**Study Problem and Hypothesis:**

Using a dataset collected from Carvana’s online site, can machine learning models, such as Decision Trees (DTC) and Multiple Linear Regression (MLR), predict car prices based on the research dataset?

Hypothesis: Null hypothesis - The selected machine learning models, including DTC and MLR, cannot accurately predict the prices of cars from the research data with a mean absolute percentage error (MAPE) of less than 20%.

Alternate Hypothesis - The selected machine learning models, including DTC and MLR, can accurately predict the prices of cars from the research data with a mean absolute percentage error (MAPE) of less than 20%.

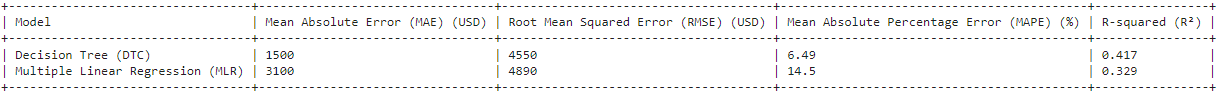
**Data Analysis Process:**

A simple cleaning process of removing null values and false years was done to remove outliers and make the predictive models easier to function. Scatterplots of the cleaned data were created to visualize price against Mileage and Years of the cars.

**A graph of a graph with numbers and dots

Description automatically generated with medium confidence**A graph of blue dots

Description automatically generatedThe study design involves using DTC and MLR for categorizing car prices and predicting them based on features of year and mileage. Visual assessment tools like scatter plots and histograms of residuals will provide insights into data distribution and model performance. The dataset will be split into 80% training and 20% testing. A paired T-tests will compare Mean Absolute Errors (MAE) between models to ensure significance between the two models. Calculations of MAE, Root Mean Squared Error (RMSE), MAPE, and R-squared (R2) values were run for both DTC and MLR models for statistical comparison.

****

**Study Findings:**

The analysis reveals notable differences between the DTC and MLR in predicting car prices. DTC shows a lower MAE, RMSE, and MAPE compared to MLR. The R² values for both models indicate that DTC better explains the variance in the target variable compared to MLR. Due to the DTCs lower MAE, RMSE, and MAPE we see a lower margin of loss regarding USD when placing our predictive prices to use. As a car company, the need to accurately predict car prices with minimalist margin of error is essential to stay competitive and within market values. Based on the MAPE, DTCs showcase a 6.49% error from the actual price, whilst MLRs hold 14.5% from the actual. Essentially, DTCs mean percentage of error will hold a mean loss of 6.49% in car value if incorrectly predicted than compared to MLRs 14.5%.

A blue dotted graph with white text

Description automatically generatedA diagram of a blue dotted line

Description automatically generated with medium confidenceThe rejection of the paired t-test null hypothesis suggests a significant difference in predictive performance between the two models, making the analysis statstically signifanct between the predictive models. In the context of the research question—to what extent can machine learning models, such as DTC and MLR, predict car prices on the research dataset?—the analysis concludes that both DTCs and MLRs both hold credibility as both were able to produce MAPEs of less the 20%.

**Study Limitations:**

The data lacks other variables that can influence car pricing, such as color, package details, and information about damages. These details, while not included in this dataset, can be considered in future studies to potentially improve the accuracy of price predictions. Additionally, the dataset is specific to the United States market and does not account for variations in other global markets. Consequently, the predictive models may have limited applicability beyond the U.S. market, although the underlying methodology can be adapted for different datasets. Due to the simplicity of the data, no delimitations would be set.

The advantage of sourcing my data from Kaggle is due to its readily availability and usability. The data is downloadable as .csv file which is easy to open and load into Pandas as a dataframe. With that, Kaggle users tend to clean and preprocess their datasets before uploading, thus making that process easier for this and other further analysis.

The disadvantage that arises is that the data lacks updated information. Consequently, trends, seasonalities, and the market might be missed until a new upload is provided. This does not alter the functionality of the models and only affects the dollars that might be gained or lost at the time of the analysis.

The selection of the analysis techniques, Decision Tree Model (DTC) and Multiple Linear Regression (MLR), was based on their suitability for predicting the 'Price' variable in the dataset. One advantage of using the DTC lies in its ability to handle non-linearity and interactions among features effectively. It can capture complex relationships in the data without assuming a specific functional form, making it a flexible choice for predictive modeling. On the other hand, a disadvantage of the DTC is its tendency to overfit the training data, especially with a deep tree. Overfitting can lead to excellent performance on training data but poor generalization to unseen data, impacting the model's predictive accuracy.

MLR, on the other hand, is advantageous for its interpretability and simplicity. It allows us to understand the relationship between the predictor variables and the target variable in a straightforward manner, expressed through coefficients. However, a limitation of MLR is its assumption of a linear relationship between predictors and the target variable. When this assumption is violated, the model's predictions may be inaccurate.

Considering these factors, both DTC and MLR were chosen as they offer complementary advantages, with DTC being capable of capturing complex relationships while MLR providing interpretability. The ensemble of these models allows for a robust evaluation of the dataset and adds to Jin (2021) previous studies of machine learning models.

**Recommended Action:**

From the analysis, the best course of action would be to use DTCs for predicitve car prices due to it lower margin of error in comparison to MLRs models. For further analysis, a recommendation of increasing variables of car determinates and a larger dataset would be essential to improving the accuracy of of future models and lowering margin of loss. Another recommendation is to experiment with other models to determine if they will outpreform DTCs.

**Expected Benefits:**

Expected benefits of the adopted predictive DTC model is increased accuracy of car prices with a 6.49% mean margin of loss if incorrectly forecasted. This provides an overall increase in profits for Carvana and other car companies that adopt this model of prediction. As a consumer, buyers will appreciate the knowledge that they are purchasing cars at relative market value, thus furthering increasing sales.