UE23CS352B : MACHINE LEARNING LABORATORY

Caesar's Taxi Prediction Service

NAME	Ria S Nair	Reanna Netto
SRN	PES2UG23CS476	PES2UG23CS470
SECTION	H	H

INTRODUCTION

The efficiency of urban taxi services and ride-sharing platforms hinges on the accurate prediction of customer demand across different geographical areas and times. Misallocated taxis lead to passenger wait times and increased operational costs due to unnecessary cruising. This project develops a predictive service to forecast the volume of taxi trips (demand) in the New York City (NYC) area using historical trip data, with the ultimate goal of generating actionable dispatch recommendations.

We leverage a month of NYC Yellow Taxi trip data from March 2016. The approach combines spatio-temporal aggregation with advanced ensemble methods. We transform the raw data into a structured time series for each spatial zone, engineer features incorporating historical demand lags, and compare the performance of multiple regression models - Linear Regression, Random Forest, and XGBoost - to determine the optimal strategy for short-term demand prediction. The final outcome is a prescriptive recommendation of high-demand taxi placement zones.

METHODOLOGY

1. Dataset and Data Cleaning

The primary source was the **NYC Yellow Taxi Trip Data** (March 2016), containing original trip records.

Data Cleaning: The dataset was filtered to ensure data quality and geographic relevance:

- Invalid trips (passenger count 0 or >6) and zero-distance trips were removed.
- Trips outside the typical NYC operational bounds (latitude 40.5◦ to 41◦ and longitude −74.3◦ to −73.7◦) were excluded.
- **Result:** The cleaning process yielded a final working dataset of 11,968,907 rows.

2. Aggregation Strategy and Feature Engineering

To convert sparse trip data into a continuous demand signal, we employed an aggregation strategy focused on both space and time:

- **Spatial Discretization:** The pickup locations were binned into a grid using a 0.01◦×0.01◦ resolution, creating `lat_bin` and `lon_bin` features, approximating a 1 km^2 zone size.
- **Temporal Grouping:** The data was aggregated by `lat_bin`, `lon_bin`, `hour` (0−23), and `weekday` (0−6).
- **Target Variable:** The target variable, **`demand_count`**, represents the total number of trips originating from a specific spatial bin during a specific hour and weekday.

Key Features Engineered:

- **Spatio-Temporal Features:** `lat_bin`, `lon_bin`, `hour`, `weekday`.
- **Lagged Demand:** To capture historical patterns, three time-series lag features were created for each unique grid cell: `demand_lag_1h`, `demand_lag_2h`, and `demand_lag_3h`. These represent the demand count in the same grid cell for the preceding 1, 2, and 3 hours, respectively.
- **External Feature (Extension):** A simulated binary weather variable, `weather_rain`, was added to the extended XGBoost model to test the value of external data.

3. Modeling and Validation Approach

The aggregated data (`demand_df`) was split temporally, with the first 80% used for training (Xtrain,Ytrain) and the final 20% for testing (Xtest,Ytest).

Models Implemented:

1. **Linear Regression:** Baseline model.
2. **Random Forest Regressor:** Non-linear ensemble model.
3. **XGBoost Regressor:** Advanced gradient boosting machine (selected as the primary advanced model).
4. **ARIMA and LSTM:** Explored as pure time-series models aggregated by hour (ignoring spatial context) for comparison.

