Student Name: Fahad Mulla

Professor: Dr. Krishna Bathula

Course: Analytics Capstone Project

11 December 2024

NYC Yellow Taxi Demand Prediction

My capstone project deals with the New York City's yellow taxi's demand prediction based on various demographic features by making use of machine learning and deep learning models and techniques to fulfill the supply and demand problems faced by New York city yellow taxi service providers. New York City Taxi and Limousine Commission (TLC) is the only New York's government agency authorized to provide licenses to private cab companies for street-hailing and pre-arranged services. Moreover, companies that are specifically online booking based such as Uber and Lyft are also licensed by New York's TLC agency, but are not authorized for street-hailing. Loss of business and inconvenience to the people travelling in New York city could be solved by providing real time insights to the drivers. This project makes use of the official data acquired by Taxi and Limousine commission of New York.

Introduction

New York City is said to have the most taxis across the U.S.A and is also titled a taxi capital of America. These taxis are operated by private cab companies which are licensed by the taxi and limousine commission (TLC) agency of New York. The cab companies lease their cars to the drivers and drivers get to keep 100% of the fares and tips. Fares for these cabs and amount for leasing the car to the drivers are set by TLC to avoid unfair practices and exploitation. There's another scheme offered by private cab companies to drivers where the cars are leased for lower prices, where a portion of the fare goes the cab company. These Yellow Cabs are the only

taxi services that are authorized for street-hailing. Any unauthorized street-hailing services offered by anyone without the license is subjected to a penalty of $2000 and 60 days in jail and license and registration gets revoked for 60 days. Since there are heavy penalties and restrictions on street-hailing there's no trouble from private car owners to offer street-hailing services. The major concern for my project is how is the yellow taxi cabs revenue and business gets affected? Here comes into picture the online taxi services such as Uber and Lyft. These two taxi services also need licenses to run their taxis in New York city. But they are only restricted to online taxi services. Since these companies run their services online, they have been capturing the taxi services business much more than yellow taxi services. Due to the easy of availability of booking a cab over any smartphone, it has become the most popular means of taxi commute for the commuters.

Here I'm trying to solve this major problem faced by yellow taxi cab services which is lack of availability at the required location and times of the day. If the yellow taxis are spread across New York city evenly and not get concentrated at only a single zone of the city, it would be much convenient for the commuters to make use of this cab services. Having a proper supply for the rise in demand of the taxi services at most needed regions is the solution the to my problem statement.

By making use of my machine learning techniques and selecting the right model can help solve this demand supply problem. I have to choose a model that fits right for my dataset and gives me the desired predicted output. Here, I'm expecting to get a predicted region demand beforehand so that the drivers could get to the locations of high requirement zones to be available for the commuters. Using the attributes in the dataset like pickup, drop-off time, passenger count,

trip distance, pickup and drop off location ids are used to feed the model to get the predicted demand requirement output.

Apart from yellow taxi cabs drivers and private cab companies, this work can be used by online taxi service providers like Uber, Lyft to predict the most demanding location beforehand. This can also be used by government agencies for traffic control and maintenance. Moreover, this work could also be utilized by business owners to get an insight on which time and day of the week they could expect a greater number of people going around their business to attract them by giving out discount and various marketing strategies to boost their business revenues. Logistic companies to minimize the delivery timings and schedule accordingly. Researchers and data scientists studying time-series forecasting or urban transportation patterns. This could lead to minimizing the carbon emissions, lower the wait times and higher profits for the drivers.

There are various GitHub, Kaggle and other repositories and resources available across the internet where people have worked on this taxi demand prediction in New York City on NYC taxi and limousine services dataset. These works include the use of time series models like ARIMA, Linear Regression, XGBoost, LSTM, and Random Forest Regression. Moreover there are various papers on demand prediction of Yellow taxi of New York City, of which few are mentioned in the literature survey of this section. My work of using the LSTM model is also done previously which did showcase promising results in forecasting the predicted rides with better accuracy than the conventional models. Some of which are experimented with spatial-temporal models by using location based features, but also has challenges with big datasets and factors such as weather and holidays. My approach to the model training makes use of latest NYC taxi trip dataset, which is of 2024, which ensures that it is related to current mobility landscape. Moreover the use of Hybrid Architecture and Bidirection LSTM’s makes it different

from previous works where the demand prediction would completely rely on either CNN's or LSTM's. Here we know that CNN's are best for extracting spatial or temporal patterns, while LSTM's are good at handling sequential dependences. And the bidirectional LSTM captures relationships from both past and future sequences, enhancing prediction accuracy. With this changes in works from the earlier works I expect better accuracy for such a big dataset , faster convergence and applicable to the real-time situation.

Literature Review

In 2017, Spatio-Temporal Modeling with STAR models demonstrated a significant advancement in capturing the dynamic behavior of taxi services. By employing LASSO-type penalization, these models effectively handled high-dimensional datasets, surpassing the predictive accuracy of traditional Vector Autoregressive models. Building upon this success, a 2019 study introduced Deep Survival Analysis for Dispersal Events, a novel two-stage framework that integrated deep learning techniques with survival analysis. This innovative approach proved highly effective in predicting urban dispersal events, particularly when driven by high taxi demand. Notably, this model exhibited superior forecasting performance compared to existing deep learning methodologies, showcasing the potential of this combined approach for improving urban transportation planning and management.

