Big Data_Individual Report

In the innovative NYC TLC Taxi Fare Prediction project, my primary responsibility was the training of machine learning models using AWS SageMaker, focusing on the inbuilt XGBoost algorithm. This critical role aimed to create a reliable and efficient model for predicting taxi fares in New York City. The project, at the intersection of Big Data and machine learning, required a sophisticated approach to data handling and model training. My task was to harness the vast data available and apply advanced machine learning techniques to develop a model that could accurately estimate taxi fares, taking into account the unique dynamics of the city's taxi services.

The project presented several challenges, particularly in managing and processing the extensive dataset. The main task was to optimize the XGBoost model, balancing accuracy with computational efficiency. This required meticulous data preprocessing, including feature engineering and the elimination of redundant data, to ensure the model was trained on high-quality, relevant datasets. This part was done by one of my other teammates. Through careful tuning of the XGBoost parameters and leveraging SageMaker’s capabilities, I was able to overcome these challenges, resulting in a model that provided accurate and reliable fare predictions.

The outcome of my efforts was highly successful, as evidenced by the model achieving an RMSE value of 5.033, indicating a high level of predictive accuracy. The low error margin in fare predictions underscored the efficacy of the model, particularly in a dataset with complex variables such as unpredictable tips. This project was an invaluable opportunity for me to deepen my expertise in machine learning and Big Data processing. Working with AWS SageMaker and XGBoost, I gained significant insights into real-world applications of these technologies. My contribution not only propelled the project towards its goals but also reinforced my skills in model development and data analytics.