Car Sales Prediction using Linear Regression

Team Members:

- G. Karthikeyan
- M. Deepak
- P. Shravya
- A. Vakula
- CH. Bhavani

The data set we used for this project, i.e., the 'Car Sales data for prediction.' from Kaggle, contained the prices, body, mileage, and other attributes over 4300 rows. The linear regression methodology was used to predict the car sales. In the dataset, we used the dependent variable - 'price' and the other attributes like 'body', 'year', 'mileage', 'engine type, 'brand,' 'enginev,' 'registration', and 'model' are considered independent variables.

Our learning outcomes from the project were as follows:

- We learned different cloud service techniques, trained the model from scratch, and verified the prediction using various parameters associated with the dataset.
- As beginners to the AWS cloud platform, we learned how to use services like AWS S3, AWS Glue, and AWS SageMaker and the importance of integrating them, which leads to greater results in less time.
- Performed visualization using graphs for the dataset and observed that the data is not linear, made few modifications such as applying log, scaling the data, and transforming the data by creating dummy variables.
- Lastly, calculated all parameters needed to know more information on prediction and how accurately the model can predict.
- Throughout the project, we learned how to use different libraries using Python and the functions that the library can provide.
- This project created a strong interest in machine learning, which is definitely a positive beginning toward learning more in the future.

The challenges we faced during this project are as follows:

- The problem in each stage is integrating each AWS service into one another; while doing this, initially, we came across connectivity issues.
- In AWS Glue, while referring to the dataset file in AWS S3, we were unable to connect it due to misconfiguration, and later, we had to debug it step by step. At last, we connected it to the Glue and imported the dataset.
- While visualizing the graphs for the dataset, generating the visual graphs with the given data set was a little confusing, but after doing the standard cleansing and validation, it was resolved.
- The interpretability of linear regression model predictions is important.

- Balancing model complexity and interpretability was challenging in dynamic and evolving environments.
- AWS SageMaker was the same as Jupyter Notebook, but using the cloud platform was a new and better experience than using it on-premises.
- Creating the AWS crawler was really good learning since we have used offline ETL tools like Informatica Power Center, but in the cloud, making connections is quite a new experience.
- Overall, we came across various challenges, but with a good understanding and a lot of research and exploring the internet, we were able to complete it till the end smoothly.