE = rmse, R2=summary(model)$r.squared)

par(mfrow=c(1,1))

plot(test\_df$MPG, pred1)

##################################################

#### Model Round 2 ###

##################################################

data2 <- tibble::rowid\_to\_column(data, "ID")

# # Outlier Detection and Replacement

col\_num <- c(2:7)

for(i in col\_num ){

plot(data2$ID,data2[,i]) # by entry before cleanup

mean<- mean(data2[,i])

std <- sd(data2[,i])

two\_sd <- (2\*std) + mean

two\_sd\_below <- (-2\*std) + mean

data2[,i][data2[,i] > two\_sd] <- two\_sd

data2[,i][data2[,i] < two\_sd\_below] <- two\_sd\_below

}

#summary Stats

summary(data2)

# Create Test and Train Datasets

set.seed(125)

allrows2 <- 1:nrow(data)

trainrows2 <- sample(allrows, replace = F, size = 0.8\*length(allrows2))

train\_df2 <- data[trainrows2,c(1:8)]

train\_label2 <- data[trainrows2, 8]

table(train\_label2)

test\_df2 <- data[-trainrows2, c(1:8)]

test\_label2 <- data[-trainrows2, 8]

table(test\_label2)

# Modeling with Cleanup

model2 = lm(MPG~., data=train\_df2)

summary(model2)

alias(model2)

model2$coefficients

plot(residuals(model2))

## Accuracy Measures

# Adjusted R^2 = 0.8446

aic2 <- AIC(model2) # aic = 1645

mse2 <- mean(model2$residuals^2) # 9.12 MSE

pred2 <- predict(model2, newdata = test\_df2)

rmse2 <- sqrt(sum((pred2 - test\_df2$MPG)^2)/length(test\_df2$MPG))

c(RMSE = rmse2, R2=summary(model2)$r.squared)

par(mfrow=c(1,1))

plot(test\_df$MPG, pred2)

##################################################

#### Model Round 3 ###

##################################################

# Stepwise Selection with Standardized Data

library(caret)

library(leaps)

library(MASS)

#Select variables using stepwise regression

step.model <- stepAIC(model2, direction = "both",

trace = FALSE)

summary(step.model)

aic3 <- AIC(step.model)

mse3 <- mean(step.model$residuals^2)

pred3 <- predict(step.model, newdata = test\_df2)

rmse3 <- sqrt(sum((pred3 - test\_df2$MPG)^2)/length(test\_df2$MPG))

c(RMSE = rmse3, R2=summary(step.model)$r.squared)

par(mfrow=c(1,1))

plot(test\_df$MPG, pred3)

##################################################

#### Model Round 4 ###

##################################################

# Stepwise Selection with unstandardized Data

#Select variables using stepwise regression

step.model2 <- stepAIC(model, direction = "both",

trace = FALSE)

summary(step.model2)

aic4 <- AIC(step.model2)

mse4 <- mean(step.model2$residuals^2)

pred4 <- predict(step.model2, newdata = test\_df)

rmse4 <- sqrt(sum((pred4 - test\_df$MPG)^2)/length(test\_df$MPG))

c(RMSE = rmse4, R2=summary(step.model2)$r.squared)

par(mfrow=c(1,1))

plot(test\_df$MPG, pred4)

**Appendix B: Full Model Results & Accuracy Statistics**

**Table

Description automatically generated**

**A picture containing text, newspaper, receipt

Description automatically generated**