Fuel Efficiency Prediction using Random Forest Regression

Thomas Strade, Dylan Bruno, Adni Onoh

November 26, 2024

EAS 345 – Introduction to Data Science

Dr. Sabato

Table of Contents

Summary

The Fuel Efficiency predictive model is based on the dataset “Car Specification Dataset 1945 – 2020” [1]. The original dataset contains 78 columns, each with approximately 70,000 rows consisting of different car specifications such as make, model, years of production, fuel economy, weight, and more. From this set, the data will be cleaned and reduced to 18 columns of interest that will be used to train a Random Forest Regression model. This model is a supervised machine learning algorithm that randomly selects subsets of the training data to build decision trees that are then combined to form the complete model, outputting a single prediction value. The target variable of the model is fuel efficiency, measured in units of miles per gallon, corresponding to the 18th column of the cleaned dataset. The original and cleaned datasets, relevant R script, and all documentation are contained in the Fuel Efficiency GitHub Repository.

Data Cleaning

The original date of completion set for the data cleaning was November 11th; however, the data was not fully cleaned until November 21st. This sets the project about 2 weeks behind schedule. The original data links and their cleaned csv-formats are found in the “Data Sources” directory of the GitHub repository linked above. Additionally, the R script used to clean this data can be found in the “Codebase” directory. It has been determined that the “UCI Auto MPG” dataset will not be used for the training or testing of this model for two reasons: lack of satisfactory variable overlap and poor labeling. Second, the setback in the project schedule demands that only one dataset be used to attempt to build, train, and test the model by the deadline. The choice will not greatly impact the training of the model – the “Car Specification Dataset 1945 – 2020" data, after cleaning, still contains over half a million points of data, which can easily be expanded by considering other variables of interest.

The Random Forest regression model will be trained on a data set with the following variables: Body type, number of seats, (body) length in millimeters, curb weight in kilograms, maximum torque in Newton-meters, injection type, cylinder layout, number of cylinders, compression ratio, engine type, valves per cylinder, boost type, engine placement, engine horsepower, drive wheels, number of gears, transmission, fuel grade, and fuel economy (the variable to be predicted). These were chosen based on hypothesized factors that impact fuel economy. Additionally, the amount of present and missing data was considered for any variables that would potentially be used to predict the fuel economy. After selecting parameters to use for training, the missing data was handled in two ways: rows with missing numerical data were simply removed, and rows with missing string data were filled in with an “Other” label. There are also two versions of the cleaned data in the GitHub repository: one with the string values as-is and the other with the string values converted to unique integers. This decision was made to ensure that the data could be manipulated in multiple ways, regardless of data types.

Exploratory Data Analysis

The exploratory data analysis is also behind schedule, but it is nearly done as of November 24th, 2024. Each of the 18 variables have been plotted against the Fuel Economy values so that the average point of each variable could be observed visually. Data that consisted of densely packed values (a large quantity of whole numbers through a range of several hundreds or thousands of units) was also fit with `geom_smooth()`. Additionally, these columns of dense data were graphed on a density plot. The maximum torque (N-m) and horsepower data exhibited a more normal distribution on the logarithmic scale, whereas the original values were skewed left. All these graphs can be found on the GitHub repository under the "Visuals/Density_Distributions" directory.

Additionally, preliminary regression models were created for the dataset using linear regression for one model and logistic regression for the other. The dataset used here was not the regular cleaned dataset, but a modified version of the "Model_CarSpecs_1945to2020.csv" that replaced any remaining N/A values with the column's average value. This method was not used in the original cleaning process, which will be revisited before the initial random forest model is built. Through this, more rows of data will be recovered, and more variables can be included to predict the fuel economy value. The linear regression summary showed that with all variables included, the model had a promising R-squared score of 0.76. The linear regression summary using the log of the torque and horsepower had an R-squared score of 0.77. Visually speaking, the relation between the data does not appear linear. To build the logistic regression model, the Fuel Economy column was normalized to sit between 0 and 1. The summary and graph are as follows:

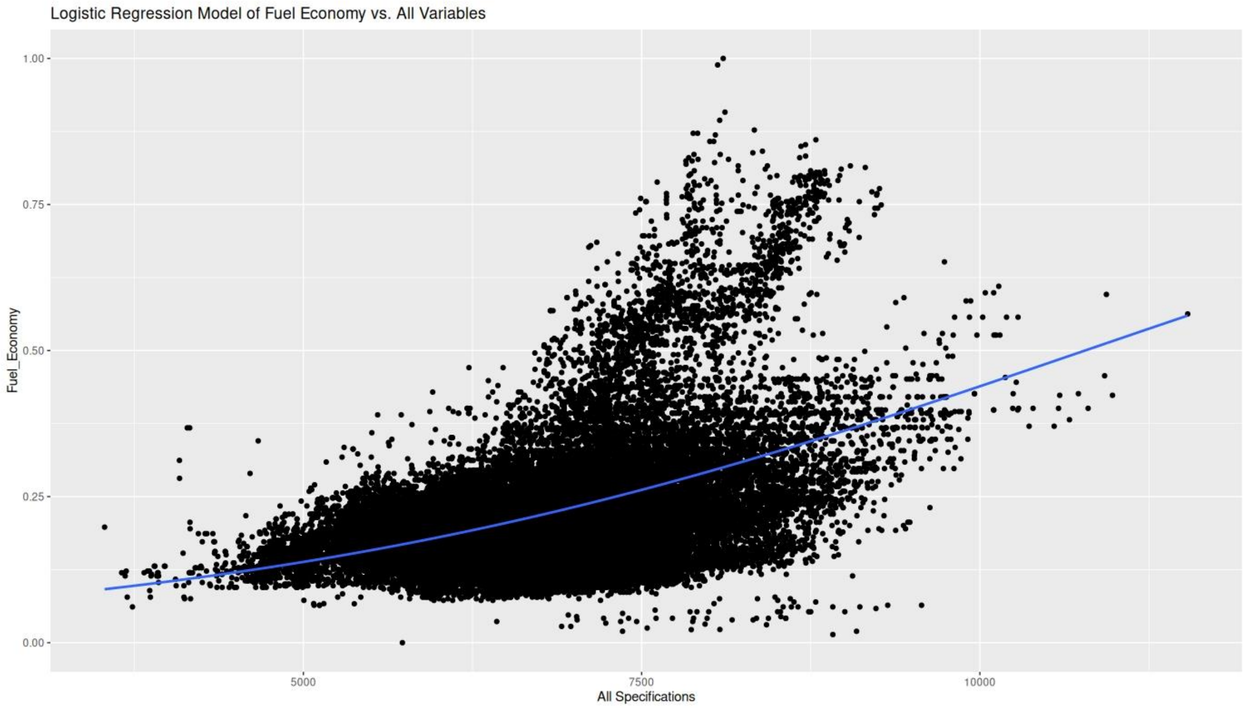
```
Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)         -2.575e+00  1.782e-01  -14.447  < 2e-16 ***
Body_type           -2.288e-03  5.946e-04   -3.848 0.000119 ***
number_of_seats      9.296e-03  1.851e-03    5.021 5.16e-07 ***
length_mm            2.012e-04  5.731e-06   35.108  < 2e-16 ***
curb_weight_kg       3.796e-04  8.139e-06   46.639  < 2e-16 ***
maximum_torque_n_m  -8.061e-05  2.972e-05   -2.712 0.006685 **
injection_type       5.778e-03  2.601e-04   22.212  < 2e-16 ***
cylinder_layout      1.274e-02  2.167e-03    5.879 4.17e-09 ***
number_of_cylinders  8.411e-02  1.975e-03   42.586  < 2e-16 ***
compression_ratio    1.201e-03  1.154e-04   10.405  < 2e-16 ***
engine_type         -7.856e-02  2.305e-03  -34.076  < 2e-16 ***
valves_per_cylinder -7.950e-02  1.868e-03  -42.548  < 2e-16 ***
boost_type          -1.444e-03  2.781e-05  -51.908  < 2e-16 ***
engine_placement     8.656e-04  5.701e-05   15.183  < 2e-16 ***
engine_hp            1.992e-03  4.048e-05   49.217  < 2e-16 ***
drive_wheels         2.111e-02  1.760e-03   11.991  < 2e-16 ***
number_of_gears     -2.180e-01  1.790e-03 -121.803  < 2e-16 ***
transmission         4.392e-02  2.536e-03   17.322  < 2e-16 ***
```

```
fuel_grade        -1.339e-02  7.967e-04  -16.806  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.01295602)

    Null deviance: 2377.68  on 39093  degrees of freedom
Residual deviance:  504.29  on 39075  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```



Modeling

The goal is to build a predictive model to estimate fuel efficiency (measured in miles per gallon) based on car specifications. This model aims to help identify key factors influencing fuel economy using the "Car Specification Dataset 1945 – 2020." The ultimate goal is to provide reliable predictions by leveraging the Random Forest Regression method. The modeling phase is behind schedule due to delays in data cleaning and exploratory data analysis, originally planned to be completed by November 11th. As of November 24th, preliminary models (linear and logistic regression) have been implemented for comparison purposes, though the Random Forest Regression modeling is still pending. To mitigate this delay, we plan to accelerate model optimization and testing, focusing exclusively on finalizing the model by the deadline. This is a regression problem, as the target variable (fuel economy) is a continuous value. After exploring preliminary models like linear regression, the final model will use Random Forest Regression because it is well-suited for the complexity of this dataset for the following reasons:

- Handles Non-Linearity and Interactions:

  Preliminary linear and logistic regression models showed promising R-squared values but did not fully capture non-linear relationships between car specifications and fuel efficiency. Random Forest models excel in such cases because they are non-parametric, meaning they do not assume a linear relationship between input variables and the target variable.

  For example, variables like engine horsepower or body type may interact with other factors (e.g., curb weight or drive wheels), creating complex patterns in the data. Random Forest can naturally capture these interactions.