A 2020 study investigating the "Impact of COVID-19 on Taxi Demand" in New York City revealed significant shifts in the spatiotemporal patterns of yellow taxi demand following the pandemic. These findings underscored the profound impact of COVID-19 on the yellow taxi cab business, highlighting the need for adaptive strategies to navigate the changing urban mobility landscape. Building upon this understanding, researchers in 2023 developed "Hybrid Models for Demand and Fare Prediction," a platform integrating Long Short-Term Memory

(LSTM) networks and Mixture Density Networks (MDN). This innovative approach aimed to enhance the user experience within the urban transportation system by providing more accurate and reliable predictions of taxi demand and fare rates. By leveraging the power of these advanced machine learning models, this research strives to improve the efficiency and accessibility of urban transportation services for both passengers and drivers.

The "Machine Learning Case Study on Demand Prediction" is a publicly available GitHub project that explores the application of machine learning techniques to forecast taxi demand. This project mentions the crucial role of feature engineering and model selection in providing better prediction accuracy. By selecting and transforming relevant data features, researchers can significantly enhance the performance of their chosen machine learning models. Complementing this work, the "Real-Time Demand Prediction Model" project demonstrates the development of a real-time prediction pipeline for taxi demand. This innovative approach lets streaming data to continuously update demand forecasts, enabling more dynamic and responsive decision-making. With a combination of regression and neural network models, this project aims to provide accurate and up-to-the-minute insights into fluctuating taxi demand, thereby improving the efficiency and responsiveness of urban transportation systems.

A Master's thesis delved into the realm of spatial and temporal demand forecasting, employing supervised learning techniques to achieve this goal. The study leveraged both Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) networks, enabling a comparative analysis of classical and advanced machine learning methods for predicting demand accuracy. In parallel, another research endeavor focused on "Time Series Regression for Pickup Prediction." This work utilized time series data on the number of pickups within specific regions of New York City, employing a combination of linear and non-linear regression models

alongside spatial clustering techniques. By integrating these methodologies, the researchers aimed to generate region-specific pickup predictions, offering valuable insights into the nuanced demand patterns across different areas of the city.

A comprehensive "Fare and Duration Analysis" of taxi trip datasets explored the predictive power of machine learning algorithms, including decision trees and neural networks, to forecast both trip fares and durations. This analysis demonstrated a fascinating cross-applicability of these models, highlighting their potential to effectively address both demand and trip dynamics within the urban transportation system. In parallel, a 2011 study, "Spatio-Temporal Clustering for Taxi Demand," delved into the identification of high-demand zones within the city. By employing k-means clustering and density-based clustering algorithms, this research effectively categorized regions with significant taxi demand, revealing crucial hotspots for transportation services. These findings provide valuable insights for drivers, enabling them to strategically allocate their resources and optimize their service delivery to better meet passenger needs and improve overall operational efficiency.

In 2013, researchers employed "ARIMA Models for Short-Term Prediction," focusing on time-series modeling and parameter tuning to effectively capture temporal patterns in demand. The simplicity of ARIMA models made them a valuable baseline for establishing initial demand forecasts. Recognizing the limitations of linear models in capturing complex demand behaviors, a 2017 study introduced "Neural Networks for Nonlinear Demand Patterns." This research leveraged the power of Artificial Neural Networks (ANNs) to effectively model nonlinearities within demand data. By incorporating weather and event-related features as input variables, the researchers designed deep neural networks capable of producing more accurate forecasts in scenarios characterized by intricate and dynamic demand fluctuations. This research

demonstrated the significant potential of neural networks in enhancing the accuracy and robustness of demand forecasting systems.

In 2017, Conditional Random Fields (CRFs) emerged as a powerful tool for capturing interdependencies within taxi demand across both space and time. By increasing the strengths of CRFs researchers were able to enhance their ability to model complex spatiotemporal relationships, leading to more accurate and informative demand forecasts. In 2009, the "Dynamic STARIMA for Traffic Flow Forecasting" study introduced a novel approach that integrated dynamic STARIMA models with spatial autoregressive techniques to effectively predict traffic flow. This innovative framework successfully incorporated temporal variability by incorporating time-varying coefficients into the model, resulting in improved forecasting accuracy. Furthermore, the 2014 study on "Hierarchical Vector Autoregression" pioneered the use of Hierarchical VAR models to address the challenges of multi-region demand forecasting. By employing a layered modeling approach, this research effectively captured interdependencies across the various boroughs of New York City, providing a robust solution for overcoming the complexities associated with high-dimensional forecasting problems.

