Large-Scale Data Analytics on NYC Taxi Limousine Commission(TLC) Data for Fare Prediction and Demand Forecasting

*Note: Sub-titles are not captured in Xplore and should not be used

1st Makoto Ono
Computer Science
University of Nottingham
Nottingham, UK
psxmo3@nottingham.ac.uk

2nd Raonak Shukla
Computer Science
University of Nottingham
Nottingham, UK
psxrs15@nottingham.ac.uk

3rd Helin Lei
Computer Science
University of Nottingham
Nottingham, UK
alyhl45@nottingham.ac.uk

4th Hande Simay Telatar
Computer Science
University of Nottingham
Nottingham, UK
psxht3@nottingham.ac.uk

5th Palak Bajaj
Computer Science
University of Nottingham
Nottingham, UK
psxpb13@nottingham.ac.uk

6th Kumar Harsh
Computer Science
University of Nottingham
Nottingham, UK
psxkh2@nottingham.ac.uk

***Abstract*—Urban transportation is facing unprecedented challenges: traffic congestion, pollution, and inefficient resource allocation. Taxi and ride-sharing companies generate massive amounts of data on trip patterns, driver behavior, and rider preferences. This study is an effort to provide one platform solution by leveraging these datasets to optimize urban mobility, improve service efficiency, and ultimately create a more sustainable transportation network. The intent behind the project is to empower management in data-driven decision-making. By harnessing the power of data, companies can make smarter decisions, optimize operations, and gain a competitive edge in the marketplace. Weekly demand forecasting is performed using SARIMA models, which scored the lowest on the Akaike Information Criterion (AIC) at 1390, compared to other models. The study aims to predict base fares using various regression models, employing both global and local approaches. Additionally, it will examine the performance of distributed algorithms, with a primary focus on computational power (i.e., the number of available cores) and the size of the data. The Root Mean Square Error(RSME) obtained is the least for 16 partitions (9.34), which is very close to the global model (9.10). However, the time required to train and calculate RSME is 8.21 times less than the global model.**

***Index Terms*—fare prediction, demand forecasting, global and ensemble models, size-up, scale-up, speed-up, time series modelling**

I. Introduction

Amid the rapid growth of urban transportation, conventional data analysis methods are no longer capable of efficiently handling the massive amounts of data generated. For instance, there are about 10 billion instances of hire vehicle trips in the New York City TLC dataset [1]. Also, combining other datasets like NOAA's temperature and precipitation data along with the public transport dataset for the State of New York makes the total number of features to 39 [2]. The quantum of data requires leveraging clusters of machines to process in parallel. For our study, Pyspark is being used which provides an excellent platform for data distribution and fault tolerance to ensure the reliability and robustness of big data pipelines created.

The first part of the paper will discuss the challenges in data preparation and feature selection. Next, visualization is done to extract valuable insights from the data. This will be followed by the methodology adopted for Regression and Time Series Modelling. The last section deals with regression modeling for fare prediction using global and local approach and evaluate the performance of distributed models using speed-up, size-up and scale-up as metrics.

II. Problem Description and Challenges

The first challenge is the volume of the dataset. The dataset contained more than 10.11 billion instances of trip data from 2019 to 2024, which means that it cannot fit into the driver's memory. So a technique called "column pruning" [3] is used to eliminate unnecessary columns before concatenating the parquet files making importing of data more efficient, faster and thereby reducing the amount of data movement. For cleaning redundant columns such as 'request_datetime' and 'on_scene_datetime' were removed and substituted with 'pickup_datetime'. We also used "repartition" to control the partitions of dataset so that it can optimize query performance [4]. Subsequently, NULL values in the dataset and column type mismatches were examined and the values were converted to match the appropriate formats, completing the data imputation process.

The second challenge is feature engineering and feature selection. In the original dataset from the Taxi Limousine Commission(TLC), which contained 10.11 billion instances and 25 features, only a few features were found to be related to the fare during the exploratory data analysis. Therefore, it was considered to use datasets of other public transports, including daily subway, buses and railway rides. A new feature 'pt_taxi_ratio' was calculated by taxi use as a percentage of all urban transport. Besides, after considering previous research [5], the weather data was combined with the original dataset to increase the related features including precipitation, and temperature. To account for dynamic pricing, more features including 'isRushhour', 'isOvernight', 'isWeekend' were added based on the rules of taxi fare calculation [1]. For feature selection, local approach is combined with embedded method Random Forest for feature selection. After feature selection was done for each partition, a user-defined function was used to select features by max voting. In this method, each tree of the random forest can calculate the importance of a feature according to its ability to increase the pureness of the leaves. The higher the increment in leaves' purity is, the higher the importance of the feature is.

