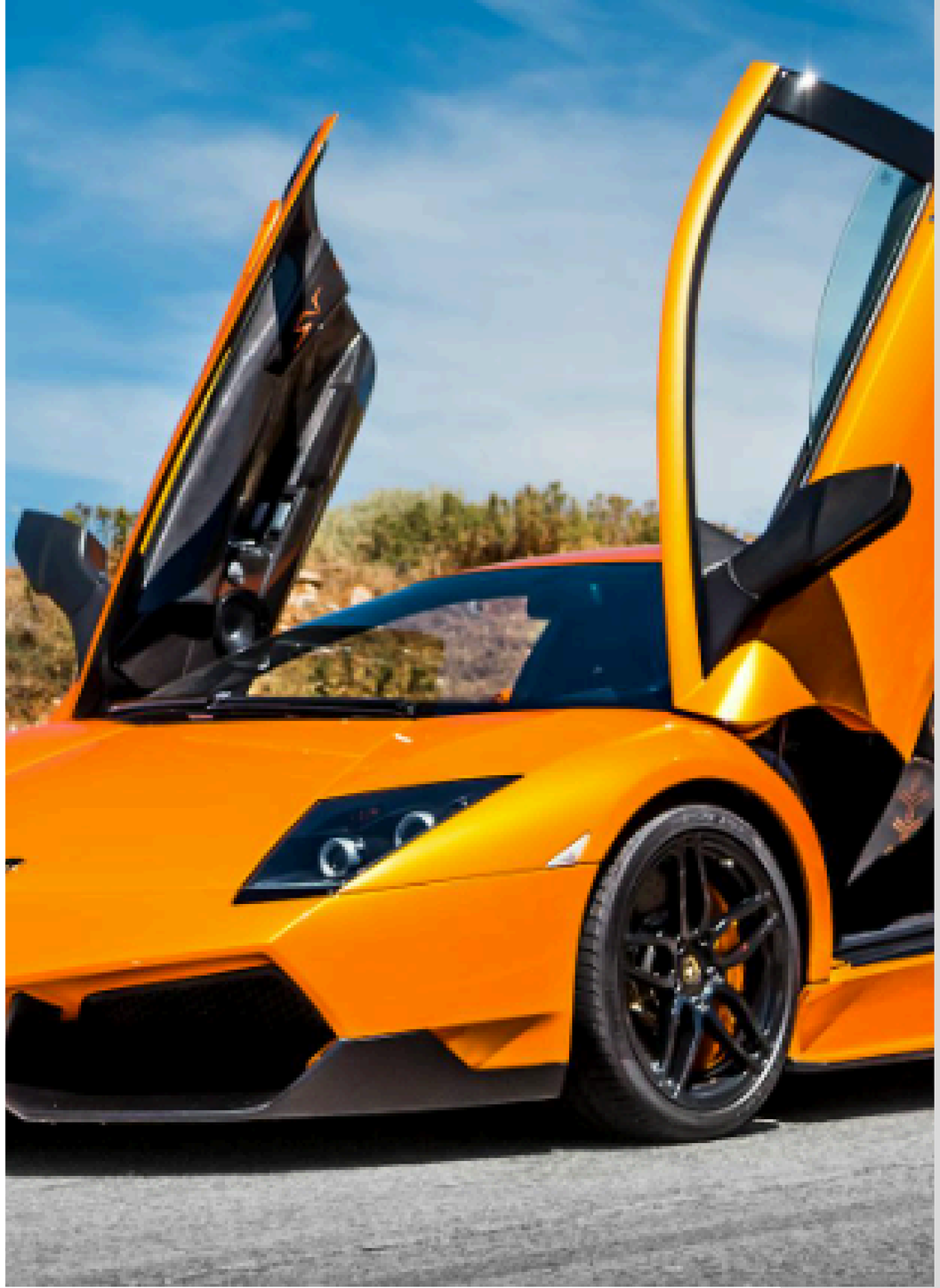
PROJECT REPORT

PRESENTATION

Introduction of Topic

Project tile:
MPG_Predictor

- Problem Statement: The challenge of accurately predicting a vehicle's fuel efficiency (MPG) based on its technical specifications.
- Project Goal: To develop a machine learning regression model that predicts MPG and deploy it as an interactive web application using Streamlit.
- Relevance: For consumer choice, environmental impact, vehicle design.



