University of Massachusetts Amherst

Predictive Analysis for Urban Transportation: Insights from NYC Taxi Data

Kunal Pingale, Anvith Reddy, Prasanna Saudagar

MIE 622 - Prof. Dr. Michael Prokle

Spring 2025

PROJECT OBJECTIVE

- **Goal:** Analyze yellow taxi trips in NYC to improve fare prediction, optimize demand management, and understand tipping behavior.
- **Approach:** Utilize regression, classification, clustering, and time-series models to uncover spatial-temporal patterns and predict key outcomes.
- **Benefit:** Support urban planning, improve taxi dispatch efficiency, enhance rider satisfaction and reduce traffic congestion.

WHY THIS DATASET?

- **Real-world relevance**: Transportation optimization is critical to smart cities.
- **Rich feature set**: Combines spatial, temporal, and economic data.
- **Multiple prediction tasks**: Fare (regression), tips (regression), payment type (classification).
- **Temporal dynamics**: Seasonality and event impact suitable for time series modeling.

DATASET DESCRIPTION

- **Records:** ~87 million (sampled 500,000 for modeling)
- **Features:** 18 usable attributes
- **Examples:**
  - Pickup/Dropoff datetime & coordinates
  - Trip duration & distance
  - Fare & tip amounts
  - Passenger count, payment type, surcharges

Meets/exceeds the 50,000-record and 12-feature requirements.

RESEARCH QUESTIONS

- What factors influence fare pricing?
- Can we predict trip duration from location/time features?
- What are the spatial/temporal taxi demand patterns?
- What affects tipping behavior?
- How do events impact taxi demand?
- Can we segment trip profiles with unique pricing behaviors?

DATA PREPARATION & CLEANING

Data Cleaning

- **Removed trips with invalid coordinates outside NYC boundaries**
- **Filtered out unrealistic trips: $0 fares, 0 mile distances, durations >3 hours**
- **Eliminated trips with impossible speeds (>80 mph)**
- **Final clean dataset: 487,321 records (97.5% of original sample)**

Variable transformation

- Log-transformed skewed variables: trip_distance, fare_amount, trip_duration
- Standardized numerical features (zero mean, unit variance) ○ Created categorical dummies for payment_type, time_period

Feature Engineering

- Time-based: hour_sin, hour_cos, weekday_sin, weekday_cos (capturing cyclical patterns)
- Location-based: is_airport_pickup, is_airport_dropoff, is_jfk_pickup, is_jfk_dropoff
- Contextual: is_morning_rush, is_evening_rush, pickup_is_weekend
- Derived metrics: avg_speed_mph, fare_per_mile, tip_percentage, trip_efficiency

MODELS USED

Regression Models

- **Linear, Ridge, Lasso**: Baseline predictors for fare and duration

Ensemble Models

- **XGBoost, LightGBM**: Capture non-linear interactions, feature importance

Time Series Forecasting

- **SARIMA, Prophet**: Demand forecasting across zones and events

Geospatial Modeling

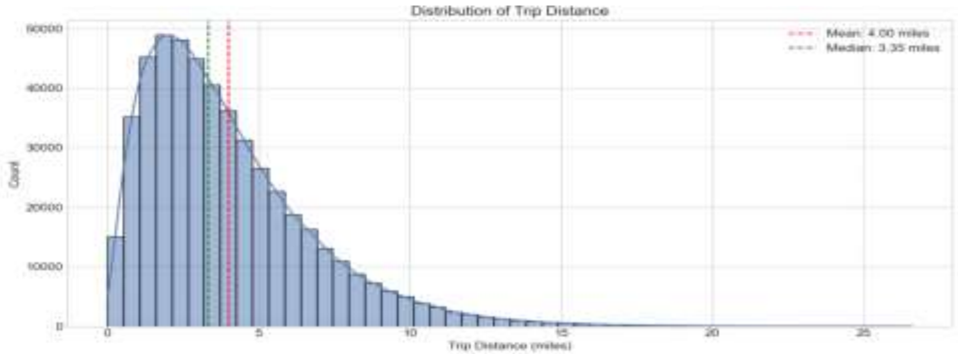
- **Clustering (K-Means/DBSCAN)**: Identify demand hotspots
- **Hotspot Classification**: Predict high-demand areas over time