The third challenge is the COVID period included in the data set for time series modelling. The demand plunged amid 2019 and recovered only after December 2020 due to movement constraints. The adapted solution is to filter out the stable part of the time series, first 100 weeks are dropped for demand forecasting.

A. Related work on Taxi Transport Prediction

An exemplary prior study on predicting yellow taxi demand effectively utilized a combination of linear regression and ARIMA modeling techniques. This innovative approach decreased the prediction error of linear regression compared to the typical ARIMA modeling. Significantly, the integration of these models achieved an enhancement in model performance, evidenced by an increased R-squared value, despite a reduction in the number of explanatory variables utilized. [6]

Another study delved into fare prediction by transforming a typical regression problem into a classification challenge. [7] This transformation involved discretizing the fare into four distinct expenditure quartiles, thereby simplifying the prediction process into manageable categories. The researchers employed clustering techniques to generate eight to nine distinct groups based on taxi pickup and drop-off locations. This methodological approach not only facilitated a more structured analysis but also tailored the fare prediction model to reflect geographical nuances in fare variations, enhancing the predictive accuracy for different geographical areas.

A complementary research conducted by other researchers, which is the study on spatial prices and search frictions, suggests that spatial pricing based on originating points and enhancing search efficiency can increase consumer surplus of taxi fares without adversely affecting the profits of drivers [8].

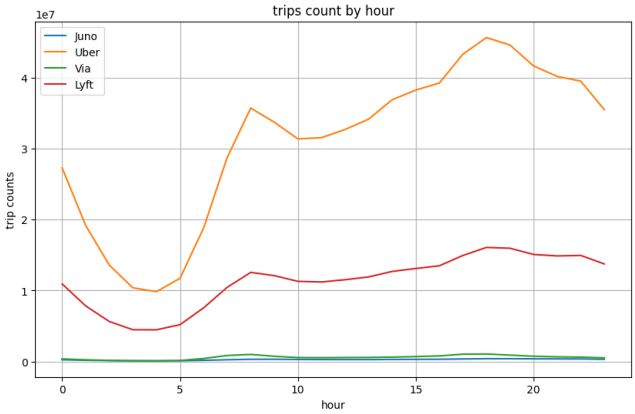

Fig. 1. Demand Variation by hour.

These previous studies gave a deeper insight into taxi demand and fare prediction. Together, these studies underpin the importance of tailored, context-specific models in enhancing the efficiency and profitability of taxi services.

III. Data Preprocessing and Visualization

a) Data Preprocessing:

The dataset contained a different set of parquet files, which includes Yellow Taxi trip dataset and High Volume For-Hire Vehicle (HVFHV) trip dataset. After removing redundant columns such as 'request_datetime' and 'on_scene_datetime' and substituting with 'pickup_datetime', NULL values in the dataset and column type mismatches were examined and the values were converted to match the appropriate formats. Also, percentage comparison columns against a comparable pre-pandemic date are considered redundant and removed from MTA Daily Ridership Data, completing the data imputation process.

In the initial exploratory data analysis and data visualization, we introduced new columns in the taxi dataset by aggregating rows daily and hourly, which include 'is_Weekend' (New York State public holidays included), 'is_Rushhour' (4-8pm on weekdays), 'is_Overnight' (8pm-6am), which represents the category of the time in which the instance recorded respectively, and 'TMAXExtremity', 'TMINExtremity' in the weather dataset, each of which represents the deviation of daily maximum and minimum temperature from the ideal temperature (set at 68°F). The formula for the temperature extremity columns is $E = |t - 68|$.

After processing each table, left join was performed on calendar and hour columns. However, in the process of examining the correlations between each column, little correlation between a few columns was observed, hence they were dropped in the model-building process. To further create a line chart by each ride-share company, the instances on the HVFHV dataset are aggregated based on four different affiliated ride-share companies (i.e., Uber, Lyft, Via, Juno).