- Handles Noise and Outliers:

  Random Forest reduces overfitting by aggregating predictions from multiple decision trees built on bootstrapped samples of the data. This approach minimizes the influence of noisy or extreme data points, making the model more robust compared to individual trees or simpler regression methods.

- Features Important Insights:

  Random Forest models provide a ranking of variable importance, showing which car specifications (e.g., engine horsepower, transmission type, fuel grade) contribute most to fuel economy predictions. This can offer valuable insights into the key drivers of fuel efficiency.

- Handles High Dimensionality:

  The cleaned dataset currently includes 17 independent variables, many of which are categorical (e.g., body type, drive wheels, injection type). Random Forest can easily handle a mix of categorical and numerical data, making it a good fit for the problem at hand.

- Flexibility and Predictive Accuracy:

  Random Forest models tend to achieve high predictive accuracy because they reduce variance (overfitting) without greatly increasing bias. This makes it an effective tool for real-world datasets like ours, where relationships between variables are complex.

Main R Tools (Libraries/Functions) to be used:

- Libraries
    - ggplot2
    - dplyr
    - randomForest
    - caret
- Functions
    - randomForest()
    - train() from caret
    - geom_smooth() from ggplot2
    - predict()

Function Purposes

- randomForest(): Fits a Random Forest model by training multiple decision trees on bootstrapped samples of the data and aggregating their predictions to improve accuracy and reduce overfitting.
- train(): Optimizes the hyperparameters of the Random Forest model (e.g., the number of trees, depth) to improve prediction performance.
- predict(): Generates predicted fuel economy values for new data.
- ggplot(), geom_smooth(): Provides visual insights into variable relationships with fuel economy.

RandomForest() Inputs / Outputs Interpretations [2]

Inputs:

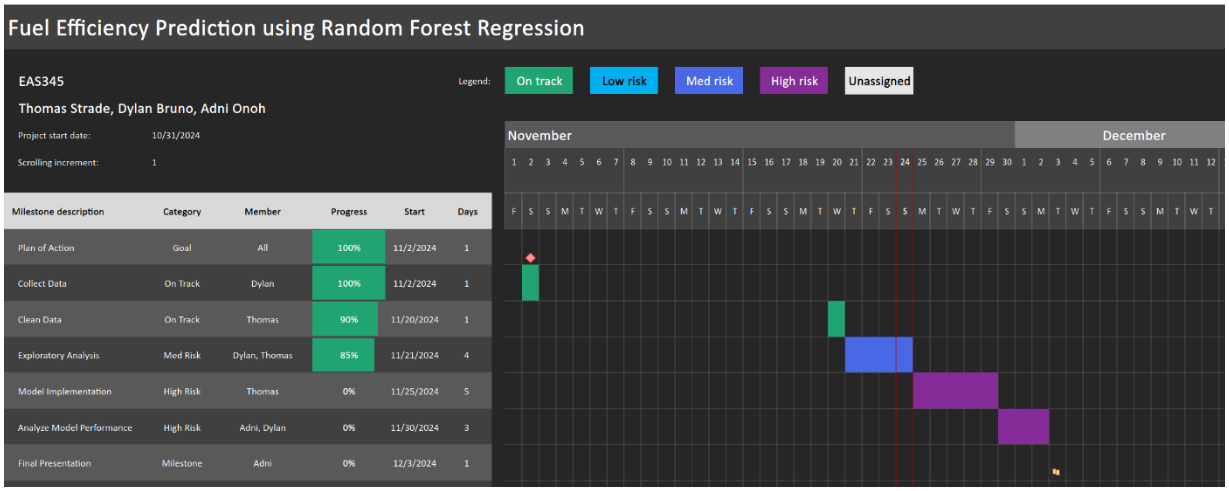
- formula = y ~ x (where x will be all parameters in the data frame)
- data = data.frame
- xtest = data frame containing predictors for the test set
- ntree = number of trees to grow (should be large, such as > 1000)
- keep.forest = boolean (whether to retain the forest in the output or not)
- importance = boolean (whether to evaluate the importance of the predictors or not)
- proximity = boolean (whether to calculate the proximity among the rows or not)
- oob.prox = boolean (whether proximity should be calculated only on “out-of-bag" data or not)
    - “out-of-bag" data is the test data, which can be used to evaluate the performance of the model

Outputs:

- call = original function call
- type = regression
- predicted = predicted values of the input data
- importance = two-column matrix where the first column is the mean decrease in accuracy and the second is the mean decrease in the mean square error
- importanceSD = vector of the standard errors of the permutation-based importance measure
- localImp = matrix of casewise importance measures, corresponding to the importance of the *i*-th variable on the *j*-th case