2. Data Sourcing and Exploration (EDA)

- Dataset: Introduce the "Auto MPG" dataset from the UCI Machine Learning Repository.
- Features: List and describe the key features used (e.g., Cylinders, Displacement, Horsepower, Weight, Acceleration, Model Year, Origin).
- Exploratory Data Analysis (EDA):
  - Distributions: What did the histograms of key features like MPG and Horsepower look like?
  - Correlations: Show a correlation heatmap. Discuss the strong relationships you found (e.g., the negative correlation between Weight and MPG).
  - Key Insights: "EDA revealed that Weight and Horsepower are the strongest predictors for MPG."

Data Preprocessing

- Handling Missing Values: Actually there were no outliers

- Feature Engineering: I have converted the float columns into integer and also cleared some of the columns which were having question marks to normal. And i have done the labelEncoding
- Data Splitting: I have splitted the data in the ratio of 8 : 2 because it avoids the overfitting, first I thought to take the data in the ratio of 5:5 but then the accuract was too low so I have taken in this way.

Model Development and Training

- Model Selection: I have tried with all the possible algorithms but the accuracy was not that good. So i have selected lasso as the regression model..
- Training Process: After doing the standardScaler we have done polynomialfeatures then we gave the 80% data for training and the remaining data for testing.

Final Model: After making a lot of tries i have selected the best model that is the lasso. It not only avoids the overfitting but also does the feature engineering.

Model Evaluation and Results

- The model's performance was evaluated on the unseen test data using two key metrics: R-squared (R^2), to measure how much of the MPG variance the model could explain, and Mean Absolute Error (MAE), to measure the average prediction error in miles per gallon.
- Results: R-squared (R^2): 0.905
  Mean Absolute Error (MAE): 1.70 MPG
- Analysis: These results are excellent. The R^2 value of 0.905 indicates that the model can explain 90.5% of the variance in fuel efficiency, demonstrating a high level of accuracy. The MAE of 1.70 means that the model's predictions are, on average, off by only 1.70 miles per gallon. This confirms the model is highly reliable and accurate for its intended purpose.

6. Challenges and Solutions

- Challenge 1: "Missing Horsepower data."
- Solution 1: , "Imputed the missing values using the median of the Horsepower column to avoid skewing the data."
- Challenge 2: "Origin was a non-numeric feature."
- Solution 2: , "Applied one-hot encoding to convert the categorical Origin (USA, Europe, Japan) into a numerical format the model could understand."
- Challenge 3: "Deploying the model on Streamlit."
- Solution 3: "Used joblib to save the trained model into a single file (model.joblib), which could then be easily loaded into the Streamlit script."

Conclusion

This project successfully developed an accurate machine learning model to predict vehicle MPG based on the "Auto MPG" dataset. After essential data preprocessing, the final model demonstrated high R-squared value, proving its reliability. This model was then successfully deployed as a practical, user-friendly Streamlit web application. Future enhancements could include training on more modern datasets and exploring advanced regression techniques to further improve accuracy.

References:

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. (Dataset: Auto MPG)
2. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp.2825-2830.
3. Streamlit Inc. (2024). Streamlit: The fastest way to build and share data apps. The primary data source for this project was the "Online Retail Data Set" from the UCI Machine Learning Repository. This dataset, originally donated by Dr. Daqing Chen, contains over 540,000 transnational records from a UK-based online retailer between 2010 and 2011. It is a widely cited and standard benchmark for tasks involving customer segmentation, cohort analysis, and predictive modeling based on RFM (Recency, Frequency, Monetary) principles.
4. Citation: Chen, D., Sain, S.L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.
5. .
6. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, pp.56-61.