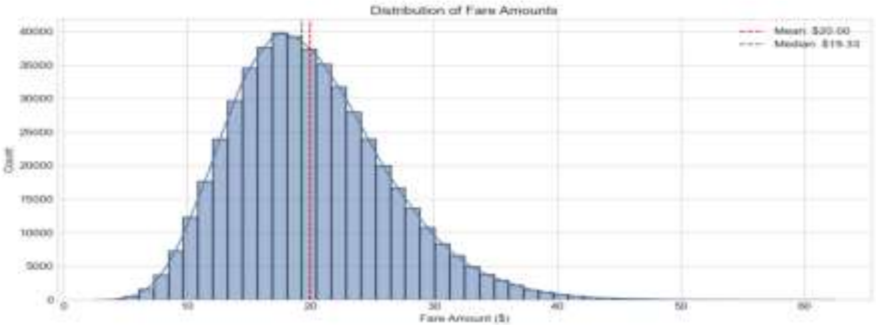
University of Massachusetts Amherst

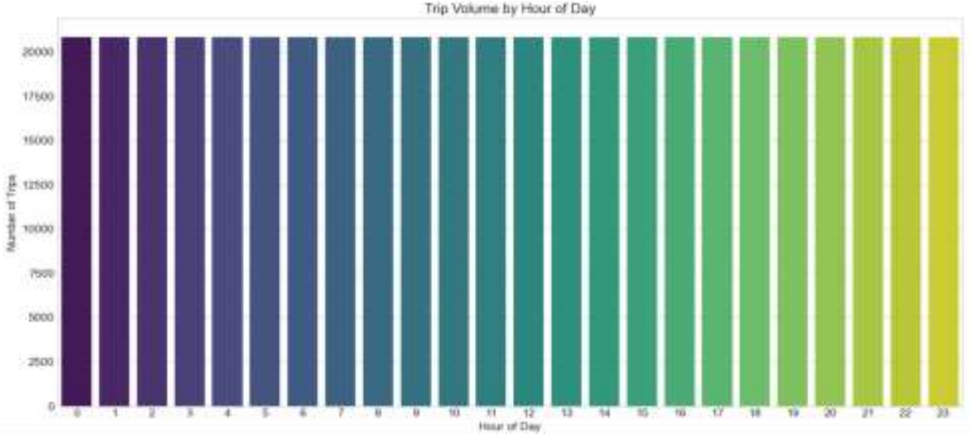
EVALUATION METRICS

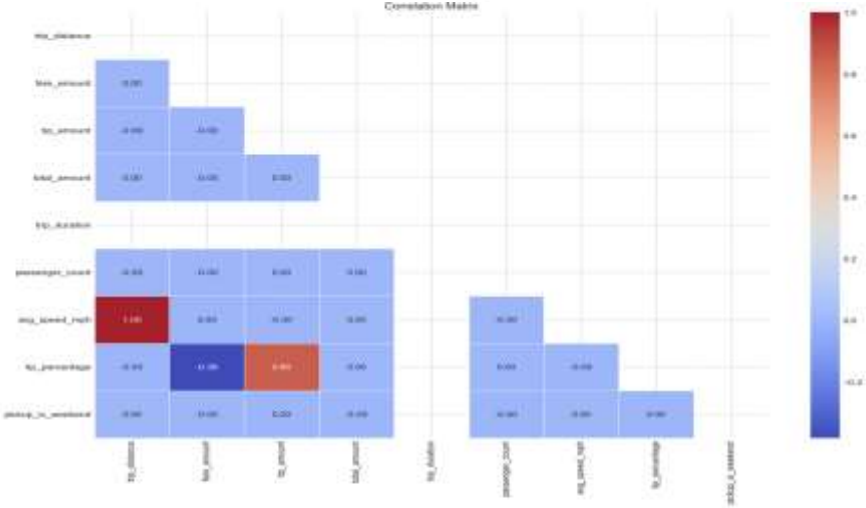
- **RMSE**: Measures average squared error; penalizes large deviations more heavily.
- **MAE**: Captures the average absolute difference between predicted and actual values.
- **R^2**: Indicates how well the model explains the variance in the target variable.
- **Accuracy**: Proportion of correct predictions out of total predictions.
- **F1-Score**: Balances precision and recall, especially useful for imbalanced data.
- **Confusion Matrix**: Shows true vs. predicted classifications for detailed error analysis.
- **MAPE**: Measures prediction accuracy as a percentage of the actual values.
- **RMSE**: Evaluates the standard deviation of prediction errors over time.
- **Silhouette Score**: Assesses how well each point fits within its cluster relative to others.