- ntree = number of trees grown
- mtry = number of predictors sampled for splitting at each node
- forest = list containing the entire forest
- oob.times = number of times cases are “out-of-bag"
- proximity = matrix of proximity measures among the input (based on the frequency that pairs of data points are in the same terminal nodes)
- mse = vector of mean square errors (sum of squared residuals divided by *n*)
- rsq = pseudo R-squared (1 – mse / Var(y))
- test = list of the corresponding predicted, mse, rsq, and proximity for the test set

Project Prognosis

Overall, the project is behind the originally planned schedule, but it is still likely to be completed on time. By November 30th, 2024, the cleaned dataset will be revised, and the regression modeling will be almost, if not entirely, complete. Based on the preliminary linear and logistic regressions, it is anticipated that the model will yield useful results. There will be enough time to optimize and evaluate the model’s performance by including more data and comparing the resulting test statistics from each iteration of the model. Figure X shows the revised Gantt chart.



Fuel Efficiency Prediction using Random Forest Regression

EAS345

Thomas Strade, Dylan Bruno, Adni Onoh

Project start date: 10/31/2024

Scrolling increment: 1

Milestone description	Category	Member	Progress	Start	Days
Plan of Action	Goal	All	100%	11/2/2024	1
Collect Data	On Track	Dylan	100%	11/2/2024	1
Clean Data	On Track	Thomas	90%	11/20/2024	1
Exploratory Analysis	Med Risk	Dylan, Thomas	85%	11/21/2024	4
Model Implementation	High Risk	Thomas	0%	11/25/2024	5
Analyze Model Performance	High Risk	Adni, Dylan	0%	11/30/2024	3
Final Presentation	Milestone	Adni	0%	12/3/2024	1

Attachments

References

[1] Islam, J. (2023, February 25). *Car specification dataset 1945-2020*. Home Page. https://doi.org/10.34740/KAGGLE/DS/2938279

[2] Breiman , L., Cutler, A., Liaw, A., & Wiener , M. (2024, September 22). *RandomForest.pdf*. Cran R-Project. https://cran.r-project.org/web//packages/randomForest/randomForest.pdf

October 23, 2024

Fuel Efficiency Prediction using Machine Learning

Dylan Bruno, Thomas Strade

EAS 345

Dr. Sabato

Table of Contents:

Summary

This project aims to develop a predictive model for fuel efficiency based on various vehicle characteristics, such as engine type, fuel type, horsepower, number of cylinders, and manufacturer. The outcome will be a concrete and reliable estimate on the vehicle's rating of miles per gallon (mpg), where a higher mpg typically indicates a more environmentally friendly vehicle. Publicly available datasets are the main source of data that will be used to build a random forest regression (RFR) algorithm, training the supervised machine learning model (MLM) to predict the mpg rating for gasoline-based vehicles from 1996 onwards. The primary datasets used to train and test the supervised MLM are the UCI Auto MPG dataset (Quinlan, 1993) and the Car Specification Dataset 1945-2020 (Islam, 2023). These sets will be cleaned using the R library collection *Tidyverse,* which consists of packages such as *dplyr* and *ggplot*2. The data collected will be cleaned such that the supervised model will only be concerned with the type of car (sedan, SUV, truck, etc.), engine size, number of cylinders, type of transmission, CO2 rating, and fuel type. Other data in the sets, such as make, model, and year, will not be used by the model. The goal is to provide a reliable tool that can predict fuel efficiency for gasoline vehicles, which can be useful for both manufacturers and consumers.

