***Automobile Mileage Prediction***

Project Report submitted by

***Ajay Dyavathi (1308963)***

***Aniket Jadhav (1319758)***

***Mairaj Mohsin Mohammed (1309404)***

***Nandini Bhusanurmath (1305500)***

For the course

**DTSC 701 – M02 Big Data Analytics**

**A picture containing text

Description automatically generated**

Under the supervision of

**Prof. Xueqing Huang**

On

Dec 26, 2022

**Abstract**

In this paper, we analyze the application of supervised machine learning techniques with big data tools like pyspark to predict automobile characteristics such as MPG based on various features such as acceleration, horsepower, displacement, etc. The data is collected from the UCI Machine Learning Repository. Different techniques like multiple linear regression analysis, decision tree regressor have been used to make the predictions. The predictions are then evaluated and compared to determine which ones provide the best outcome. Both methods produced comparable results. In the future, we intend to use more sophisticated algorithms to make predictions.

1. **Introduction**

In this project, we analyze the data set and discover the fundamental relationships between the variables and the performance of an automobile. We'll start with a quick experimental analysis with the help of a graphical display and summary of the data. We created a linear regression model using a machine-learning pipeline to forecast the mileage of an automobile. In addition, we will conduct testing on the process underlying the study variable. Furthermore, we want to examine if we can predict how well the cars will perform based on the data we have.

1. **Technical Details**

In this research, the presented methodologies are applied to predict the mileage of an automobile. Machine learning algorithms, such as Linear Regression and Decision Tree Regressor are compared to see which one predicts the best and has the minimum margin of errors.

**2.1 Dataset**

Here in this project, we have used the data from UCI Machine Learning Repository.

<https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/>

Also, the pandas is used for data structure design, which is fast, flexible, and expressive.

Table

Description automatically generated with medium confidence

*Fig 2.1.1 Reading the dataset using pandas*

Spark Machine learning libraries need the data in the form of vectors to be compatible with models. Therefore, we used “VectorAssembler” to convert features to vector.

Graphical user interface, text

Description automatically generated

*Fig 2.1.2 Vectorizing the features*

Categorical variables are converted to One Hot Encoded vectors to be compatible with machine learning models. This is done by String-Indexer and One-Hot Encoder.

**Table

Description automatically generated**

*Fig 2.1.3 Variables converted using String-Indexer*

**Table

Description automatically generated**

*Fig 2.1.4 Variables converted using One-hot Encoder*

**2.2 Methods Utilized:**

The methods implemented in this project to make the predictions are:

* Linear Regression
* Decision Tree Regressor

Linear Regression:

Linear regression is a type of supervised machine-learning model that is commonly used in forecasting. Supervised machine learning models are those in which we use training data to build the model and then use the loss function to test its accuracy.

Linear regression is a well-known time series forecasting technique that is used in predictive modeling. As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

In this project, the linear regression model we have trained predicts the mileage of an automobile with label column as “MPG” using the input features Cylinders, Displacement, Manufacturer, Model Year, Origin, Weight, Acceleration, and Horsepower.

Graphical user interface, text, application, email

Description automatically generated

*Fig 2.2.1 Linear Regression.*

Decision Tree Regressor:

Decision Tree is one of the most used, practical approaches for supervised learning. It can be used to solve both Regression and Classification problems, with the latter being more practical. Decision tree regression observes an object's features and trains a model with a tree structure to predict data in the future to produce meaningful continuous output.

Graphical user interface, text, application

Description automatically generated

*Fig 2.2.2 Decision Tree Regression.*

**2.3 Model Evaluation:**

**RMSE: Root Mean Squared Error**

Root-Mean-Square-Error or RMSE is one of the most popular measures to estimate the accuracy of our forecasting model’s predicted values versus the actual or observed values while training the regression models or time series models. It measures the error in our predicted values when the target or response variable is a continuous number.

**R2 Score: R Squared score**

The R2 score is one of the performance evaluation measures for regression-based machine learning models. It is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.

Graphical user interface, text, application, email

Description automatically generated

*Fig 2.3.1 RMSE & R2 Scores for Linear Regression Model.*

Graphical user interface, text, application, email

Description automatically generated

*Fig 2.3.2 RMSE & R2 Scores for Decision Tree Regressor Model.*

**Discussions and Evaluations:**

Confront team members gave responsible direction and made the required preparations in order to actively discover the issues that customers were having and seek out options to help. We decided on a time slot and individual assignment where everyone contributed to resolving and improving the project.

There are numerous performance metrics that can be used across the models, but those used specifically for automobility include: vehicle miles traveled, vehicle hours of delay, travel time index, duration of congestion, average vehicle occupancy, vehicle miles traveled in congestion, level of service, auto speed, auto travel time reliability, planning time index, etc.

As we previously discussed, a performance car excels in the following areas:

1. Acceleration

2. Top speed

3. Cornering

4. Braking

**Conclusion:**

In this paper, we have examined the data, displayed the characteristics, and determined the relationships between them. The data set is split into a training set and a testing set. After that, we analyzed different predictive models. By the end of the project, we understand how the implementation of data handling techniques varies in the PySpark framework and Python Language. Using machine-learning algorithms such as linear regression and decision tree regressor, we have successfully predicted the MPG of automobiles and compared the models by evaluating RMSE and R2 scores.

**Individual member contributions to the group project**

1. Ajay Dyavathi (1308963)
2. Aniket Jadhav (1319758)
3. Mairaj Mohsin Mohammed (1309404)
4. Nandini Bhusanurmath (1305500)