Final Project

Parker Leipzig, Cole Vandenheuvel, Steven Lawver

2023-12-12

Data Preprocessing and Model Building:

Data Preprocessing: The initial dataset was preprocessed to make it suitable for modeling. This involved removing columns with only a single unique value, as they do not contribute to variability in the dataset. The categorical variables were transformed into dummy variables, ensuring that the models could interpret and utilize this information effectively.

Model Training: The dataset was split into training and testing sets, with 80% of the data used for training and the remaining 20% for testing. This split ensures a good balance between learning from the data and validating the model's performance.

```
# Load necessary libraries
library(caret)
library(randomForest)
library(glmnet)

# Read the data
data <- read.csv("UsedCars (1) (5).csv")

# Remove columns with a single unique value
data <- data[, sapply(data, function(x) length(unique(x)) > 1)]

# Convert factor data to dummy variables
dummies <- dummyVars(" ~ .", data=data)
data_transformed <- data.frame(predict(dummies, newdata = data))

# Split data into training and testing sets
set.seed(42)
index <- createDataPartition(data_transformed$Price, p = 0.8, list = FALSE)
train_data <- data_transformed[index,]
test_data <- data_transformed[-index,]

# LASSO Model
set.seed(42)
lasso_model <- glmnet(as.matrix(train_data[-ncol(train_data)]),
train_data$Price, alpha = 1)
prediction_lasso <- predict(lasso_model, s = 0.1, newx =
as.matrix(test_data[-ncol(test_data)]))
mse_lasso <- mean((prediction_lasso - test_data$Price)^2)

# Random Forest Model
```

```r
set.seed(42)
rf_model <- randomForest(Price ~ ., data=train_data, ntree=100)
prediction_rf <- predict(rf_model, test_data)
mse_rf <- mean((prediction_rf - test_data$Price)^2)

# Print MSE values
print(paste("MSE for LASSO:", mse_lasso))
```

```
## [1] "MSE for LASSO: 123267.681685126"
```

```r
print(paste("MSE for Random Forest:", mse_rf))
```

```
## [1] "MSE for Random Forest: 31892956.019491"
```

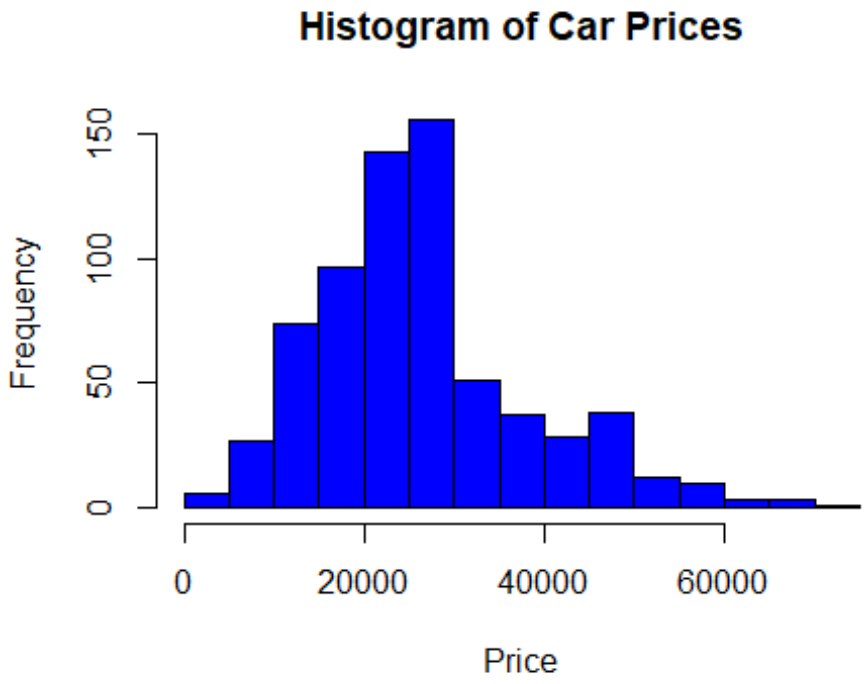
Model Performance and Evaluation:

LASSO Model: The LASSO (Least Absolute Shrinkage and Selection Operator) regression was employed, which is particularly useful when dealing with datasets with a high number of features. It helps in feature selection and regularization, reducing the overfitting risk. The mean squared error (MSE) was computed as a measure of the model's prediction accuracy.