In 2017, the "Gaussian Conditional Random Fields" study introduced a significant advancement by incorporating Gaussian assumptions into the traditional CRF framework. This innovative approach, known as GCRF, resulted in smoother and more refined predictions, particularly for continuous demand variables. This improvement in precision enhanced the practical applicability of CRFs for real-world scenarios. Building upon this foundation, we can trace the roots of spatiotemporal modeling back to the 1980s, when Pfeifer and Deutsch introduced the seminal "Spatio-Temporal Autoregressive Models" (STAR models). These pioneering models, which focused on capturing both spatial and temporal interdependencies

within data, laid the groundwork for subsequent research in this field. The STAR models represent a foundational methodology that has profoundly influenced the development and refinement of more sophisticated spatiotemporal modeling techniques employed in contemporary research.

A 2010 study on "Taxi Demand Hotspot Prediction" lead to the development of context-aware methods for identifying areas of high demand. By including external factors such as weather conditions, special events, and time-of-day information, this research enhanced the accuracy of hotspot predictions. These insights proved of no value for optimizing taxi dispatch strategies, allowing for more efficient resource allocation and improved service delivery. Building upon these advancements, a 2016 study, "Spatio-Temporal Kriging for Demand Estimation," introduced a novel geostatistical approach to demand forecasting. This innovative method effectively leveraged spatial correlations and historical data to generate accurate demand estimates. By including spatial autocorrelation into the forecasting process, Spatio-Temporal Kriging demonstrated a significant improvement in the accuracy and reliability of demand predictions, providing valuable insights for urban transportation planning and management.

Formula for Data Aggregation:

$$D(t,z) = \sum_{i=1}^{n} x_i$$

Where D(t,z) is the aggregated demand at time t for zone z, and x_i represents individual taxi pickups.

LSTM Network: Captures temporal dependencies across hourly/daily intervals.

- Hidden state h is updated using:

$$h_t = \sigma(W_h.[h_{t-1}, x_t] + b_h)$$

Where x_t is the input at time t, and W_h and b_h are weight and bias matrices

CNN Layer: Learns spatial patterns from different geographic zones.

- Convolution operation:

$$y = f(W * X + b)$$

Where W is the filter, X is the input feature map, b is the bias, and f is the activation function.

Transformer Layer: Enhances long-range temporal dependencies.

- Uses multi-head self-attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q, K, and V are query, key , and value matrices, and d_k is the scaling factor.

Algorithm Pseudo-code

1. Import Necessary Libraries
2. Load and Preprocess Dataset
3. Apply Clustering
4. Exploratory Data Analysis (EDA)
5. Prepare Data for LSTM-CNN Model
6. Define Advanced LSTM-CNN Model
7. Train the Model

8. Evaluate and Visualize Results

Methodology

My Capstone project for NYC Yellow taxi demand prediction deals with predicting the future demand of yellow taxi in the different regions of New York City based on the historical values. The methodology includes data preprocessing, exploratory data analysis, clustering, time-series forecasting, and use of Hybrid LSTM-CNN model. The novelty her is combining the spatio-temporal clustering with advanced sequence modelling for granular demand prediction. In the data preprocessing part, I have cleaned the data to remove anomalies and ensure high quality input for analysis and modeling. Here temporal features like trip duration and pickup, dropoff time and time bins are utilized. I have removed the irrelevant features from the data such as tips, mta tax, congestion charges, vendor Ids, etc… Moreover null values are removed. Passenger count with 0 or more than 6 are removed because a taxi with 0 passengers is not a taxi. Trip distance more than 35 miles are also removed because the NYC region we are woking on is within that range of miles. I have worked on the outliers part in the dataset. I have removed the negative and zeros in the distance column.

Figure 3 provides a visual representation of the demand heatmap across different clusters and time bins. A careful examination of this heatmap reveals distinct patterns, with certain clusters exhibiting significantly higher demand during specific hours and in particular regions. These observed variations in demand intensity across both time and space strongly suggest the presence of pronounced temporal and spatial patterns within the data. Recognizing and understanding these inherent patterns is crucial for developing accurate and effective demand prediction models. By incorporating these insights into our forecasting algorithms, we can

significantly improve the precision and reliability of our predictions, ultimately leading to more efficient and responsive urban transportation systems.

In the model development I have made use of hybrid LSTM-CNN deep learning model to record spatio-temporal dependencies in demand data. Here the temporal patterns are learned using LSTM layers, whereas spatial dependencies are recorded using Conv1D Layers. I trained 200 epochs with patience 50 in which around 83 epoch only processed as the losses didn't changes. In the Evaluation and Visualization part, the performance is evaluated using Root mean squared error and mean absolute percentage error. A chart of predicted vs true demand patterns is visualized to see the models accuracy.

Results and Analysis

The result from my work shows different trends in the trips and predicted demands in yellow taxi. I have performed various EDAs on the trip dataset to analysis the demands of taxis during various hours of the day and day of the week. I have noticed from the analysis in Fig. 1 the maximum distances of trips are for 2 to 3 miles of distance. People choose to travel smaller distance in cabs making it the most used for that distance. Mostly the longer distances are covered using public transport or it could be that the people residing are populated only in that radius of distance.

