Group Member: Jingkai Wang, Qixiang Jiang, Ruisi Wang, Xiaoli Fang
DS 4400 - Machine Learning
Prof. Bilal Ahmed
07/25/2023

Vehicle Analysis Project Proposal

I. Overview

This project aims to analyze the Vehicle Manufacturing Dataset available on Kaggle. This dataset contains valuable information about different vehicle types, manufacturers, and various production metrics, which can provide valuable insights when analyzed.

II. Objectives

The main objectives of the project are as follows:

Exploratory Data Analysis (EDA): Uncover the basic structure and relationships within the data.

Predictive Modeling: Train a machine learning model to predict vehicle manufacturing numbers based on other variables.

Visualization: Create clear and concise visualizations to display the results of the analysis.

Report: Create a final report that includes an in-depth explanation of the data analysis, methodologies, and results.

III. Methodologies

1. Data Cleaning and Preprocessing:

The initial step would involve cleaning and preprocessing the data to make it suitable for analysis and modeling. This may include handling missing data, dealing with outliers, and converting categorical data into numerical data, if necessary.

2. Exploratory Data Analysis (EDA):

We will explore the data to understand the relationships and patterns in the data. This will include statistical analysis and visualization to understand correlations and other relationships in the data.

3. Feature Engineering and Selection:

Based on the findings from the EDA, we will create new features that could enhance the performance of our predictive models. We will also select the most relevant features to avoid overfitting.

4. Machine Learning Modeling:

Based on the problem at hand, we will employ suitable machine learning algorithms. If the problem is a regression task (predicting manufacturing numbers), we might use Linear Regression, Decision Trees, Random Forest, Gradient Boosting, or Support Vector Machines. If it's a classification task (predicting vehicle types), we might use Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, or Neural Networks.

5. Model Evaluation and Optimization:

Each model's performance will be evaluated using appropriate metrics (Accuracy, F1-score). We will also fine-tune our models using techniques such as Grid Search and Cross-Validation to maximize their performance.

6. Visualization:

The EDA and predictive modeling results will be visualized using libraries such as Matplotlib, Seaborn, or Plotly.

7. Reporting:

A comprehensive report will be created that includes all steps of the analysis, from data preprocessing to the final model evaluation. The report will be written in a way that is accessible to both technical and non-technical audiences.

IV. Tools and Libraries

The project will mainly use Python for data analysis and modeling. The specific libraries will: We will use Pandas for data manipulation and analysis, and NumPy for numerical computations. Besides, for data visualization, we will use Matplotlib, Seaborn, Plotly, and Altair to create the graphs. As for the Machine Learning part, we will use Scikit-Learn to predict modeling and model evaluation, which will include Linear Regression, etc.

V. Expected Outcomes

At the end of this project, we aim to deliver a detailed analysis of the Vehicle Manufacturing dataset, a reliable and robust predictive model that can predict vehicle manufacturing numbers as well as visualizations of our findings and model results. Finally, we will generate a comprehensive final report to include all the details of the projects.

Reference

Car ID Serial No.	Brand	Model	Year	Color	Mileage	Price	Location
1 2000	Toyota 22% Honda 21% Other (1151) 58%	Civic 5% Camry 4% Other (1809) 90%	2015 2020	White 21% Blue 19% Other (1206) 60%	25.0k 70.0k	12.0k 29.0k	Los Angeles 12% New York 11% Other (1535) 77%
1	Toyota	Camry	2018	White	45000	18000	Los Angeles
2	Honda	Civic	2019	Blue	35000	16000	New York
3	Ford	Focus	2017	Silver	55000	14000	Chicago
4	Chevrolet	Cruze	2016	Red	60000	12000	Miami
5	Hyundai	Elantra	2018	Black	40000	15000	San Francisco
6	Toyota	Corolla	2020	Gray	25000	19000	Dallas
7	Honda	Accord	2019	White	30000	18000	Atlanta
8	Ford	Mustang	2015	Yellow	65000	22000	Phoenix
9	Chevrolet	Impala	2017	Black	55000	16000	Houston
10	Hyundai	Sonata	2016	Blue	50000	14000	Seattle

https://www.kaggle.com/datasets/arnavsmayan/vehicle-manufacturing-dataset