COURSE PROJECT –DATA MINING 2025

Lecturer - Professor Chen Hajaj

DEAN ISRAEL AND ESTER CARMIEL

our problem

Taxi companies face ongoing challenges in understanding how trip patterns, geographic locations, and pricing structures influence demand and revenue. A lack of clear insights into passenger behavior can result in inefficient fleet utilization, suboptimal driver allocation, and reduced profitability. Without data-driven analysis, it is difficult for companies to make informed operational and pricing decisions.

Project goals

This project aims to analyze real-world taxi data in order to understand trip patterns and support data-driven operational decisions.

Identify distinct types of taxi trips based on time, distance, location, and pricing characteristics

Uncover hidden patterns in passenger nehavior using unsupervised learning techniqes.

Analyze how trip characteristics influence demend and profitability.

Provide actionable insights that can support better operational decisions for taxi companies and drivers.

methods/techniques

- Converted and standardized time-related fields
- Created temporal features to capture time-of-day and weekly patterns
- Engineered trip efficiency and pricing-related features
- Applied cyclic encoding for time variables
- Derived behavioral indicators such as tipping patterns
- Filtered unrealistic values using domain-based rules

Dataset and a short list of features

Dataset

Taxi trip records from San Francisco (2023)

Over 4.1 million trips, collected via a mandatory city Taxi API

Reliable, large-scale real-world transportation data

High data quality, including validation flags that allowed filtering invalid or unrealistic trips

Main Feature Groups

- Temporal: trip time, duration, time-of-day indicators
- Spatial: pickup & drop-off locations, core demand area
- Trip characteristics: distance, hail type
- Financial: fare amount, fare per km, tips, tolls

Methodology

Unsupervised Learning

- K-Means – used to identify distinct trip types based on distance, duration, and pricing characteristics
- DBSCAN – used to detect dense pickup areas and identify core demand zones without predefining the number of clusters

Supervised Learning

- Logistic Regression – selected as a simple, interpretable baseline for trip profitability classification
- Random Forest – chosen for its ability to capture non-linear relationships and handle feature interactions
- XGBoost – selected for its strong performance on structured data and robustness to feature scale
- Neural Network (MLP) – used to model complex patterns and compare performance with tree-based models

How did we do it?

Unsupervised Experiments

Applied K-Means to identify distinct trip types
Applied DBSCAN to detect dense pickup areas and core demand zones
Evaluated clusters using cluster profiling and spatial visualization

Supervised Experiments

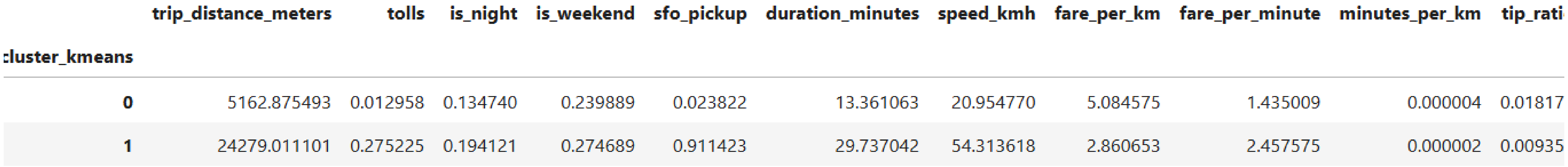
Defined a binary target: trip profitability
Trained multiple classification models for comparison
Evaluated performance on a held-out test set