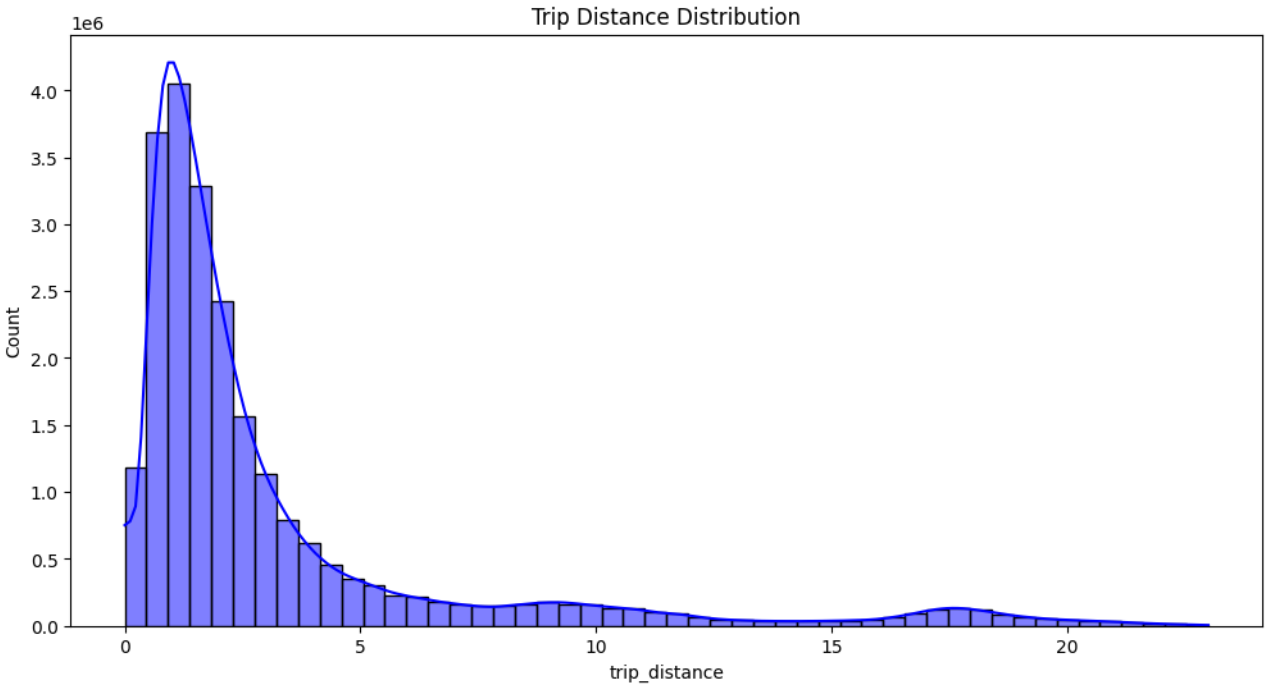

Fig. 1. Trip Distance Distribution

In Fig. 2 shows the hourly demand distribution. In this analysis I notice the most requirement of the taxi is during the late afternoons and evening hour of the day, and with a noticeable dip in the morning. I could derive from this graph that this aligns with usual human activity where a person goes to work during the work hours and other such as leisure hours. My model might predict these peak hours thereby making sure there's an availability of taxis during these demanding hours.