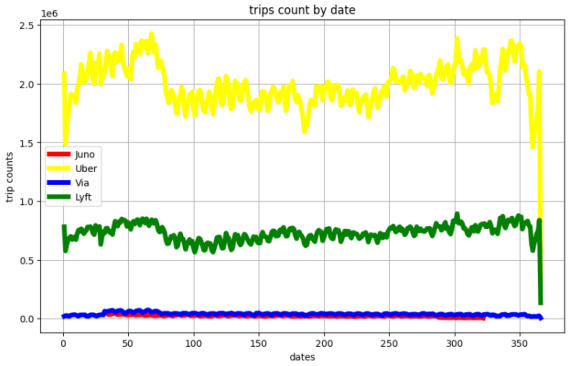

Fig. 2. Demand Variation by Week

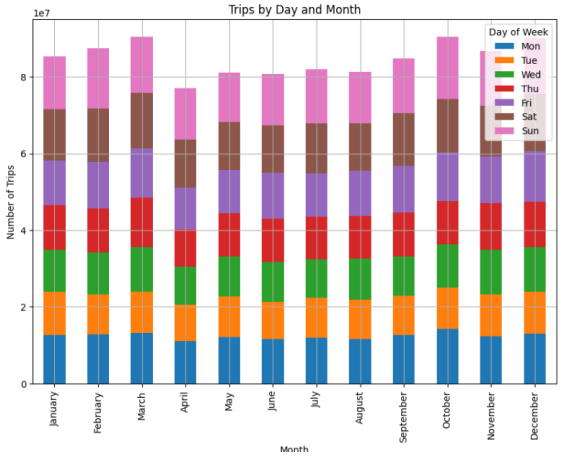

Fig. 3. Demand Variation by Week and Months

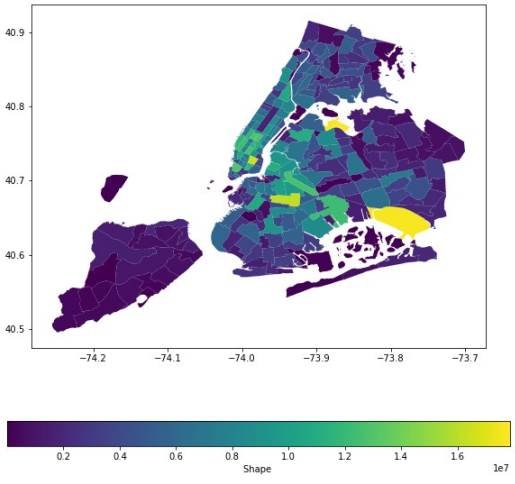

Fig. 4. Taxi Pickup Heatmap for State of New York

b) Data Visualization:

From "Fig. 1" we can see there in higher demand in between 4:00 pm to 08:00 pm and it falls in between 12:00am to 06:00am. Our hypothesis is that with increase in demand the cost of ride will also increase. To make price prediction more dynamic we have incorporated two new features in the dataset 'is_Rushhour', 'is_Overnight' in our analysis.

The Line chart segmented by date "Fig. 2" reveals that the trip counts maintained around 2 million daily for Uber. Via and Juno control only a small portion of the ride share industry in Newyork. A marginal increase was observed during the winter months. This motivated us to include temperature and precipitation data from NOAA.

"Fig. 3" provides insights into the distribution of trip counts by month and day of the week, showing some variation from Monday to Friday. Nonetheless, trip counts from April to September were slightly lower than during other months. This motivated us to create new feature, such as 'is_weekend'.

Last section deals with zone wise segmentation of demand, for this heat map is being used. The purpose of creating a heatmap "Fig. 4" is to identify zones with the highest number of pickups and drop-offs during working days, weekends, rush hour, and nights. Such heatmaps can help higher management to have a bird eye view of the source of demand and can help in better allocation of resources. It was found that LaGuardia Airport, JFK Airport, East Village, and Crown Heights North have the highest number of pickups in total, while Liberty Island, Riker Island, and Jamaican Bay have the lowest demand over five years. During rush hours, the highest demand comes from East Village, JFK Airport, Lower East Side, Times Sq/Theatre District, which are also some of the prominent business centers in the State of New York. However, over the weekends, the highest demand comes from East Village, Crown Heights North, JFK Airport, and Lower East Side. Airport dominates in terms of demand in all the categories because, according to National Car Ownership Statistics 91.7% of households had at least one vehicle in 2022. Only 8.3% of households did not have vehicles.

