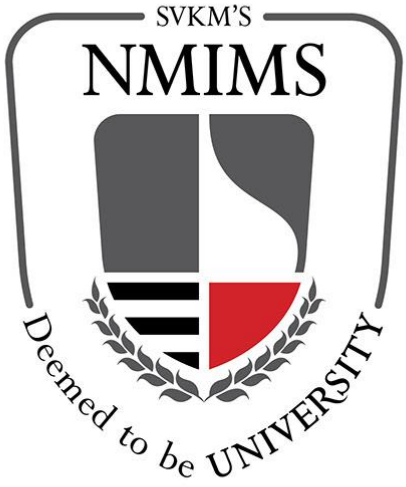
Predictive Analysis

Name: N. Abhinav	Sap Id: 70572200031
Roll.No : L035	Date: 06-04-2025

NYC Taxi Fare Prediction using Machine Learning - April 2024

Submitted To: Prof. Rajesh Prabhakar



1. Project Overview

This project focuses on building a predictive model to estimate taxi fares in New York City using machine learning algorithms. It is developed using April 2024 data collected from the NYC Taxi and Limousine Commission (TLC) for Green Taxis. An interactive front-end is developed using Streamlit for real-time fare predictions.

2. Objective

The main goal of this project is to accurately predict taxi fare amounts based on relevant ride features like distance, pickup and drop-off locations, and time of travel. This helps improve transparency and supports fare estimation for customers.

3. Dataset Description

- **Source**: NYC TLC (https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page)
- **Dataset**: Green Taxi Data - April 2024
- **Format**: Parquet
- **Key Columns**:
    - lpep_pickup_datetime
    - lpep_dropoff_datetime
    - trip_distance
    - fare_amount
    - passenger_count
    - payment_type
    - PULocationID, DOLocationID

4. Data Preprocessing

Performed the following steps:

- Removed null or invalid rows (negative distances or fares).
- Converted datetime columns to datetime format.
- Extracted day, hour from pickup time for feature engineering.

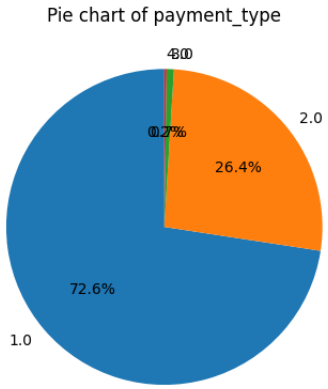
5. Feature Engineering

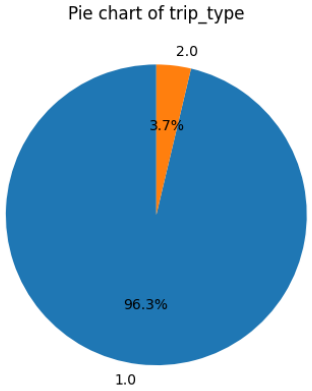
- Derived trip_duration using pickup and dropoff timestamps.
- Extracted temporal features like hour, day, weekday from pickup datetime.
- Removed outliers from trip_distance and fare_amount.

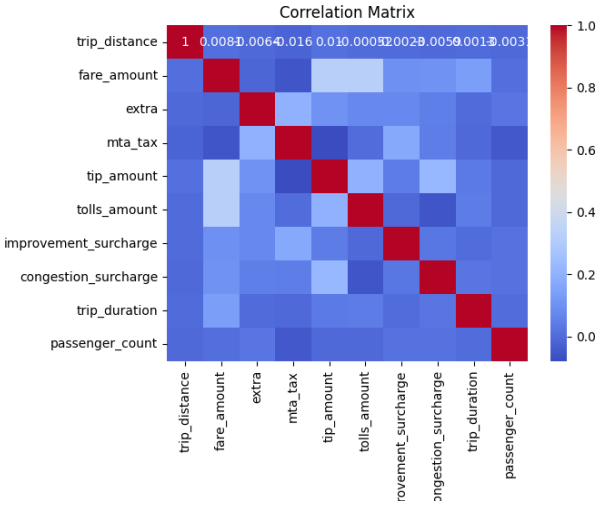
6. Exploratory Data Analysis (EDA)