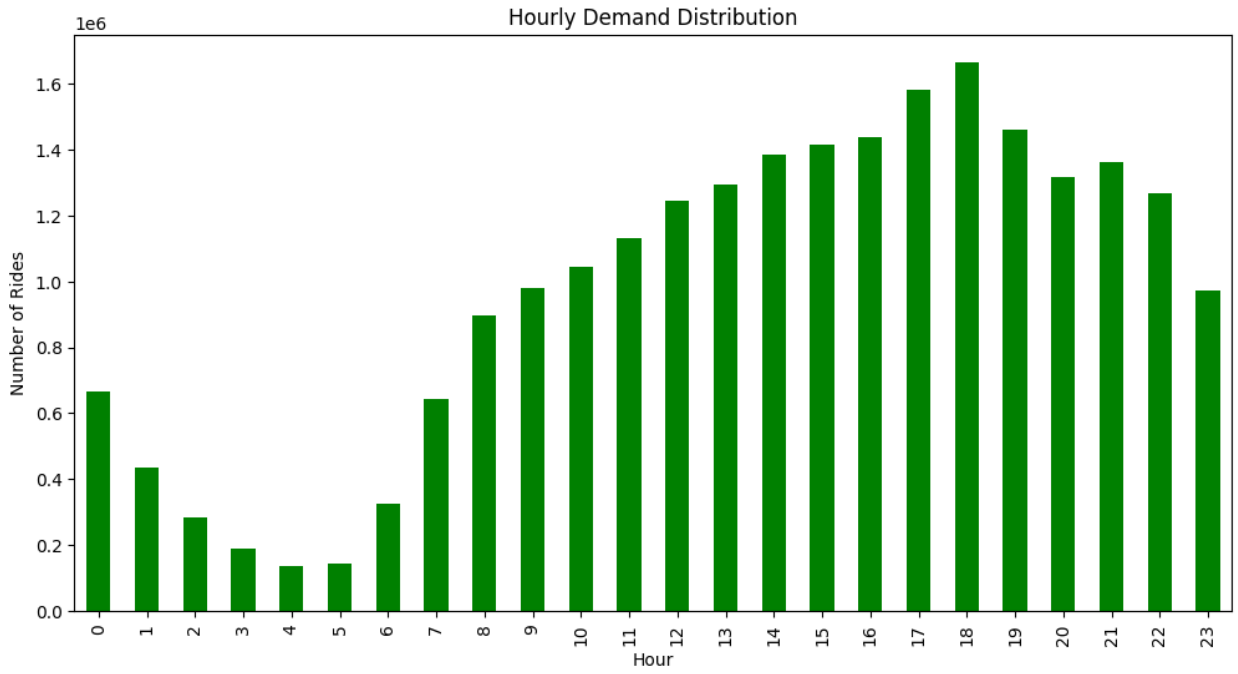

Fig. 2. Hourly Demand Distribution

Figure 3 provides a visual representation of the demand heatmap across different clusters and time bins. A careful examination of this heatmap reveals distinct patterns, with certain clusters exhibiting significantly higher demand during specific hours and in particular regions. These observed variations in demand intensity across both time and space strongly suggest the presence of pronounced temporal and spatial patterns within the data. Recognizing and understanding these inherent patterns is crucial for developing accurate and effective demand prediction models. By incorporating these insights into our forecasting algorithms, we can significantly improve the precision and reliability of our predictions, ultimately leading to more efficient and responsive urban transportation systems.

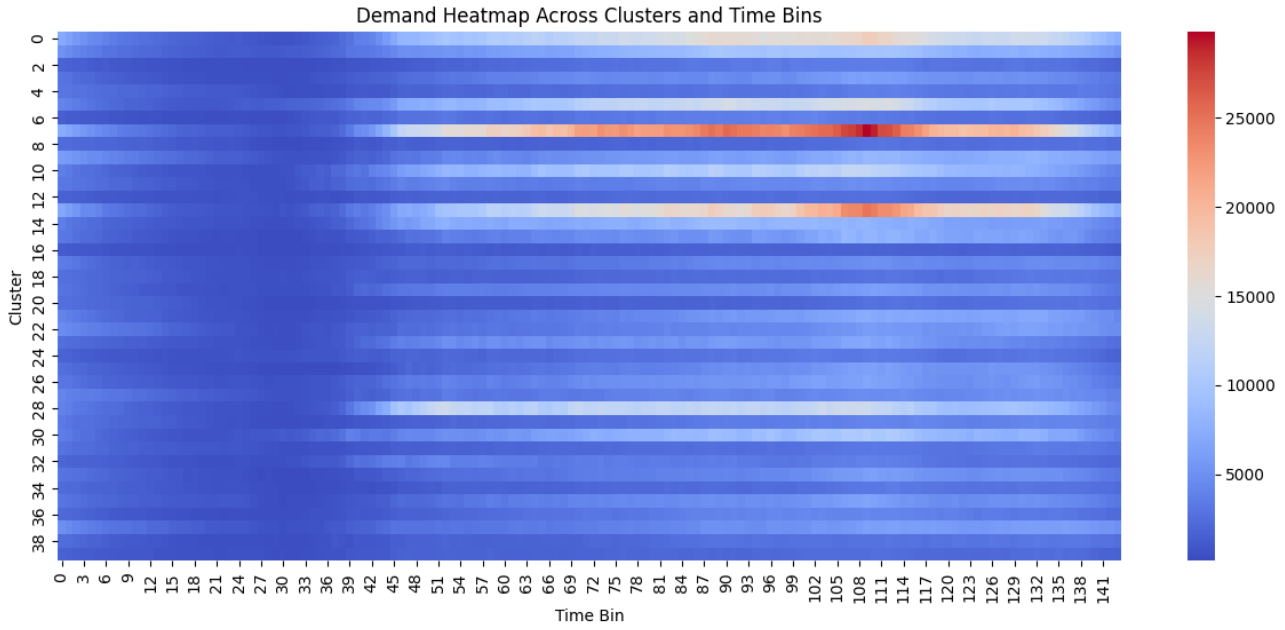

Fig. 3. Demand Heatmap across clusters and time bins

In Fig. 4 we can see the chart showing the number of rides utilized for each cluster. Clusters are the pickup and drop off locations of New York city. I'm here working with 260 clusters in the dataset. Front the chart we come to an analysis that the cluster 7 and 13 dominate the demand, while the others contribute minimally. The imbalance seen in the clusters suggests that there's a need for cluster specific models or adjustments has to be made to better capture demand patterns in underrepresented clusters