Introduction

Fuel efficiency is a critical factor in the automotive industry, impacting both economic and environmental concerns. Consumers, manufacturers, and, most importantly, policymakers' understanding of what factors affect a vehicle's fuel efficiency is of growing importance in reducing our carbon footprint on the world. Machine learning models offer opportunities to understand the intricate relationships behind common sources of carbon emissions, like the automotive industry. This proposal outlines the methods and plan of action towards training a supervised MLM, based on a RFR algorithm, that predicts a vehicle's fuel efficiency with reliable accuracy. Using a Gantt Chart to be shown later in the proposal, the project lead will ensure the team meets its scheduled deliverables. The MLM-creation process starts with the data architect, who will collect and clean the data into a set that can be used to train the supervised MLM. Once cleaned, the data scientist will be able to split the data sets into subsets, consisting of a training and a testing dataset. The MLM will then be trained based on the RFR algorithm, ultimately undergoing a series of performance tests using methods such as Mean Absolute Error

(MAE), Root Mean Square Error (RMSE), and R-squared (R^2). This data-driven model will guide manufacturers through the development of new vehicles, providing accurate and reliable feedback that predicts the vehicle's efficiency based on its specifications.

Background

Fuel efficiency plays an essential role in the design of vehicles. Its importance has recently grown due to environmental regulations and the desire to reduce environmental damage. Various factors, including engine displacement, vehicle weight, and horsepower, influence a car's fuel efficiency. Traditional approaches to predicting MPG, such as linear regression, offer basic insights, but machine learning techniques can significantly improve accuracy by accounting for non-linear relationships between variables; for example, this model will use a RFR algorithm. There have also been other models that also predicts mpg, but they focus on an older dataset while this project focuses on the modern era of vehicles. This project will employ two different datasets to enhance model performance: the Auto MPG dataset and the Car Specification Dataset 1945-2020. The range and volume of these datasets will help both test and train the model.

Roles and Responsibilities

Project Lead:

The project lead will oversee the entire project, ensuring that all team members are aligned and that the project stays on schedule. They will coordinate meetings, manage deadlines, and provide directions for the project's timely completion. The project lead will also ensure that all deliverables are met according to the project plan.

Data Architects:

The data architects are responsible for acquiring, cleaning, and organizing the datasets. They will ensure the data is consistent before handing it off for model development. They will design and implement the functions needed for efficient data processing and manage the storage and version control of both the raw and processed data.

Data Scientists:

The data scientists will focus on analyzing the datasets and developing the machine learning model. They will split the data into training and testing sets, select appropriate algorithms, and evaluate model performance using metrics such as MAE, RMSE, and R-squared. They will also perform hyperparameter tuning and optimize the model to increase prediction accuracy and reduce overfitting.

Communication Liaison:

The communication liaison will be responsible for maintaining clear and consistent communication between all stakeholders, including the team and any external collaborators or advisors. They will provide updates on the project's progress, ensure that all members are informed of changes or issues, and manage any required presentations or reports.

Scribe:

The scribe will document all key discussions, decisions, and action items during team meetings. They will maintain organized records of meeting notes, deliverables, and any changes to the project plan. The scribe will also assist the communication liaison by preparing reports and summaries for stakeholders/sponsors.

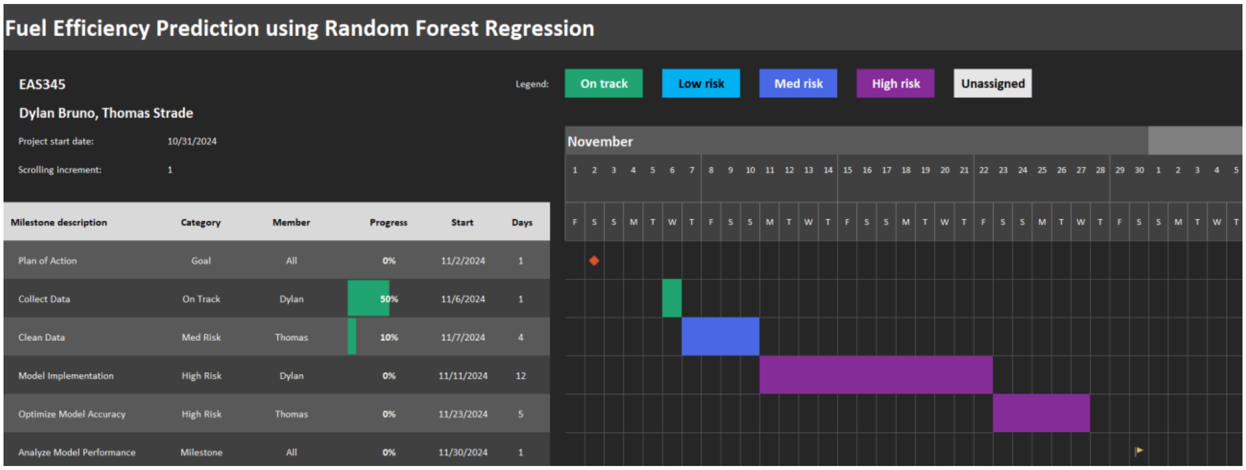
Plan of Action

The Plan of Action for this project begins with data collection and preparation. The data architects will be responsible for acquiring and preprocessing the datasets. These include the Auto MPG dataset (Quinlan, 1993) and the Car Specification Dataset 1945-2020 (Islam, 2023). The project will focus on vehicles manufactured after 1996, as this reflects advancements in automotive technology and emissions standards like OBD-II, which was introduced that year. Once the data is collected, the architects will clean and organize the information, ensuring that the dataset is of high quality. The R library collection *Tidyverse,* and the packages it includes, will be used to remove unnecessary data such as make, model and year of any given vehicle. Afterward, the data will be split into training and testing sets, allowing for robust model development and evaluation.

Once the data is prepared, the model selection and training phase will commence. The data scientists will lead this effort, implementing Random Forest Regression as the primary model due to its capacity to handle non-linear relationships and reduce the risk of overfitting. The training data will be used to build the model, while the testing data will evaluate the model's ability to generalize beyond the training set. Throughout this process, the data scientists will ensure that the model is fine-tuned, optimizing hyperparameters to maximize predictive accuracy.

In the final stage, the team will focus on model evaluation. The performance of the model will be assessed using key metrics, including MAE, RMSE, and R^2. These metrics will provide insight into how well the model predicts fuel efficiency. Additionally, the data scientists will make adjustments based on the evaluation results, while the data architects ensure the scalability of the model and its potential integration with future datasets. This comprehensive approach will ensure a reliable and accurate model for predicting vehicle fuel efficiency.