IV. Methodology

a) Regression:

Initially, a local regression model was implemented for fare prediction using linear regression and decision trees. The training dataset was distributed into partitions, applying regression locally at each chunk of the dataset. Then the results from each partition are averaged. Also, the number of partitions was increased gradually in each iteration to analyze the ensemble effect. After that, the global model was trained and tested for three algorithms linear regression, decision tree, and Gradient Boost Trees. All tree models will be compared on two metrics time and prediction error (RMSE). Finally, the best model from the Local approach will be used to conduct experiments for size-up, speed-up and scale-up on the yellow taxi dataset. [9]

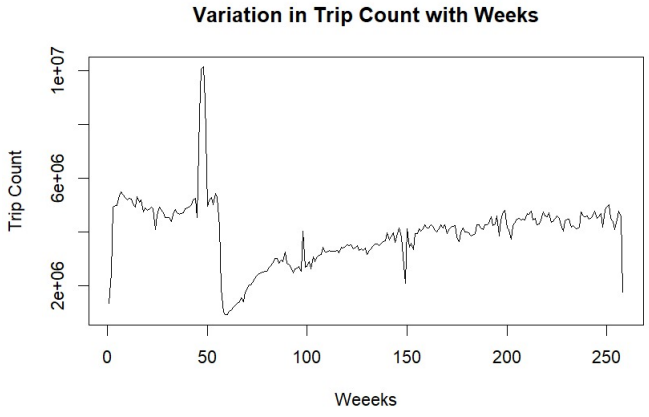

Fig. 5. Variation in trip count with weeks.

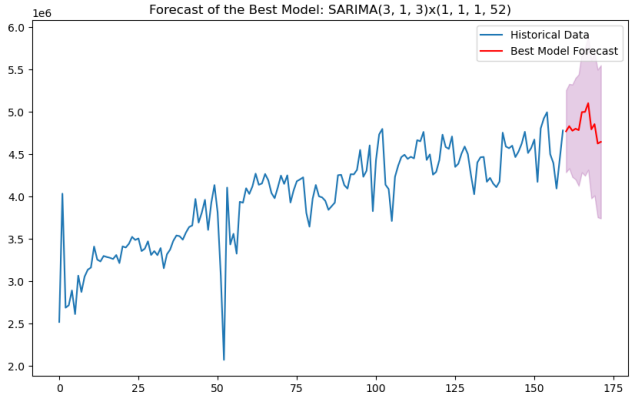

Fig. 6. Time Series Forecast for Weekly Trip Count.

b) Time Series Modelling:

Detailed time series analysis was performed on the dataset using ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal Auto Regressive Integrated Moving Average) models to predict future values effectively. The aim is to determine the model that best suits the data set by examining the performance, diagnostic checks, and statistical significance of these models.

From the time series plot(Figure 5), it can be observed that there is an effect of mobility restriction due to the COVID lockdown. The first 100 weeks are removed for our analysis to have a stable time series.

The original data is first subjected to preliminary stationarity testing using the Augmented Dickey-Fuller (ADF) test. This test is a statistical test used to determine whether a time series is stationary. It is important to ensure that the data does not have time-dependent structures such as trends or seasonality. ADF statistic is -2.191 and p value is 0.209; This indicates that the series is non-stationary as the p value is above the 0.05 threshold and transformations such as differencing may be needed to make the data stationary.

Moving on to model analysis, various configurations of ARIMA, MA, SARIMA and Exponential Smoothing are applied. The Akaike Information Criterion (AIC) values of each model are summed to compare their relative performance. Among the models, the SARIMA model has the lowest AIC, suggesting that it may be the most appropriate model for this data set among those tested. However, since the AIC value is higher than normal and due to the necessity of taking differences, more models are developed. [10].

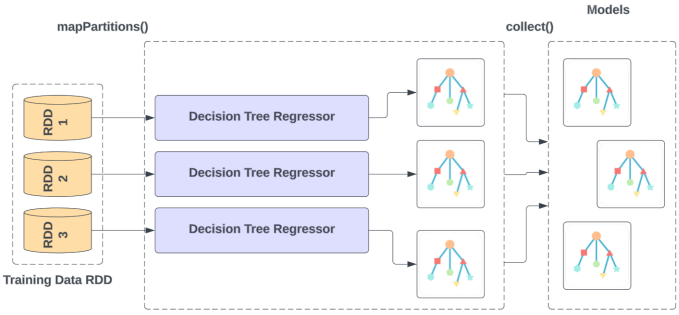

Fig. 7. Local Ensemble Model For Decision Tree Regressor.