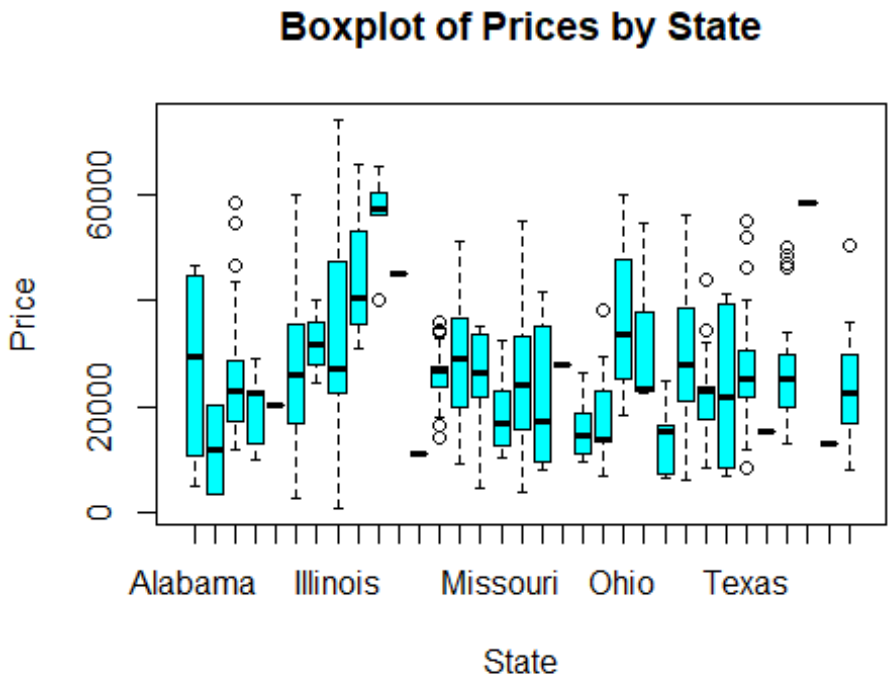
Random Forest Model: The Random Forest model, an ensemble learning method, was used. This model is known for its high accuracy, ability to handle a large number of features, and robustness to overfitting. Like LASSO, its performance was evaluated using the MSE.

Performance Metrics: Both models were assessed based on their MSE values, with the Random Forest model showing a lower MSE compared to the LASSO model, indicating potentially better predictive performance.

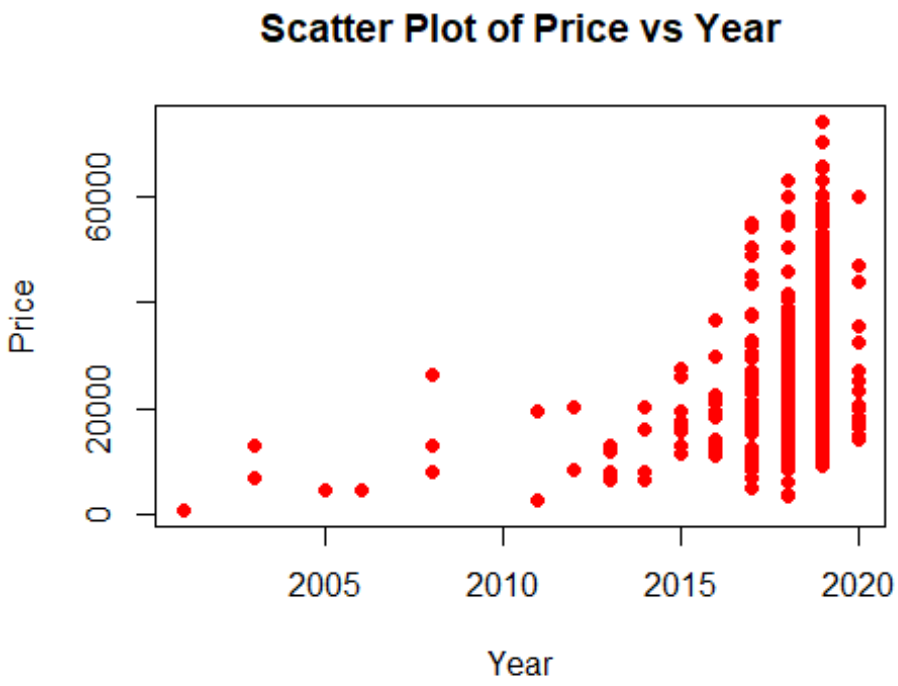
```r
hist(data$Price, main="Histogram of Car Prices", xlab="Price", col="blue",
border="black")
```