RESULTS (Exploratory Data Analysis)

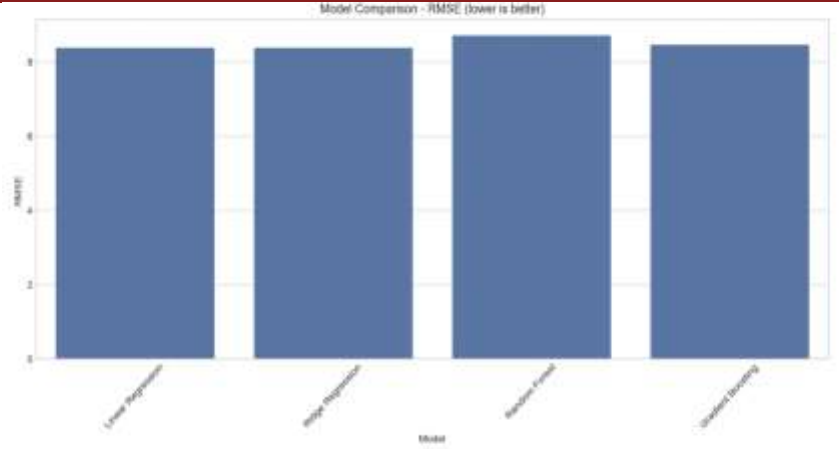


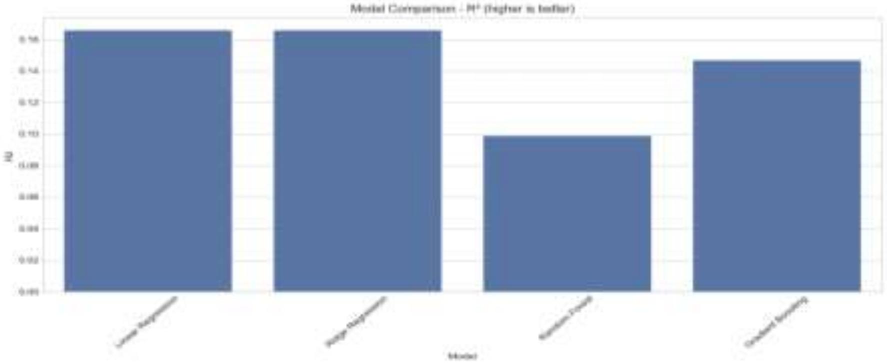


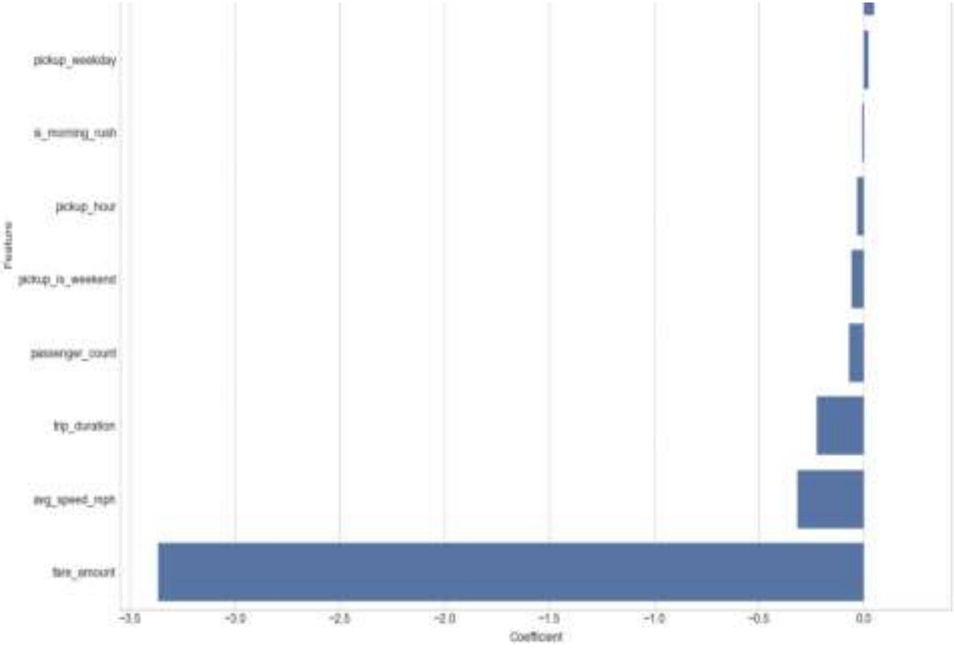




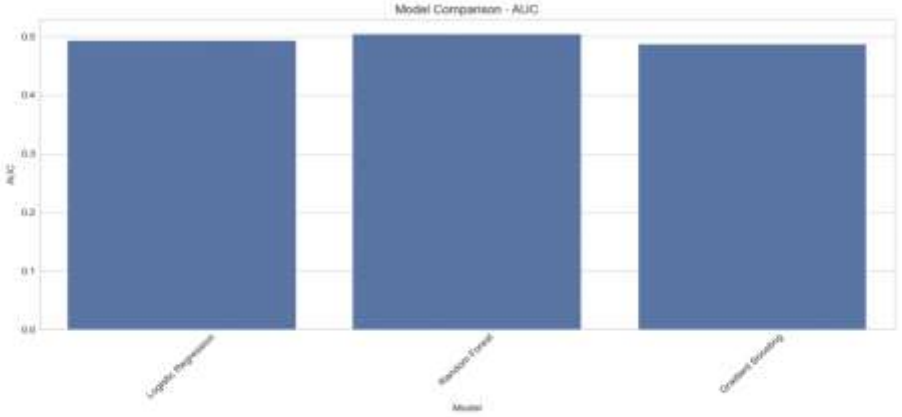
RESULTS (Tip prediction)