Looking at the Model Efficiency Ratios, all models show ratios very close to 1, indicating that their AICs are competitive, but minor differences make the SARIMA model advantageous over others. Based on this information, it is understood that the models generally perform well due to the stationarity of the residuals and the absence of autocorrelation. But there are still points that can be improved. Two things come into play here: Residual Analysis and Overfitting Control.

SARIMA$(3,1,3)(1,1,1)_{52}$ model stands out as the best option, with 1390.2 AIC value, to predict this time series data based on the given metrics. It effectively handles complexity, data patterns, and seasonality in terms of higher orders in both autoregressive and moving average components (both seasonal and non-seasonal). Mathematically model can be described as:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-B)(1-\Phi_1 B^{52})X_t =$$

$$(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)(1+\Theta_1 B^{52})Z_t(1)$$

Here, X_t represents observed value at time t, Z_t represents white noise term at time t. Moreover, selecting "52" as the seasonal period in SARIMA model configurations is of great importance in effectively analysing and forecasting weekly time series data showing annual patterns. At the end, a 12-week forecast is created based on the SARIMA$(3,1,3)(1,1,1)_{52}$ model (Figure 6)

V. Experiments and Results

a) Local Approach and Ensemble Effect:

The whole dataset is partitioned into resilient distributed datasets (RDDs) (Figure 7) in the local ensemble approach, and a separate model is trained for each partition. The idea here is to distribute the training process among the workers through parallelization, in this way each worker has to deal with less data and it will also reduce runtime and memory constraints. After training the models, a user-defined function

is used to aggregate the results from each model. The metric used to evaluate the prediction is RSME. The hypothesis for experimenting is to check whether the ensemble effect; many weak learners obtained due to partitions can produce results equivalent to a global model, as depicted in the following equation:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x) \tag{2}$$

Here:

- B is the number of partitions

From "Fig. 8", it can be seen that the RMSE error decreases as the number of partitions increases, but later on, it increases as each model will have less data to train on. The ensemble effect is visible because: each decision tree is grown in depth by recursive partitioning of the feature space in order to reduce the bias, and aggregating the results for each partition reduces the total variance in predictions [11]. The RMSE obtained is least for 16 partitions (9.34), which is very close to the global model (9.10). However, time required to train and calculate RMSE is 8.21 times less than the global model. This proves our hypothesis that, due to ensemble effect, the approach of divide and conquer can help get results close to the global model faster since distribution and parallelisation reduce processing time. If the trade off can be made between accuracy and time required to run, this approach improves scalability.

b) Global Approach Vs Local Approach:

In this experiment, the performance of the globally distributed model is compared with the locally distributed model. We hypothesize that a local model will outperform the global model in terms of time required to obtain results as calculations happen on a small cluster with high-speed connections and also it minimizes data transfer between machines, which is a major bottleneck in distributed processing. For the global approach, three models were selected for fare prediction Linear Regression, Decision Trees, and Gradient Boost Trees. The RMSE for the models were 9.01, 9.10, and 10.42 respectively. The globally distributed model was slower than the local one as data must be transferred between machines in the cluster for processing and aggregation. This network communication adds latency and significantly impacts the performance, especially large datasets like ours. However, it was observed that the results obtained from a global model were robust and not affected by the partitioning of the dataset [**?**].

c) Speed-up, Size-up and Scale-up:

As is shown in "Fig. 9", we experimented with speed-up test on local-based linear regression from 1 core to 12 cores. The result of runtime from 1 core to 4 cores is stable. The sub-linear increase from 1 core to 2 cores is partly because of overhead since the progression in speed-up without overhead is much faster. When it comes to the sub-linear increase from

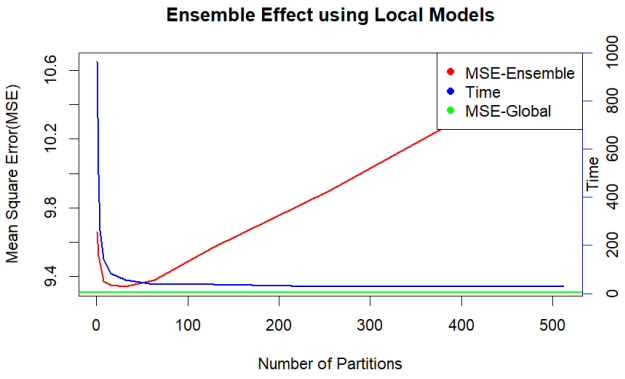

Fig. 8. Ensemble Effect.

