Taxi Demand Forecasting

Executive Summary

Ngoc Nguyen, Edward Ramirez, Sriram Raghunath, Nazanin Komeilizadeh, Noah Gillespie, Li Meng

Overview

Knowing where to go to find customers is the most important question for taxi drivers and ride hailing networks. If demand for taxis can be reliably predicted in real-time, taxi companies can dispatch drivers in a timely manner and drivers can optimize their route decisions to maximize their earnings in a given day. Consequently, customers will likely receive more reliable service with shorter wait times. This project aims to develop advanced deep learning models for time series to predict taxi demand across different pickup zones in Manhattan, NY. We will explore advanced deep learning models for time series, including Multilayer Perceptrons, Multi-series Long Short-Term Memory (LSTM), Temporal Graph-based Neural Networks, and compare them with a baseline statistical model ARIMAX.

Dataset and Preprocessing

We used a dataset from the New York City Taxi & Limousine Commission, which includes detailed records of individual taxi routes such as pickup/dropoff locations, times, pricing, and trip distances. The datasets cover all taxis and ridesharing vehicles in NYC, dating back to 2009. We processed this data into hourly time series for ride counts and average fare originating from different pickup locations. We limit our time series to yellow taxis in Manhattan taxi zones and focus on the most recent time frame, from January 1, 2022, to March 31, 2024, to sufficiently capture the seasonality patterns in the data. When cleaning the data, we impute missing values in average fare, trip distance, and passenger counts (due to zero rides during the hour) with the average of these variables across our time frame. Our training set comprises the first 80% of the time span, with the remaining 20% used for testing.

Solutions

We explored three different approaches, including a Multilayer Perceptrons model, a series of Long Short-Term Memory (LSTM) models and a Temporal Graph-based Neural Network model. These models provide a robust framework for predicting taxi demand, combining time series analysis with spatial and temporal graph-based techniques to capture the complex dynamics of taxi usage in Manhattan.

1. **Multilayer perceptron model:** This model is a basic feed forward neural network with one hidden layer. We prepared the time series into sliding windows of length 24 hours each which are used to forecast the number of pick-ups for the next hour. Each taxi zone was trained separately. We used learning rate scheduling, early stopping and dropout for robustness, and also applied hyperparameter tuning.
2. **Multi-Series LSTM Models:**
    - **Multi-Series LSTM Model:** This model is designed to train on the time series data for all taxi zones simultaneously. By incorporating a categorical variable embedding layer for the zone ID,

the model learns the unique characteristics of each zone, allowing for more accurate predictions across all areas.

- **Multi-Series Multivariate LSTM Model:** This version includes additional features such as past average fare, tip, trip distance, and time variables like the month, weekday, and hour, to capture both additional historical information and seasonality patterns.
  These models use embeddings for categorical variables, which are then concatenated with the continuous features before being fed into the LSTM layer. The models train on all zones' data simultaneously, moving through time in 24-hour steps.

3. Temporal Graph-Based Neural Network:

- **LSTM Graph Convolutional Network:** This model adds a graph structure to the LSTM architecture, leveraging both the spatial relationships between zones and the temporal nature of the data. The graph nodes in our model are represented by the 63 taxi zones in Manhattan, while the edges of the graph are the distances between zones. The model uses taxi counts as the node features, predicting the demand for the next hour based on the previous 24 hours of data for each zone.

We compared the models using two metrics: root mean squared error (RMSE) and symmetric mean absolute percent error (SMAPE). RMSE places more importance on accurately predicting the high demand zones, while SMAPE places equal importance on the predictions for all zones. Using both metrics ensures we are not improving the accuracy for high demand zones at the expense of low demand zones.

Model	KPI (Validation Set)	
	RMSE (Rides)	SMAPE
ARIMAX	40.02	69.57%
Multilayer Perceptrons	23.58	63.37%
Ride Counts LSTM	16.56	57.82%
Ride Counts + Time Vars LSTM	15.86	57.81%
Multivariate + Time Vars LSTM	16.46	57.25%
Temporal Graph-based NN	17.09	55.38%

The initial Multilayer Perceptron model outperformed the baseline ARIMAX model with a 41% reduction in RMSE. Further improvements were achieved with LSTM and graph neural networks, leading to a 60% reduction in RMSE and a 20% reduction in SMAPE compared to ARIMAX. These results suggest that temporal models could enhance prediction accuracy and potentially increase taxi rides by approximately 24 per hour per taxi zone. Encoding time variables like day of week, month, and hour improved the LSTM metrics, while including trip information like fare, trip distance, and tip ended up hurting model performance slightly. This is likely due to the data not containing any information directly connected to demand, since yellow taxi fares are determined based on a fixed rate, rather than current demand.

Overall, the temporal models performed similarly. When the model predictions are placed on a map of Manhattan side-by-side with the actual demand, the maps appear similar, with the prediction following the same general trends as the actual demand, demonstrating the viability of this model to make taxi demand predictions.

Pitfalls and Future Work

Several strategies can enhance our model's performance and extend its application. One approach is to further enhance the modeling process. Building a model that predicts demand for specific routes rather than just pickup locations could provide more granular insights into travel patterns. The model could also be modified to forecast demand for the next few hours instead of just the next hour, providing more robust predictions over longer time horizons. Additionally, we can tune hyperparameters to refine the model's accuracy.

Another direction is to extend the data inputs. For example, incorporating external features, such as weather, and expanding the training sample to include more taxi zones and ride sharing vehicle data, can improve the fit and generalizability across different scenarios.