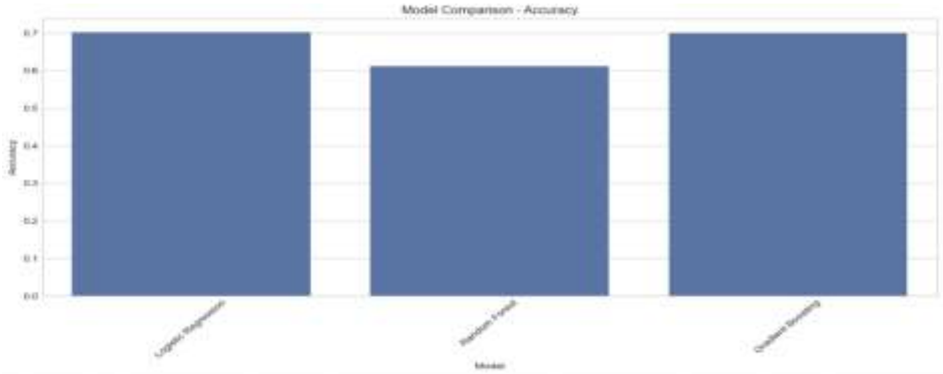


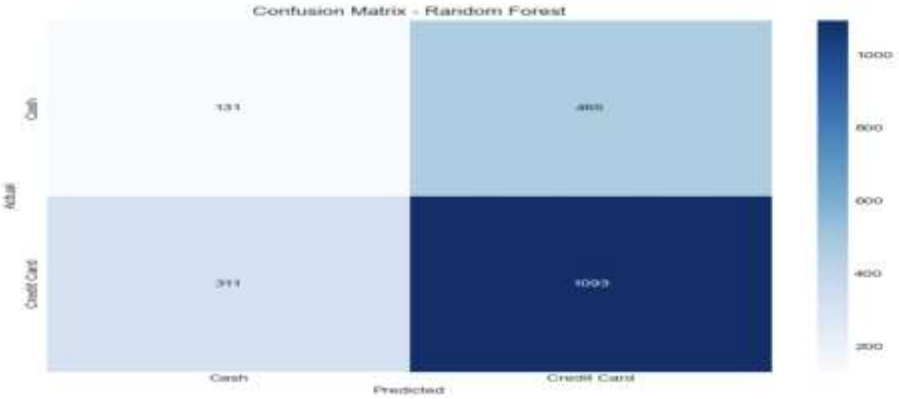


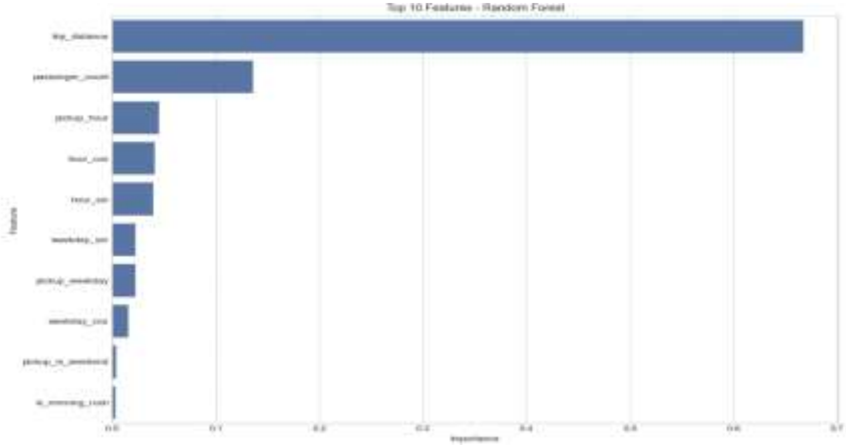


RESULTS (Payment classification)