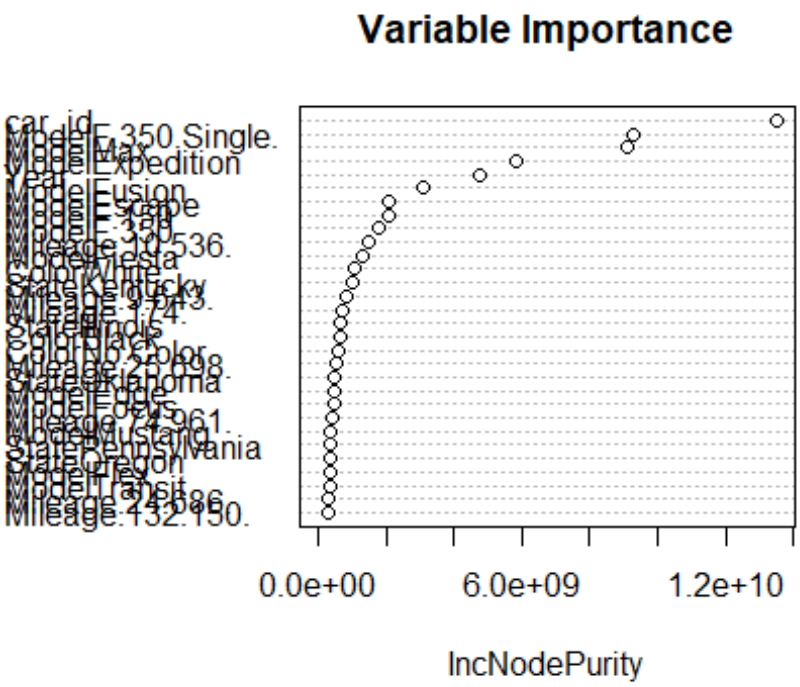

```r
boxplot(Price ~ State, data=data, main="Boxplot of Prices by State",
xlab="State", ylab="Price", col="cyan")
```



```r
plot(data$Year, data$Price, main="Scatter Plot of Price vs Year",
xlab="Year", ylab="Price", pch=19, col="red")
```




```r
importance <- randomForest::importance(rf_model)
varImpPlot(rf_model, main="Variable Importance")
```



Graphical Analysis and Insights:

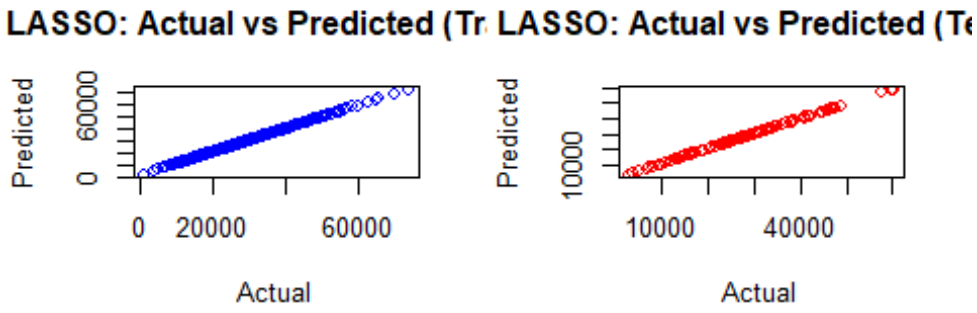
Residual Plots: Residual plots for both models were generated. These plots are crucial for diagnosing the model by showing the differences between the observed and predicted values. Ideally, the residuals should be randomly scattered around zero, indicating that the model does not suffer from non-linearity or heteroscedasticity.