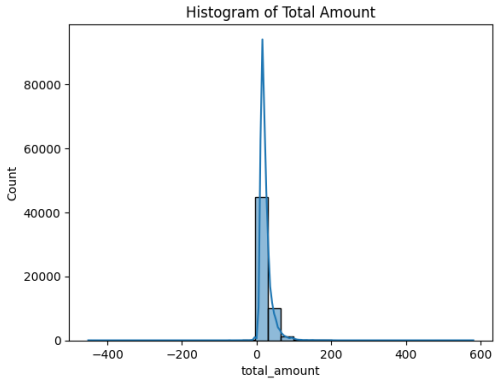
Visualized key distributions:

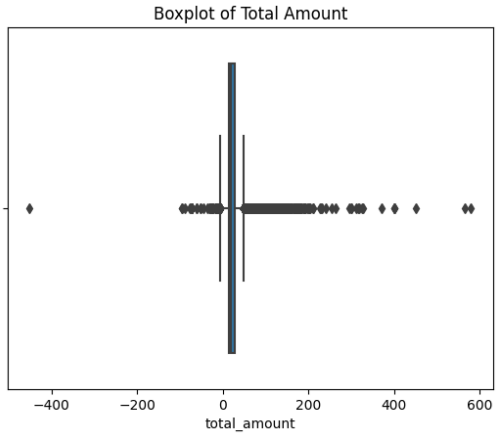
- Distribution of fare amount.
- Trip distance vs. fare amount correlation.
- Passenger count statistics.
- Hourly trip frequency

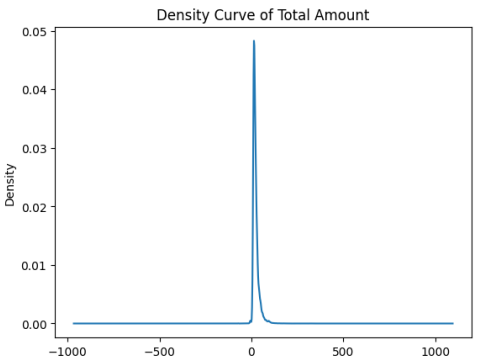












7. Hypothesis Testing

- H0: Trip distance has no effect on fare amount.
- H1: Trip distance positively impacts fare amount.
- Result: Significant p-value (< 0.05) confirming that distance impacts fare amount.

8. Model Used

Used **Linear Regression** model to predict the fare.

- Split data into training and testing sets (80-20).
- Evaluated with RMSE and R2 Score.

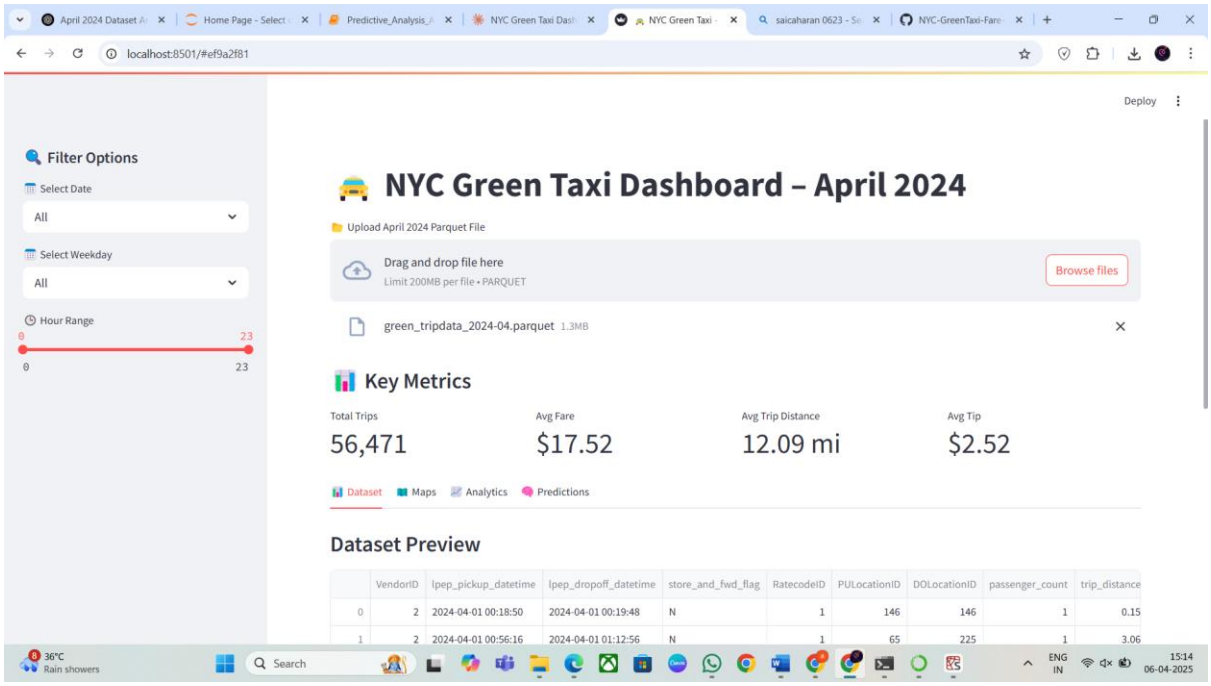
Results:

