Milestone-2 - Literature Review - Individual Report

In my assessment of the paper [Howard18], I delved into the development of a smart transportation data pipeline, focusing on IoT data derived from the New York City Taxi & Limousine Commission. The study presented diverse experiments using various machine learning algorithms, encompassing both supervised and unsupervised approaches, to evaluate their prediction accuracy and computational performance. Notably, the paper underscored the significance of algorithm selection and infrastructure optimization, highlighting the exceptional performance of Random Forest Regressors, achieving a commendable Root Mean Squared Error (RMSE) of 0.22, comparable to that of Kaggle competition winners. Moreover, the study shed light on the unexpected success of Logistic Regression in specific instances, emphasizing its efficiency in terms of training time and accuracy. The comprehensive comparison of various machine learning algorithms, particularly Logistic Regression, Random Forest, and Gradient Boosted Tree classifiers, provides a robust foundation for researchers in the transportation data analysis domain, enabling them to make informed algorithmic and infrastructural choices.

Additionally, the research paper [Ismaeil22] offered insights into the application of a decision tree classification model for predicting payment types in NYC taxi trips. By leveraging a dataset provided by the NYC Taxi and Limousine Commission, the study achieved an impressive prediction accuracy exceeding 96%. The authors meticulously discussed diverse evaluation metrics, including accuracy, test error, weighted precision, recall, and F-measure, and examined the impact of distinct splitting criteria and node impurity measures on the model's overall performance. The paper's findings suggested the potential extension of this approach to predict other facets of taxi trips, such as passenger count and peak-hour surcharges, emphasizing its applicability in diverse transportation sectors and regions, thereby enriching the realm of real-world transportation data analysis.

I also contributed to the challenges part. Regarding the challenges overcome, our team made significant strides in tackling crucial aspects. We will successfully manage the challenge of scalability by efficiently handling the substantial volumes of data from the NYC Taxi and Limousine Commission, utilizing AWS EMR and Apache Spark for effective parallel processing. Moreover, our approach to data preprocessing, integrated with an EMR cluster running Spark, ensured streamlined data cleansing and feature engineering, enhancing the overall preparedness of our data for modeling. Furthermore, we effectively addressed the computational intensity associated with training machine learning models on extensive datasets by leveraging the scalable infrastructure of AWS Sagemaker for model training, hyperparameter optimization, and cross-validation. This strategic approach guarantees the efficient and effective handling of resource-intensive tasks, thereby enhancing the robustness and accuracy of our predictive models.