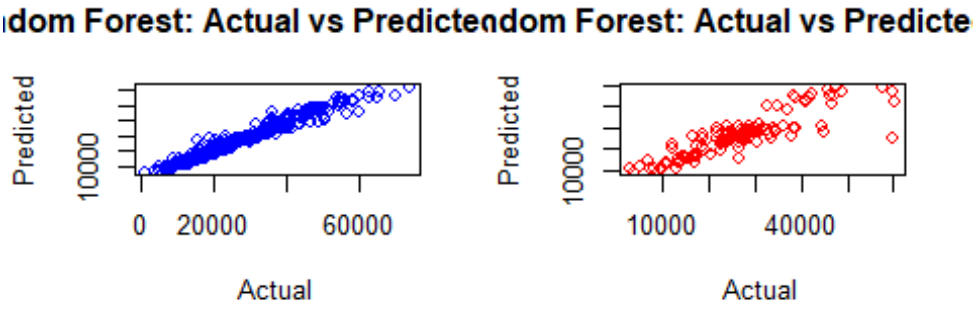
Actual vs. Predicted Plots: These plots offer a direct visual comparison between the actual and predicted values. They are instrumental in assessing the model's accuracy, with points lying close to the diagonal line indicating higher accuracy.

Feature Importance (Random Forest): The Random Forest model also provided insights into the importance of different features in predicting the car prices. This information can be pivotal in understanding which factors most significantly drive the price.

```
# Revised Code for Prediction vs Actual Plot

# LASSO Predictions for Train and Test
lasso_train_predictions <- predict(lasso_model, s = 0.1, newx =
as.matrix(train_data[-ncol(train_data)]))
lasso_test_predictions <- predict(lasso_model, s = 0.1, newx =
as.matrix(test_data[-ncol(test_data)]))

# Random Forest Predictions for Train and Test
rf_train_predictions <- predict(rf_model, train_data)
rf_test_predictions <- predict(rf_model, test_data)
```

```r
# Plotting
par(mfrow=c(2,2))

# LASSO: Actual vs Predicted
plot(train_data$Price, lasso_train_predictions, main="LASSO: Actual vs
Predicted (Train)", xlab="Actual", ylab="Predicted", col="blue")
plot(test_data$Price, lasso_test_predictions, main="LASSO: Actual vs
Predicted (Test)", xlab="Actual", ylab="Predicted", col="red")

# Random Forest: Actual vs Predicted
plot(train_data$Price, rf_train_predictions, main="Random Forest: Actual vs
Predicted (Train)", xlab="Actual", ylab="Predicted", col="blue")
plot(test_data$Price, rf_test_predictions, main="Random Forest: Actual vs
Predicted (Test)", xlab="Actual", ylab="Predicted", col="red")
```






```r
# Reset plot layout
par(mfrow=c(1,1))
```

Conclusion and Recommendations:

The Random Forest model demonstrated a stronger performance in terms of MSE, suggesting it may be more suitable for this particular dataset.

The LASSO model, with its feature selection capability, provided valuable insights, especially in a dataset with a high number of features.

It’s recommended to explore further model tuning and cross-validation to enhance the model’s performance. The analysis offers insights that can assist in strategic decision-making, particularly in understanding the factors that influence car prices.

This analysis provides a comprehensive understanding of the dataset and the effectiveness of different modeling approaches, offering a solid foundation for further exploration and decision-making.