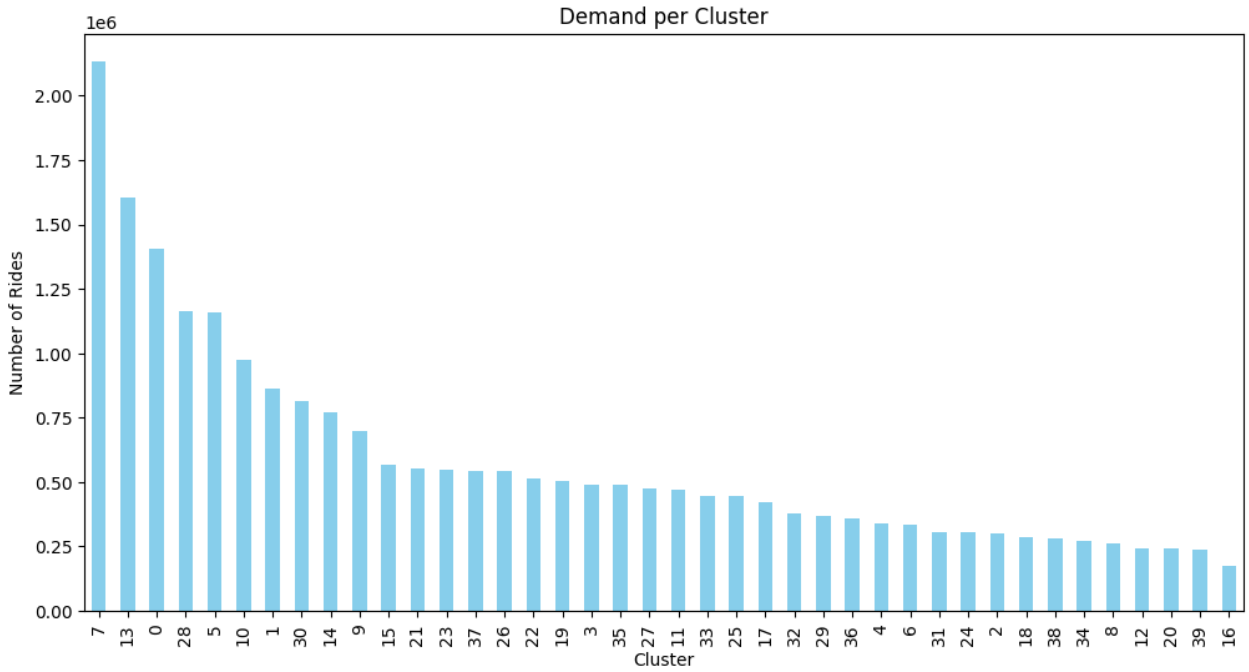

Fig. 4 Demand Per Cluster

Fig. 5 shows the boxplot visualizing the distribution of fares. Here I notice that the data points are concentrated within a reasonable range, but there is significant outlier on both the high and low ends. My analysis on this is that the outliers could present rare events like extremely long trip or the incorrect data entries. Moreover, cleaning these extreme values would be required be required if the model wouldn't show the performance expected.