Evaluation Metrics

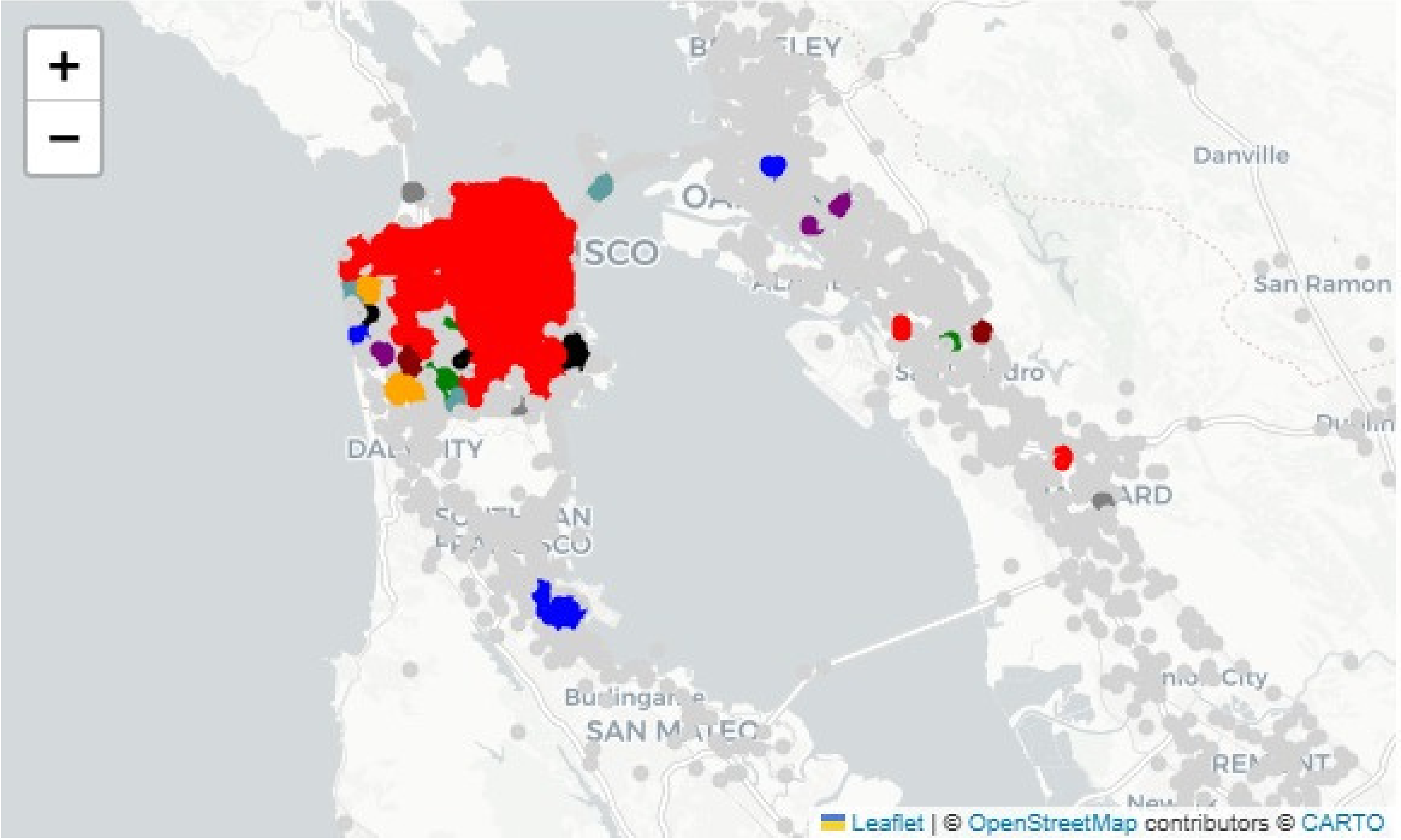
Accuracy – overall classification performance
Precision & Recall – quality of profitable trip detection
F1-score – balance between precision and recall
ROC-AUC – model discrimination ability

Results

Cluster 0 is more profitable, as it represents short trips with higher fare per kilometer, minimal toll costs, and higher tip ratios, despite lower average speed.



