Taxi Demand Prediction and Route Optimization

Abhinav Chauhan
Yukesh Raghavan
Stony Brook University
abhinav.chauhan@stonybrook.edu

Advised by: Prof. Vibha Mane

Reference Format:
Abhinav Chauhan and Yukesh Raghavan. 2023. Taxi Demand Prediction and Route Optimization. In *SBU, NY Final Project Report Reports, Fall 2023, NY, USA.* 4 pages.

1 INTRODUCTION

This report presents the findings and methodologies used in the final project, focusing on Taxi Demand Prediction and Route Optimization. The project involves predicting taxi demand in New York City and optimizing routes using machine learning techniques.

2 PROJECT OVERVIEW

Yellow Cab Concentration: Cabs are primarily concentrated in Manhattan but offer services across all five boroughs. Hailing a cab is convenient, with passengers able to flag one down from the curb or use designated taxi stands. High Taxi Usage Frequency: In New York City, taxis are utilized more frequently compared to many other locations. The prevalent practice involves spontaneous street hailing rather than pre-booking by phone, simplifying the process for both drivers and passengers. Transformation of Taxi Services: Uber's innovative approach allowed individuals to enroll as taxi drivers using their private vehicles. The simplicity of booking a ride through a mobile app reshaped the industry, establishing a new network of cabs beyond the traditional medallion system. Impact on Taxi Patronage: Recent studies reveal a notable decline in taxi patronage since 2011, attributed to the intensified competition from ride-sharing services. This shift underscores the need for strategic measures to revive and optimize traditional taxi services in response to evolving transportation trends.

This report is submitted to SBU in fulfillment of Practical ML and AI Course Requirements .

Capstone Final Project Report, Fall 2023, Stony Brook, NY
© 2023 Stony Brook University, New York.

3 DATA PREPROCESSING

- **Timestamp Conversion:**
  - Convert 'tpep_pickup_datetime' and 'tpep_dropoff_datetime' columns to datetime format for accurate time-based analysis.
- **Feature Extraction:**
  - Extract relevant features from timestamp data:
  - 'hour': Pickup hour.
  - 'day_of_week': Day of the week.
- **Spatial Partitioning:**
  - Introduce a 'Region' column, randomly assigning data to four regions. Customize as needed based on your dataset.
- **Feature Selection for Demand Prediction:**
  - Define feature variables (X) and target variable (y) for demand prediction.
  - Features (X): 'hour', 'day_of_week', 'trip_distance', 'passenger_count'.
  - Target (y): 'congestion_surcharge'.

3.1 Clustering Analysis: Optimal Cluster Selection

- **Objective:**
  - Identify the optimal number of clusters for spatial analysis based on taxi pickup locations.
- **Distance Constraint:**
  - Considering the observation that a taxi can cover up to 2 miles in 10 minutes, the inner cluster distance is set to be greater than 2 miles but not less than 0.5 miles.
- **Cluster Range Exploration:**

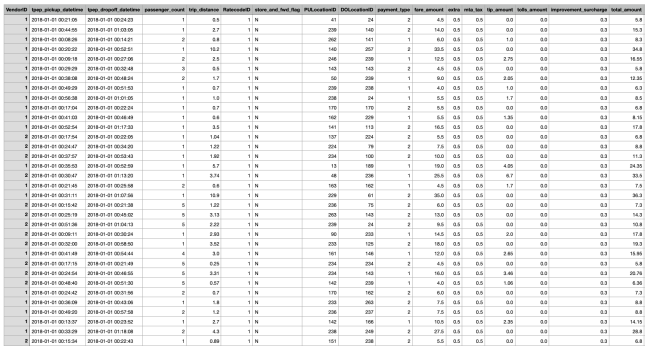

Figure 1: Dataset Preprocessing

- Explored a range of cluster values from 30 to 70 to find the optimal K value.
- Adjusted the range to align with the defined distance constraint.

- Observations:
  - Experimentation revealed that the optimal number of clusters is 30.
  - This choice adheres to the distance constraint and provides meaningful spatial partitioning.

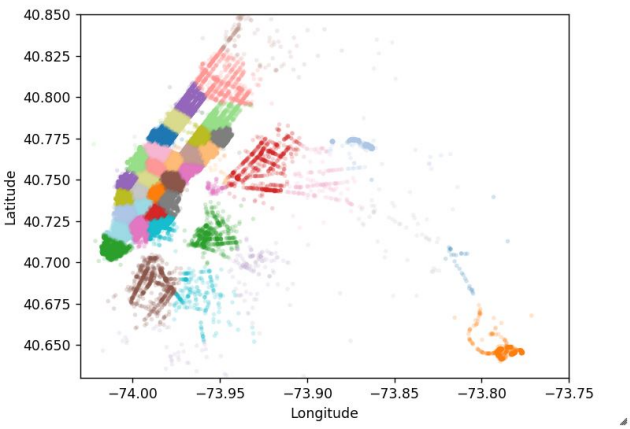

Figure 2: Spatial Clusters of New York City

3.2 Random Forest Regressor: A Powerful Choice for Predictive Modeling

- **Advantages of RandomForest:**
  - **Accuracy Maintenance:**
    - Well-suited for maintaining high accuracy across diverse datasets.
  - **Overfitting Mitigation:**
    - Robust against overfitting, especially beneficial when dealing with a large proportion of data.