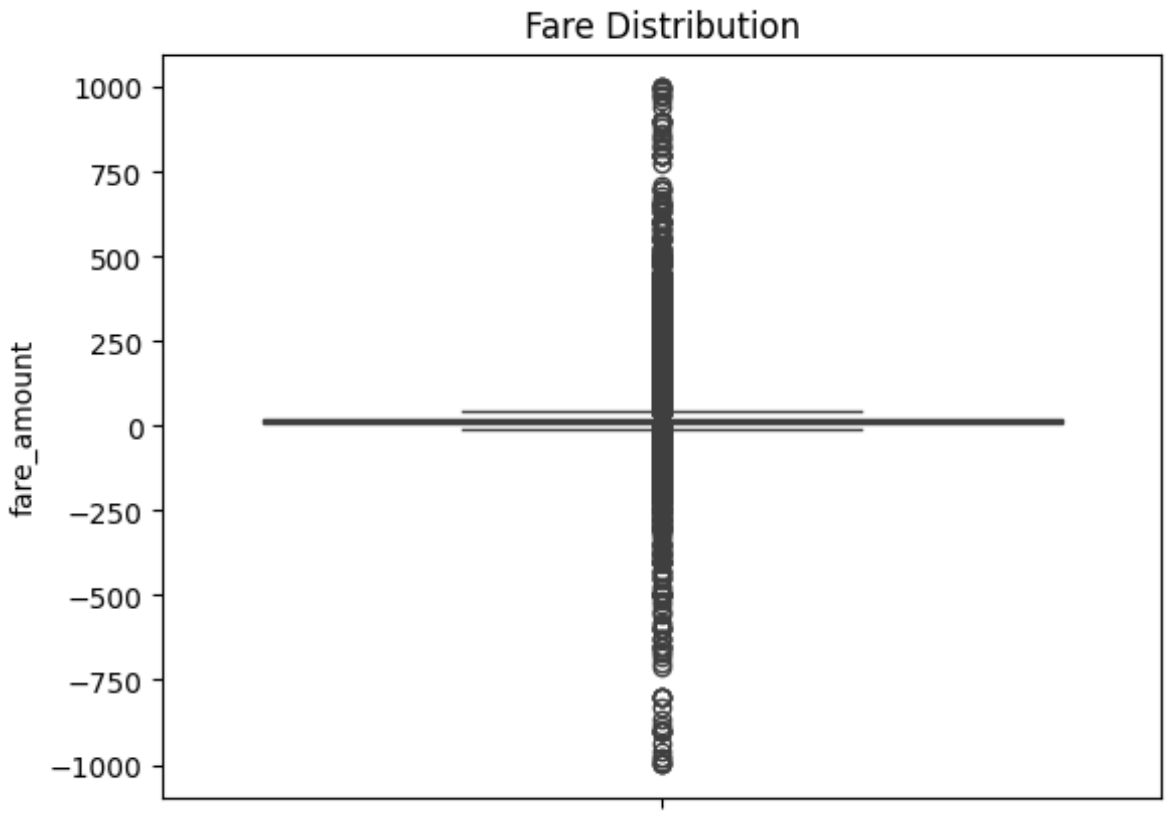

Fig. 5. Fare Distribution

Fig. 6 shows the plot tracks training and validation losses over epochs. Here I observe that the training loss decreases steadily which means that the model is learning from the training data. Moreover, validation loss also decreases initially but rises up later and stays steady which means that the model generalizes well without significantly overfitting. My analysis on this is the gap between training and validation losses are minimal which means the model is robust. Also, to include the epochs set to train was 200 but the model quit training at 83 epochs which could also be concluded from the graph that it didn't require further training, and even if further training was performed it wouldn't yield any better results.

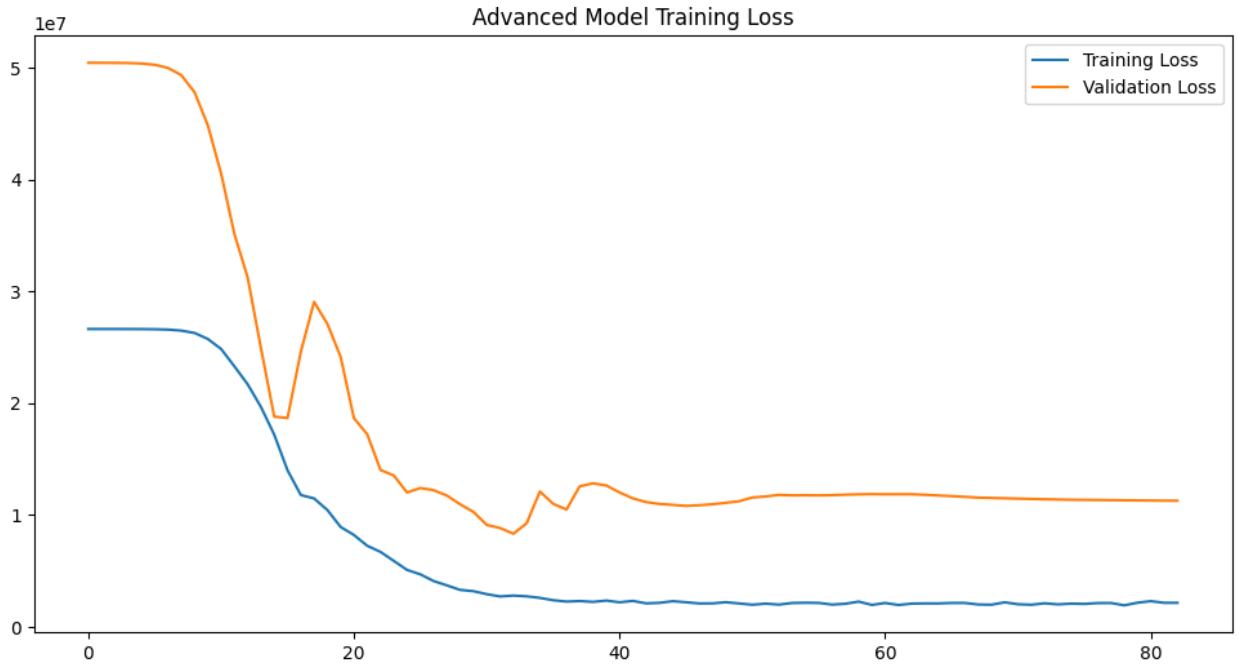

Fig. 6. Model Accuracy

Fig. 7 shows a plot of comparison between the true demand and model's predicted demand for a specific cluster over time. Here the model prediction is represented by red line that closely follows the actual demand represented by the blue line with a little deviation. I come to an analysis that the model is capable of capturing the general trends od demand fluctuations and some sharp peals in demand are slightly misaligned, which indicates that there is a room for improvement in model's accuracy to predict the extreme demanding values. Since we are working with a large dataset, accuracy achieved is very much reasonable and moreover other factors affecting this misalignment could be due to weather and holiday factors that are not taken into consideration in my work.