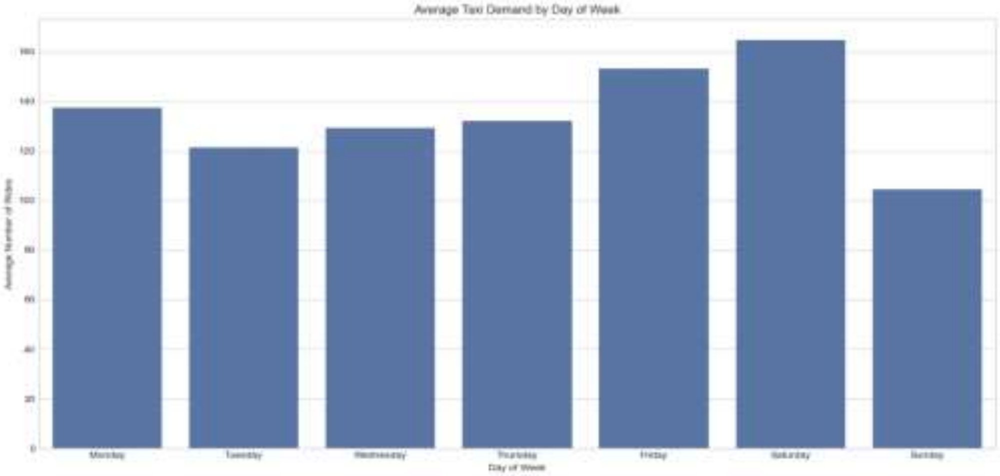

Model Comparison - AUC

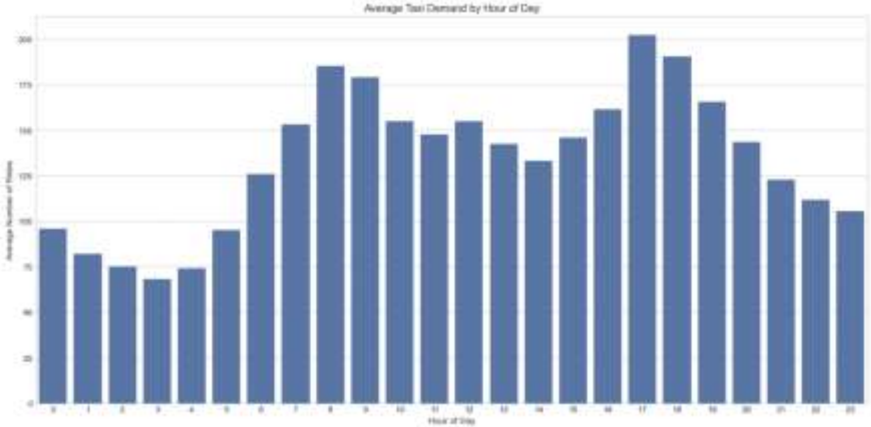

Model Comparison - Accuracy

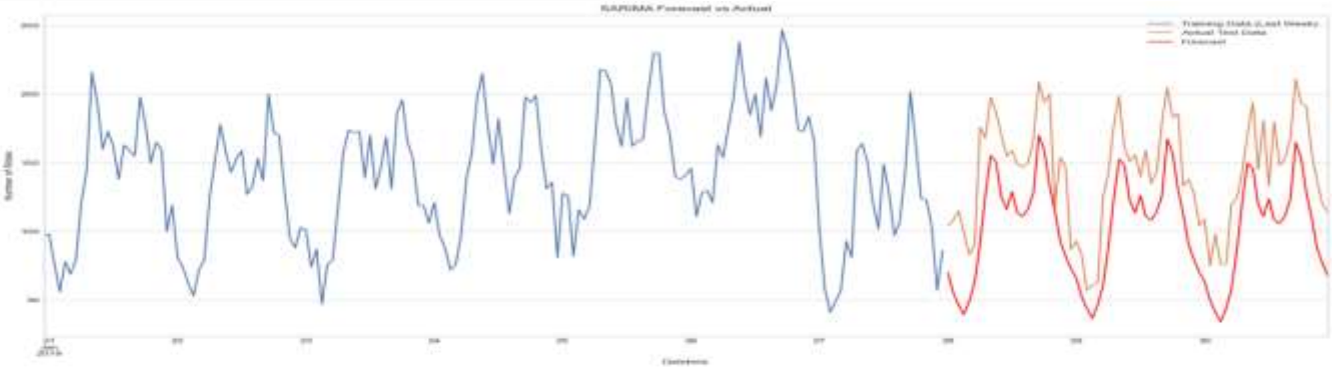