Model	Mean Absolute Error (MAE)	Computation Time (seconds)
XGBoost Regressor	0.1158	213.0978
Linear Regression	0.1165	218.2750
EMA (Exponential Moving Average) of Previous Values	0.1205	218.6057
WMA (Weighted Moving Average) of Previous Values	0.1206	220.4236
SMA (Simple Moving Average) of Previous Values	0.1231	232.9855
Random Forests Regressor	0.1261	250.1323
EMA Ratios	0.1574	473.4599
WMA Ratios	0.1580	489.5664
SMA Ratios	0.1613	524.1292

Figure 3: Results of MAPE and MAE metrics

  - **Ensemble Approach:**
    - Leverages the power of ensemble learning by combining multiple decision trees.
- **Hyperparameter Tuning with RandomizedSearchCV:**
  - Applied 'RandomizedSearchCV()' for hyperparameter tuning.
  - Randomly samples a fixed number of parameter settings from specified distributions.
  - Provides a more efficient exploration of the hyperparameter space compared to exhaustive grid search ('GridSearchCV').
- **Parameter Considerations:**
  - Adjusted key parameters like the number of trees, maximum depth, and minimum samples per leaf.
- **Cross-Validation for Generalization:**
  - Utilized 3-fold cross-validation during model training.
  - Improves generalization performance by validating the model across different subsets of the training data.

3.3 XGBoost Regressor: Next Frontier

- While RandomForest excels, the exploration extends to XGBoost Regressor.
- Gradient boosted decision trees implementation designed for speed and performance.
- Training a hyperparameter-tuned XGBoost regressor on our train data for enhanced predictive capabilities.

3.4 Route Optimization

3.4.1 *Optimization using Genetic Algorithm.*

Objective: The goal of the route optimization phase is to enhance taxi service efficiency and minimize travel distances by employing a Genetic Algorithm.

Genetic Algorithm Components:

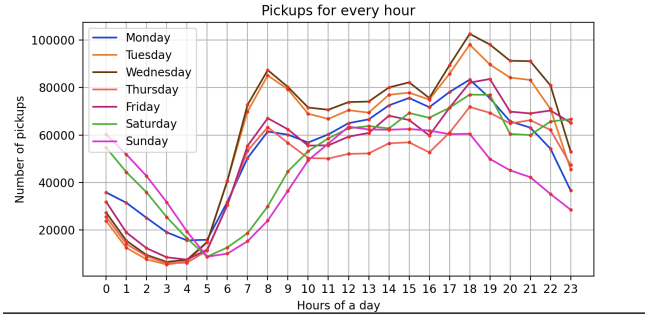

Figure 4: Pickup Demand in a Day

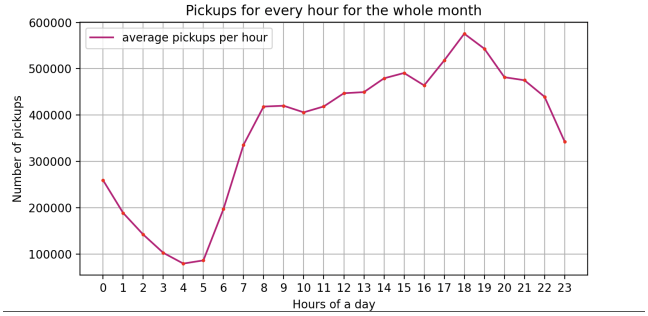

Figure 5: Pickup Demand Hourly in a month

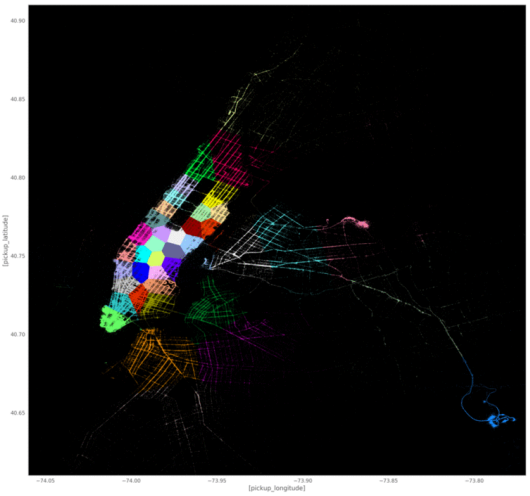

Figure 6: HeatMap of Demand Prediction

- **Individual Representation:**
  - Each individual represents a potential route, initialized with a permutation of location indices.
- **Fitness Function:**
  - The customized fitness function evaluates the total distance of a route, leveraging historical traffic data for realistic assessments.
- **Genetic Operators:**
  - Crossover (mate): Utilizing ordered crossover to generate new individuals.
  - Mutation (mutate): Employing shuffle index mutation with a 10
  - Selection (select): Utilizing tournament selection with a tournament size of 3.