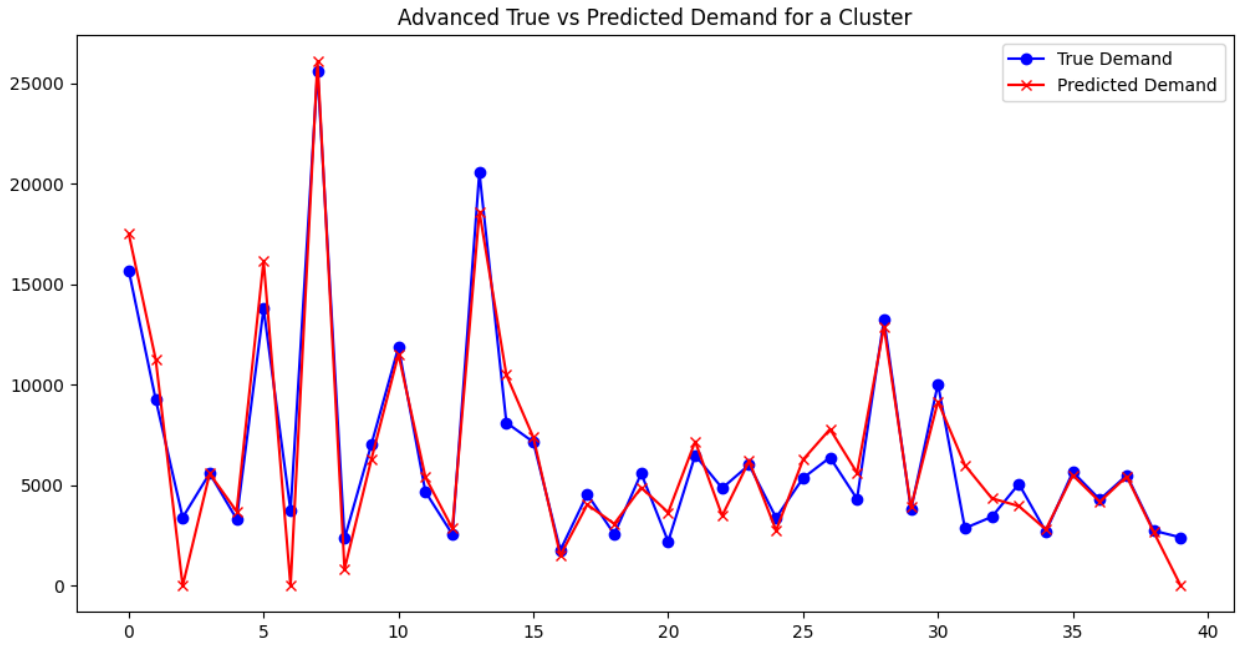

Fig. 7. Model true vs predicted

Conclusion

In this project I have addressed the problem of taxi demand prediction using machine learning and deep learning models. Systematic approach of data preprocessing, analyzing, clusters and model training are performed. I have faced a lot of issues dealing with the dataset. As it turned out to have 23 million records for the year 2024 from January to July having a lot of null values and various anomalies. The preprocessing steps consumed a lot of time because of the large dataset. I performed manual check over the data first before proceeding with the analysis. I have merged the monthly datasets from the NYC website to a single file. The dataset was in parquet format which I later formatted to csv. I performed the data cleaning first and the used the cleaned dataset file for working with the capstone project. I have also faced problems attaining the accuracy. Later over numerous amounts of time training the model with different setting I came to an accuracy of 38% which seemed to be pretty good for a dataset for 23 million

records. The training and validation loss graphs also came out precisely as they were supposed to be. Overall, from the analytics capstone project, I got to learn a lot about EDA's, data cleaning, clustering and model selection and training. This project helps various yellow cab private owners to run business efficiently without any major losses. Not only the business the drivers also get to make good amount of profits. A real time display handy for the drivers is very helpful in getting the right rides at the right location. My future scope on this project is to show highlighted regions on the map based on the predicted values.

Works Cited

https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

https://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJuniorNewYorkCityCabPricing-report.pdf

https://github.com/anafisa/NYC-Taxi-Demand-Prediction

https://github.com/pranshu1921/Taxi-Demand-Prediction-NYC

https://github.com/anafisa/NYC-Taxi-Demand-Prediction

https://github.com/hzlkat/NYC_taxi_demand_prediction

https://github.com/pranshu1921/Taxi-Demand-Prediction-NYC

https://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJuniorNewYorkCityCabPricing-report.pdf

https://www.mdpi.com/2076-3417/13/18/10192

https://findingspress.org/article/22158-changing-demand-for-new-york-yellow-cabs-during-the-covid-19-pandemic

https://arxiv.org/pdf/1711.10090

https://github.com/snowolf/Taxi-Demand-Prediction-New-York-City

https://github.com/snowolf/Taxi-Demand-Prediction-New-York-City

https://arxiv.org/abs/1711.10090

https://arxiv.org/abs/1905.01281

https://findingspress.org/article/22158-changing-demand-for-new-york-yellow-cabs-during-the-covid-19-pandemic

https://www.mdpi.com/2076-3417/13/18/10192

https://en.wikipedia.org/wiki/Yellow_cab