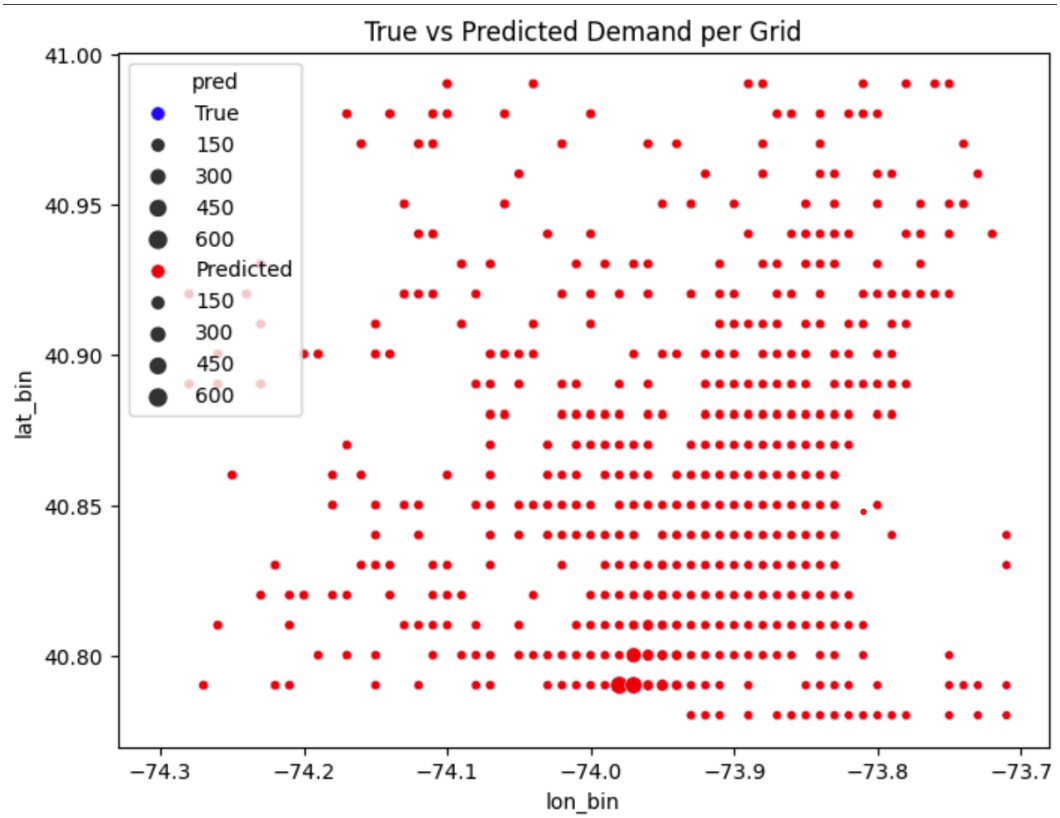
Validation:

- **Cross-Validation: Time Series Split (TSCv)** with n_splits=5 was performed on the training set to evaluate model stability against data from different time periods.
- **Hyperparameter Tuning: RandomizedSearchCV** was used to find optimal parameters for the Random Forest model, resulting in BestParameters:{max_depth:19,min_samples_leaf:2,min_samples_split:9,n_estimators: 269}.

Evaluation Metrics: The regression models were assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R2 Score).

RESULTS

Model Performance Comparison



The table below summarizes the performance of the three spatial-temporal models on the held-out test set:

Model	RMSE(Root Mean Squared Error)	MAE (Mean Absolute Error)	R2 Score(Variance Explained)
Linear Regression (Baseline)	50.91	41.21	0.903
Random Forest Regressor	36.49	12.96	0.950
XGBoost Regressor	37.07	16.27	0.949
XGBoost (Extended with weather)	40.50	17.81	N/A

Time Series Model Results (Aggregated Hourly Demand):

- **ARIMA:** RMSE 6720.15, R2 0.030.
- **LSTM:** RMSE 3838.93, R2 0.681.

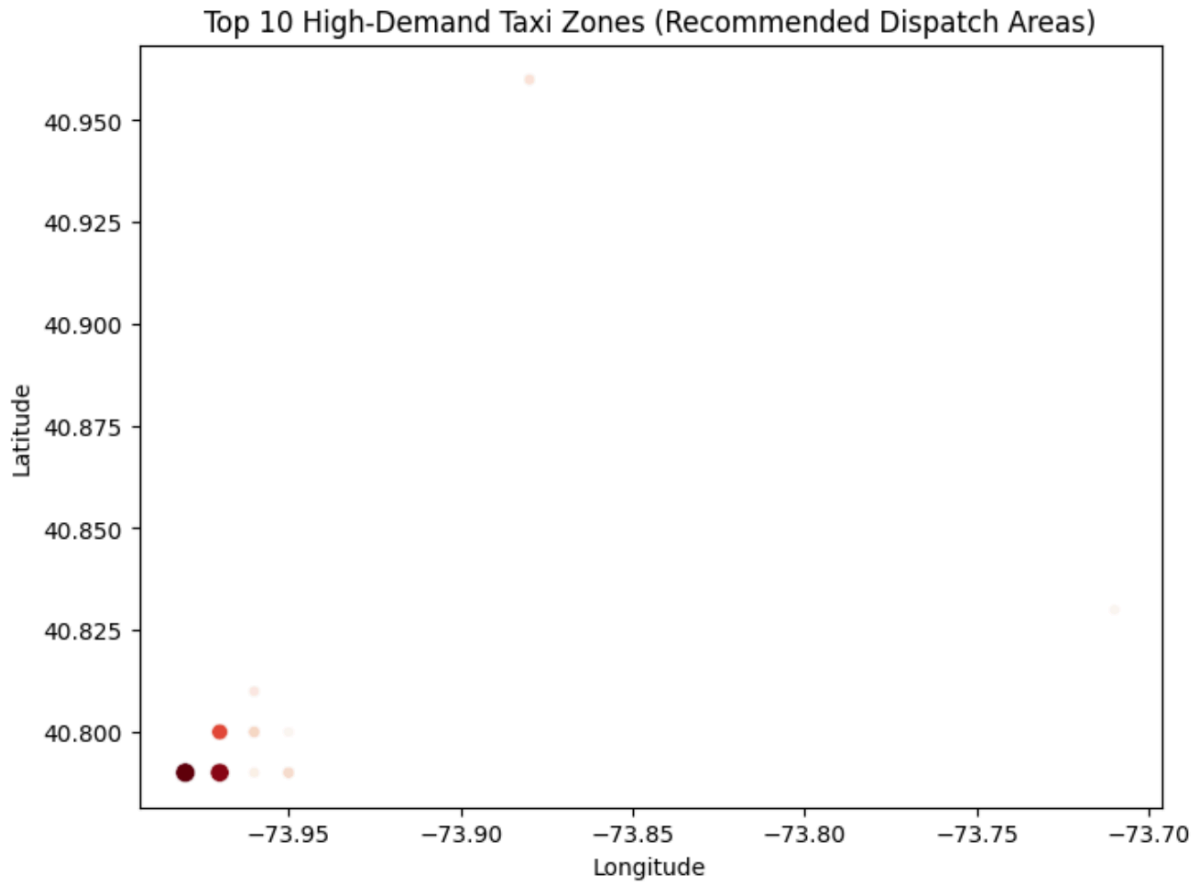
Cross-Validation Results (Random Forest)