Schedule



Fuel Efficiency Prediction using Random Forest Regression

EAS345

Dylan Bruno, Thomas Strade

Project start date: 10/31/2024

Scrolling increment: 1

Milestone description	Category	Member	Progress	Start	Days
Plan of Action	Goal	All	0%	11/2/2024	1
Collect Data	On Track	Dylan	50%	11/6/2024	1
Clean Data	Med Risk	Thomas	10%	11/7/2024	4
Model Implementation	High Risk	Dylan	0%	11/11/2024	12
Optimize Model Accuracy	High Risk	Thomas	0%	11/23/2024	5
Analyze Model Performance	Milestone	All	0%	11/30/2024	1

Schedule for Each Role

The project lead will oversee the entire development of the predictive model for fuel efficiency, starting with the proposal and continuing throughout the project. They will present and finalize the proposal while managing progression between milestones and coordinating the team during the model development phase.

The data architect is responsible for collecting, observing, and gathering data for model training and testing, ensuring its usability for analysis. Following data collection, the architect will focus on cleaning the data and preparing it for machine learning applications.

The data scientist will lead the implementation of the Random Forest Regression model, integrating the cleaned dataset into the model. They will then work on optimizing the model for accuracy, fine-tuning it to maximize its accuracy. After optimization, the data scientist will evaluate the model's performance using the following metrics: RMSE, MAE, and R^2.

Throughout the project, the communication liaison will communicate between team members and provide updates to stakeholders/sponsors regarding project progress, including any delays or risks. The scribe will document decisions made to model and the dataset. They will make sure that each team member is aligned with the project milestones and has clearly defined responsibilities related to the project's progress.

Conclusion

This project will develop a reliable machine-learning model that predicts vehicle fuel efficiency using publicly available data. By combining the Auto MPG dataset and the Car Specification Dataset, the model will capture a comprehensive range of new vehicle characteristics to provide accurate MPG predictions. This proposal should be chosen because it addresses a highly relevant concern and constraint in the automotive industry. Increasing pressure from climate scientists, public demand, and government policy push the automotive industry to prioritize characteristics such as fuel efficiency. The growing popularity and financial viability in electric vehicles make it necessary for manufacturers to take steps to keep up with advances in the industry. This project applies the predictive MLM to assist manufacturers with designing fuel efficient vehicles and to provide consumers with insights when making purchasing decisions.

References:

Quinlan, R. (1993). Auto mpg [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5859H.

Jahaidul Islam. (2023). Car Specification Dataset 1945-2020 [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/2938279