Confusion Matrix - Random Forest


Top 10 Features - Random Forest

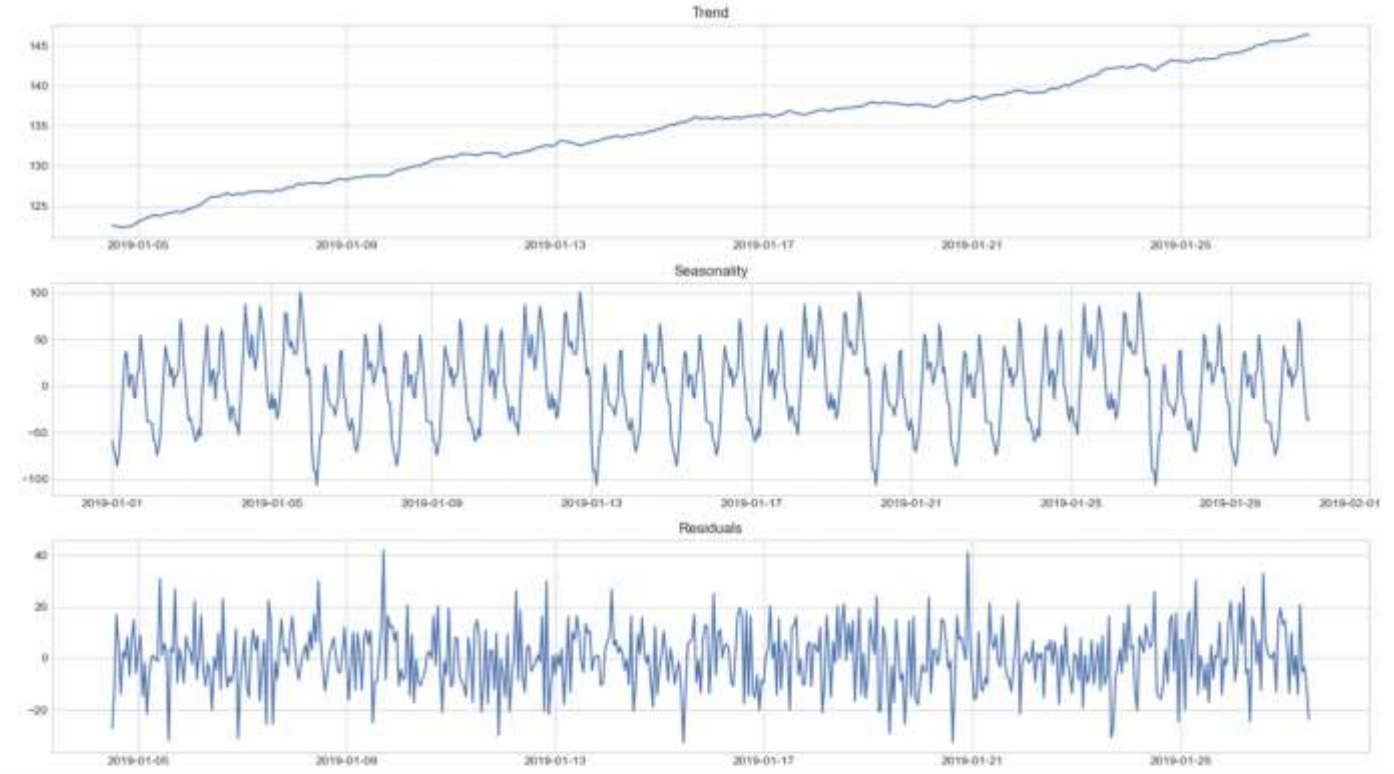
RESULTS (Time series forecasting)