The Time Series Cross-Validation showed high volatility in R2 across folds, indicating varying performance across time periods, a common characteristic in financial and mobility forecasting:

- Mean RMSE: 244.08
- Mean MAE: 77.45
- Mean R2: 0.824

Prescriptive Output: Top 10 Recommended Taxi Placement Zones

The final XGBoost model (without weather, as its inclusion increased error) was used to predict demand across the test set. By averaging the predicted demand per grid cell, the following locations were identified as the highest-demand zones, providing direct guidance for taxi dispatchers.



Rank	Latitude Bin	Longitude Bin	Predicted Demand (Mean)
1	40.79	-73.98	707.53
2	40.79	-73.97	654.44
3	40.80	-73.97	462.93
4	40.80	-73.96	215.72
5	40.79	-73.95	207.82
6	40.96	-73.88	183.43
7	40.81	-73.96	163.28
8	40.79	-73.96	145.82
9	40.80	-73.95	127.14
10	40.83	-73.71	126.93

DISCUSSION AND CONCLUSION

Model Performance Analysis

The results overwhelmingly support the use of tree-based ensemble methods over simple linear models for this spatio-temporal forecasting task.

- **Ensemble Superiority:** Both Random Forest (R2=0.950) and XGBoost (R2=0.949) achieved significantly lower prediction errors and higher variance explanation than Linear Regression (R2=0.903). This indicates that the relationship between location, time, and historical demand is non-linear and involves complex feature interactions that ensemble methods are well-suited to capture.
- **Importance of Lag Features:** The strong performance is largely attributable to the custom-engineered **lagged demand features**, which leverage the auto-regressive nature of demand within each specific geographic zone.
- **Failure of Pure Time Series:** ARIMA and LSTM models performed poorly when applied to the single aggregated hourly time series, confirming that ignoring the spatial context of demand leads to non-actionable forecasts.

Extended Features and Practical Deployment

- **External Features:** The inclusion of the simulated `weather_rain` feature slightly degraded the model performance (XGBoost RMSE increased from 37.07 to 40.50). This suggests either that the simulated feature was not correlated with the target or that its impact is marginal compared to the historical demand signals.

- **Prescriptive Value:** The project successfully moved from prediction to **prescription** by generating the **Top 10 Hotspots** based on forecast demand. This output is directly usable for tactical taxi dispatching.
- **Simulation:** The "live taxi demand updates" simulation demonstrated a potential implementation path for a real-time dispatch dashboard.

Conclusion

The **XGBoost Regressor** stands as the optimal model for this predictive service, providing a highly accurate spatial-temporal demand forecast ($R2 \approx 0.95$). The success hinges on the effective feature engineering of spatial bins and lagged demand features. This model allows for the efficient identification of future high-demand zones, enabling intelligent taxi dispatch and operational optimization.

Future Work:

1. **Refined External Data:** Incorporate *actual* minute-by-minute weather data (temperature, wind speed, precipitation) and scheduled event data (concerts, sporting events).
2. **Origin-Destination (OD) Modeling:** Predict *both* pickup and drop-off counts to understand flow dynamics, which would be essential for repositioning empty taxis.
3. **Deep Learning:** Explore advanced graph neural networks (GNNs) to model the spatio-temporal dependencies between adjacent grid cells simultaneously.