IST 5520: Data Science and ML in Python

Project Proposal: Group 4

Group Members: ***Sayantan Majumdar, Dawit Wolday Asfaw, Cong Shen, Divya Reddy Manku, Akhil Reddy Annreddy***

1. Introduction

1.1 Problem description and Research Questions

This project aims to build a machine learning model to predict the Environmental Protection Agency (EPA) emission score from different vehicles operating in the United States. We will also be analyzing the factors having the most impact on this score which varies from 1 (worst) to 10 (best). In addition, we will compare different machine learning models to assess their performance on the dataset. Essentially, this is a classification problem wherein we will utilize the EPA vehicle and emission data sets to predict the EPA emission score (a discrete quantity). Such kind of modeling and analysis if pushed to a production environment could enable policymakers in addressing critical issues in the transportation sustainability industry.In this regard, the following research question needs to be answered:

Which predictors are statistically significant for predicting the EPA emission score?

Some specific research questions include:

a. Which machine learning model works best for this data set?

b. Which cars perform well (in terms of emissions) in both city and highway driving conditions?

1.2 Classification Analysis

The vehicle and car emissions data sets can be used to model a classification problem where we need to predict the EPA emission score (a unitless quantity have integral values in the [1,10] interval) based on the explanatory variables In this project, we are dealing with a high-dimensional classification problem that is concerned with both prediction (discrete EPA emission score) and inference.

1.3 Potential problems and Challenges

The major challenge in this project is to perform feature engineering. Appropriate data-preprocessing and feature engineering need to be performed to reduce the data dimensionality required to suitably address the above research questions.

1.4 Tentative Timeline

Phase	Activities	Completion
Data Collection	Collecting the required data set for the business analytics project	September 24, 2021
Kickoff	Understanding the project requirements and elicitation	October 1, 2021
Data Management	Data cleaning, pre-processing	October 20, 2021
Data Analysis	Evaluating different ML models	November 20, 2021
Project Submission	Report writing, preparing presentation, and proof-reading	November 30, 2021

2. Data Source and Collection

2.1 Data Source

This dataset can be found at https://www.fueleconomy.gov/feg/ws/index.shtml (https://www.fueleconomy.gov/feg/ws/index.shtml). The vehicle and emission data sets need to be linked based on the vehicle ID. The vehicle data set contains 83 explanatory variables (columns) that provide detailed car specifications and has 44075 rows. As for the emissions data set, it has 8 features and 42442 rows.

2.2 Collection

We found this data set on EPA fuel economy portal (https://www.fueleconomy.gov/). The EPA has generated annual reports (https://www.fueleconomy.gov/feg/pdfs/guides/FEG2021.pdf) but there are no existing publicly available notebooks having detailed machine learning workflows. Moreover, these two data sets in conjunction satisfy all the other project requirements ($\geq$20 columns and $\geq$1000 rows) listed as part of this group assignment.