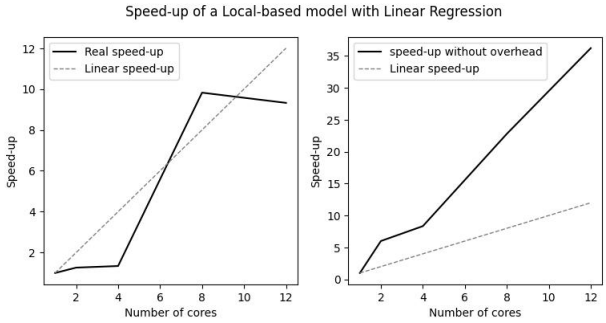

Fig. 9. Speed-up Test on Local model.

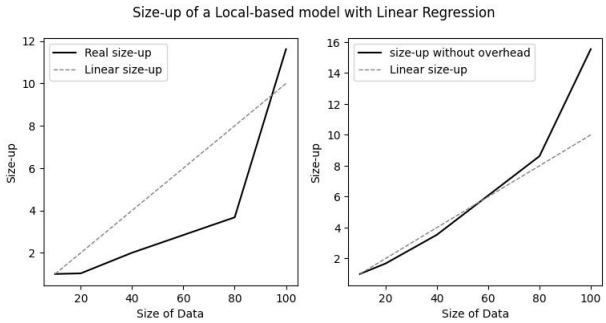

Fig. 10. Size-up Test on Local model.

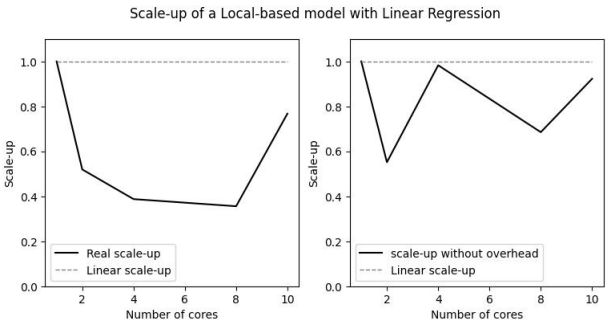

Fig. 11. Scale-up Test on Local model.

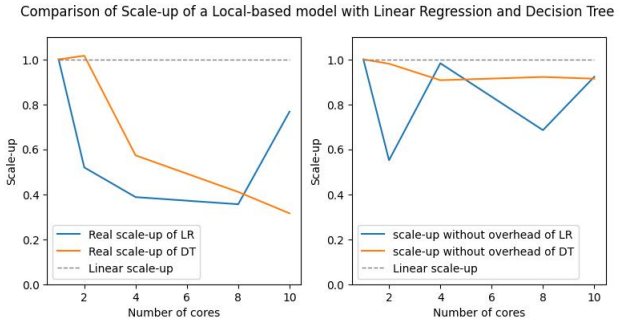

Fig. 12. Comparison of scale-up of a local-based model with Linear Regression and Decision Tree.

2 cores to 4 cores, it may stem from communication delays or poor load balancing among processors. Nevertheless, the runtime from 4 cores to 8 cores is more than 10 times as fast. One of the possible reasons for super-linear speedup is that it reduces data movement by lessening RAM usage. The processors keep more data in the fast memory so that it takes less time [13]. After that, the runtime becomes slower from 8 cores to 12 cores. It is likely that the overhead of managing additional processors outweighs their computational benefits.

Regarding the speed-up test without overhead on linear regression, the runtime from 1 core to 2 cores is more than five times faster than the baseline. The progression from 2 cores to 4 cores are much slower. However, it soared from 4 cores to 12 cores.

The second study will try to analyze the effect of size-up. Here in "Fig. 10", the number of cores is kept constant i.e. 8 cores. The whole dataset is partitioned into 10%, 20%, 40%, 80%, and 100%, and linear regression is used for this experiment. It can be inferred from the plot that as the size of the data increased from 10% to 100%, the time required to run increased to 11.62 times and 15.54 times with and without overhead, respectively. Ideally, the resources required to run the model should increase linearly with an increase in the size of the data.The sub-linear increase from 10% to 80% in test can be attributed to overhead; The super-linear increase from 80% to 100% is partly because of two reasons. The first reason is that the resource availability was constrained as a lot of people were working simultaneously on the same cluster, and the plots were obtained with only three iterations. Another possible reason is that the memory limitations. [12] It is observed that there is a huge fluctuation between iterations.