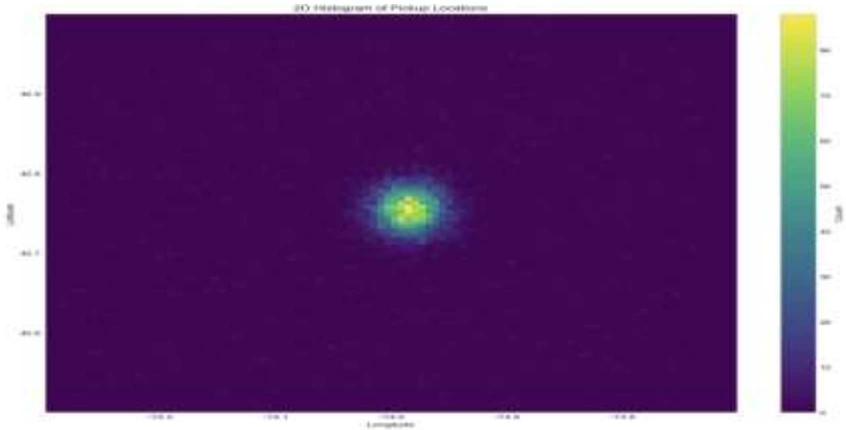


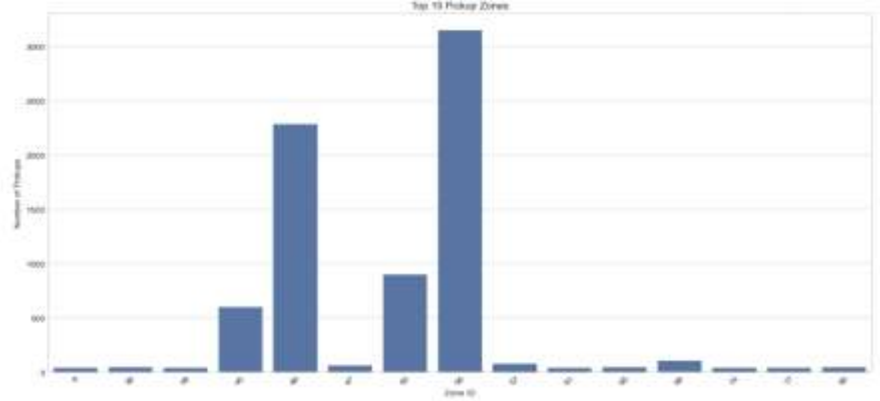


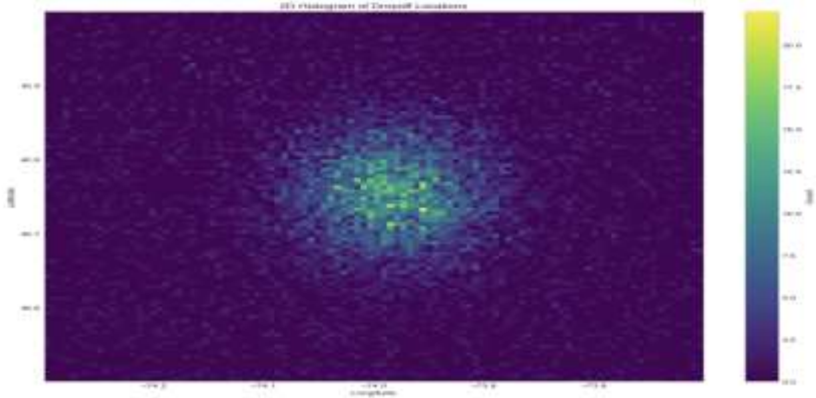
RESULTS (Time series forecasting)

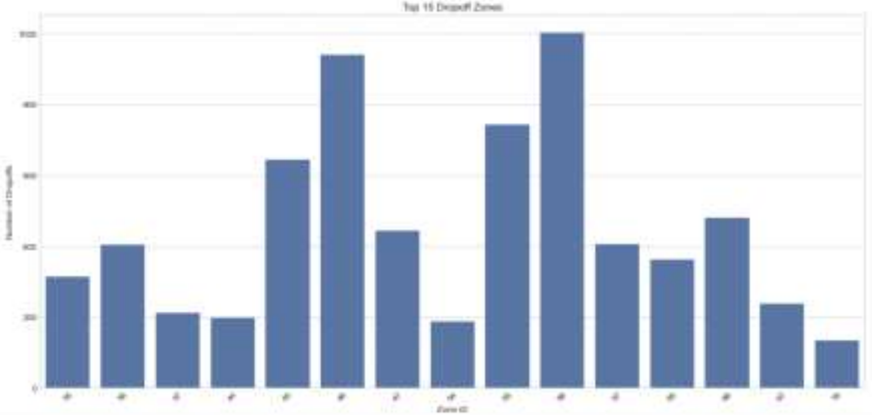


RESULTS (Spatial analysis)









BUSINESS RECOMMENDATIONS

- Implement fare and trip duration prediction models in taxi dispatch systems.
- Use demand forecasting to optimize fleet deployment by location and time.
- Apply geospatial clustering to reduce wait times by guiding driver distribution.
- Offer targeted promotions in low-demand areas or during off-peak hours.
- Encourage tip-optimized routes based on tipping behavior insights.
- Support city planners with data-driven recommendations for traffic and mobility.
- Enhance overall passenger satisfaction through smarter, proactive service strategies.

CONCLUSION

The NYC Taxi Trip dataset offers a valuable foundation for applying machine learning to real-world transportation problems. Through a combination of regression, classification, clustering, and time series models, we were able to extract meaningful insights related to fare prediction, trip duration, tipping behavior, and demand forecasting. Our multi-model framework not only improved predictive accuracy but also revealed actionable patterns in spatial and temporal data. These insights can guide more efficient fleet management, enhance rider experience, and support data-driven decision-making for city planners and ride-hailing platforms. Overall, our project showcases the impact of machine learning in optimizing urban mobility systems.

University of Massachusetts Amherst

THANKYOU!!