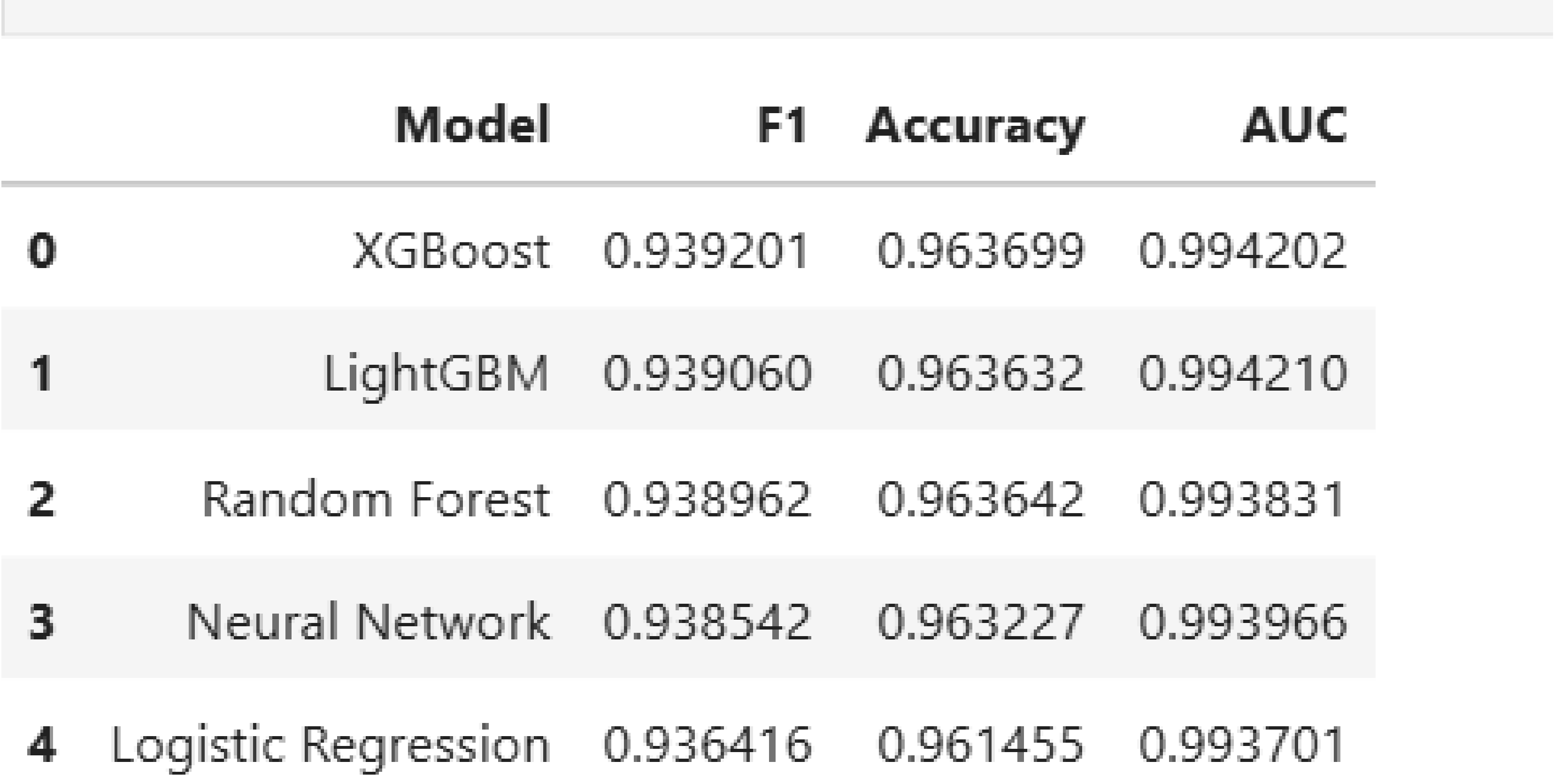
	trip_distance_meters	tolls	is_night	is_weekend	sfo_pickup	duration_minutes	speed_kmh	fare_per_km	fare_per_minute	minutes_per_km	tip_rati
cluster_kmeans											
0	5162.875493	0.012958	0.134740	0.239889	0.023822	13.361063	20.954770	5.084575	1.435009	0.000004	0.01817
1	24279.011101	0.275225	0.194121	0.274689	0.911423	29.737042	54.313618	2.860653	2.457575	0.000002	0.00935



Leaflet | © OpenStreetMap contributors © CARTO

The map clearly shows a dense central “red” area that represents the main demand hub, while smaller surrounding clusters indicate localized but consistent pickup zones rather than random noise.

COMPARISONS



	Model	F1	Accuracy	AUC
0	XGBoost	0.939201	0.963699	0.994202
1	LightGBM	0.939060	0.963632	0.994210
2	Random Forest	0.938962	0.963642	0.993831
3	Neural Network	0.938542	0.963227	0.993966
4	Logistic Regression	0.936416	0.961455	0.993701

Best overall performance: XGBoost with the highest F1 Score

conclusions

Taxi trip profitability is primarily driven by trip distance and spatial location.

Short trips originating from high-demand core areas consistently generate higher revenue per kilometer and higher tips, despite lower total fares.

Unsupervised clustering revealed a clear central demand zone, which proved to be a strong predictor of profitability.

Across all supervised models, trip distance and location-based features dominated model decisions, with XGBoost achieving the best overall performance.

These results show that combining spatial clustering with predictive modeling enables reliable identification of profitable trips and supports data-driven decision-making for drivers and taxi operators.

Thank you