The last study deals with scale-up where the size of data and the number of cores are increased proportionally to ensure that each partition deals with the same amount of data. The assumption made is that there is no communication overload within the cluster. The number of cores is increased gradually from 1 to 10 and the size of data from 10% to 100%. From "Fig. 11", it can be conferred that there is a large deviation from the ideal curve. Possible reasons can be with an increase in the number of cores, the communication overhead between nodes increases significantly, which limits the benefit of additional processing power [10]. Also, Outlab Cluster and Cluster 2 are larger clusters with more than 20 nodes, thereby introducing additional complexity in terms of distributing tasks, handling failures, and ensuring overall job coordination.

VI. Conclusion

In this paper, we have applied a big data method for taxi fare and demand prediction, using various regression models, SARIMA and ARIMA. The results have demonstrated the accuracy, effectiveness, and stability of our prediction. Since the data for analysis is increasing day by day with the advancement of data collection and data storage technologies it is better to go for algorithms that can be scaled up. In our study, we have tried to compare scaling-up options between linear regression and decision tree. From 'fig 12' it can be seen that non-linear model like decision trees provide better scalability. In terms of future work, we will consider more advanced algorithms like neural networks as they are better in capturing the non-linear relationship between the features.

References

[1] "TLC Trip Record Data," New York City Taxi and Limousine Commission, 2024. [Online]. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. [Accessed: 2-May-l 2024] .

[2] "NOAA Temperature and Perception,"NOAA , 2024. [Online]. Available:https://www.noaa.gov/weather . [Accessed: 2-May-l 2024] .

[3] M. H. Zhu and S. Gupta, "To prune, or not to prune: exploring the efficacy of pruning for model compression," arXiv preprint arXiv:1710.01878, Nov. 2017. [Online]. Available: https://arxiv.org/abs/1710.01878.

[4] G. J. Harshit and N. G., "Implementation and Performance Comparison of Partitioning Techniques in Apache Spark," in Proc. 10th Int. Conf. on Communication Systems & Networks (COMSNETS), Bangalore, India, 2019.

[5] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," Transportation Research Part C: Emerging Technologies, vol. 68, pp. 285–299, Apr. 2016, doi: 10.1016/j.trc.2016.04.005.

[6] S. Faghih, A. Shah, Z. Wang, A. Safikhani, and C. Kamg, "Taxi and Mobility: Modeling Taxi Demand Using ARMA and Linear Regression," in Procedia Computer Science, vol. 177, New York, NY, USA, 2020, pp. 186-195. doi: 10.1016/j.procs.2020.10.027.

[7] B. Rathore, P. Sengupta, B. Biswas, and A. Kumar, "Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models," Transportation Research Part E, vol. 185, p. 103530, Apr. 2024. Available: https://doi.org/10.1016/j.tre.2024.103530.

[8] B. Ata, N. Barjesteh, and S. Kumar, "Spatial Pricing: An Empirical Analysis of Taxi Rides in New York City," in Proc. of the Conference on Taxi Market Analysis, New York, NY, USA, 2022, pp. 1-12.

[9] I. Triguero and M. Galar, Large-Scale Data Analytics with Python and Spark. Cambridge, United Kingdom: Cambridge University Press, 2024.

[10] City University of New York (CUNY) CUNY Academic Works Dissertations and Theses City College of New York 2019 Understanding and Modeling Taxi Demand Using Time Series Models Sabiheh Faghih CUNY City College.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 1st ed. New York, NY: Springer, 2013.

[12] J. Zhang, Z. Yang, and Y. Benslimane, "Exploring and Evaluating the Scalability and Efficiency of Apache Spark Using Educational Datasets," School of Information Technology, York University, Toronto, Canada, 2024.

[13] V. Kumar and A. Gupta, "Analysis of Scalability of Parallel Algorithms and Architectures: A Survey," Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 1991.

[14] Kelly Anne.S,Ashlee.V “Car Ownership Statistics 2024” Forbes Advisor, Mar 28, 2024. [Online]. Available:https://www.forbes.com/advisor/car-insurance/car-ownership-statistics/#:~:text=Sources-,National%20Car%20Ownership%20Statistics%20at%20a%20Glance,98%2C573%2C935%20vehicle%20registrations%20for%20cars. [Accessed: 13-May-2024].