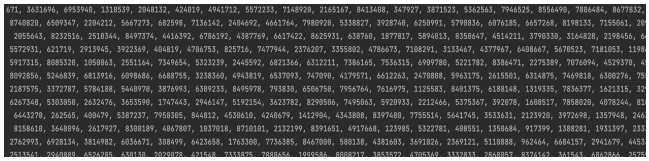

Figure 7: Optimized Route Location ID

Genetic Algorithm Execution:

- The Genetic Algorithm is implemented using `eaSimple` from the DEAP library.
- The population is initially composed of 2 individuals and evolves over 2 generations.
- The Hall of Fame (`hof`) preserves the best-performing individual throughout the evolution process.

Optimized Route: The best individual from the Hall of Fame represents the optimized taxi route. This optimized route is essentially a permutation of location indices.

Results and Future Considerations:

- The determination of the optimized route is grounded in minimizing the total travel distance.
- The consideration of historical traffic data ensures realistic and efficient route planning.
- Subsequent iterations and fine-tuning of the algorithm hold the potential for continuous improvements in route optimization.

4 CONCLUSION

The project successfully addressed challenges in the evolving taxi service landscape by predicting demand and optimizing routes. Machine learning models and spatial clustering provided valuable insights. The Genetic Algorithm demonstrated the potential for optimizing taxi routes effectively.

5 CONCLUSION: REVOLUTIONIZING TAXI DISPATCH WITH PREDICTIVE MODELING AND ROUTE OPTIMIZATION

In the dynamic landscape of New York City's taxi service, our comprehensive analysis and implementation bring forth a transformative approach to taxi dispatching, leveraging predictive modeling and route optimization techniques. Let's summarize the key insights and outcomes derived from our discussions:

5.1 Predictive Modeling for Taxi Demand:

- Utilized historical data to predict taxi ridership, offering valuable insights to dispatchers.
- The decline in taxi patronage since the advent of ride-sharing services highlighted the need for proactive strategies.

5.2 Data Preprocessing:

- Transformed raw yellow taxi datasets through essential preprocessing steps.
- Extracted relevant features, incorporated timestamps, and spatially partitioned the data for meaningful analysis.

5.3 Feature Engineering for Demand Prediction:

- Engineered features such as hour, day of the week, trip distance, and passenger count for demand prediction.
- Prepared the dataset for training predictive models by isolating input features and target variables.

5.4 Modeling and Evaluation:

- Employed Random Forest Regressor for its accuracy and resistance to overfitting.
- Utilized RandomizedSearchCV for hyperparameter tuning, optimizing both accuracy and efficiency.
- Cross-validated models to ensure robustness and generalization.
- Evaluated model performance, with XGBoost Regressor emerging as a promising candidate for further exploration.

5.5 Route Optimization with Genetic Algorithm:

- Innovatively applied a Genetic Algorithm to optimize taxi routes, considering historical traffic data.
- Defined genetic algorithm components, including individual representation, fitness function, and genetic operators.
- Demonstrated the optimization process, yielding an optimized taxi route with minimized travel distance.

5.6 Comparative Analysis of Models:

- Conducted a thorough performance comparison of various models, considering both predictive accuracy and computation time.
- Identified XGBoost Regressor as a top-performing model, balancing accuracy and efficiency.

5.7 Future Considerations:

- The route optimization framework holds potential for real-world application, considering the dynamic nature of traffic and urban environments.
- Continuous refinement and exploration of advanced models and optimization algorithms can further enhance taxi dispatching efficiency.

In conclusion, our multifaceted approach integrates cutting-edge predictive modeling with innovative route optimization, providing a holistic solution for taxi dispatch services in New York City. By embracing these techniques, taxi companies can not only adapt to changing trends but also stay at the forefront of efficiency, ultimately revolutionizing the taxi service industry.

6 REFERENCES

- Kaggle - Taxi Demand Prediction
- https://www.kaggle.com/code/ajaysh/taxi-demand-prediction
- NYC Taxi and Limousine Commission (TLC) - Trip Record Data: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Sanjay Chakraborty's Taxi Prediction Blog:
- https://sanjayc.medium.com/manhattan-taxi-demand-prediction-f16880d00fde