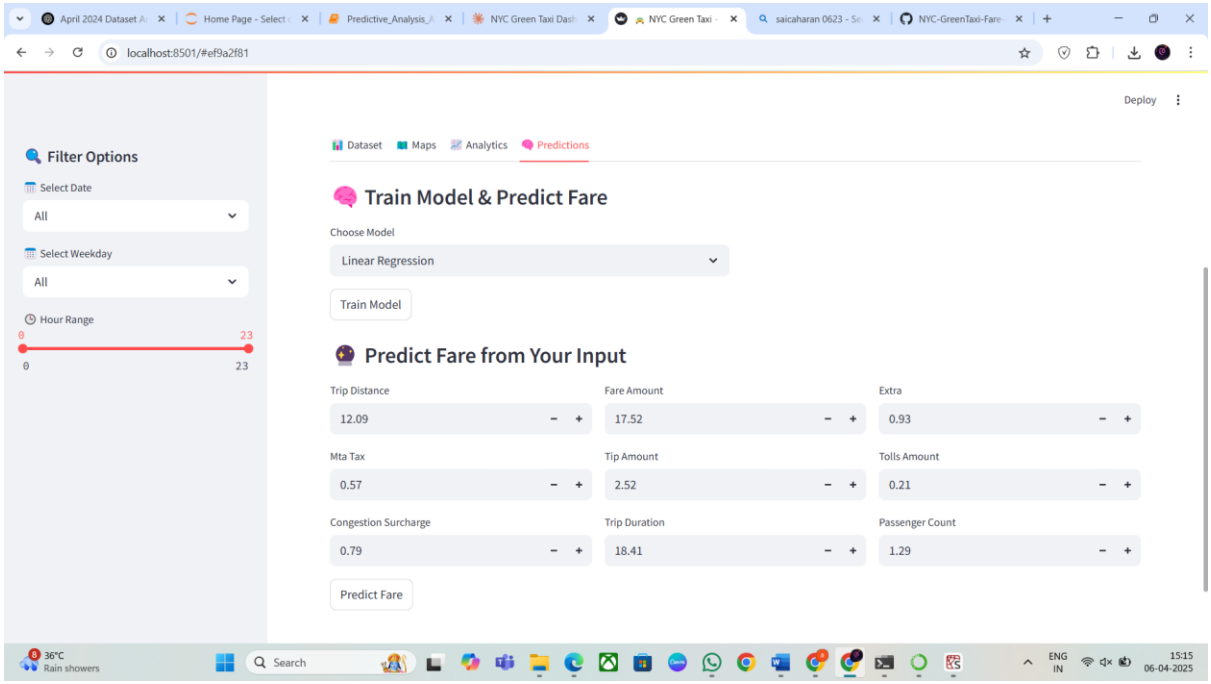
- R2 Score: ~0.82
- RMSE: Low, indicating good performance.

9. Streamlit App Features

- Built with Streamlit and integrated into Spyder and VS Code.
- Allows input of custom trip details:
    - Distance
    - Passenger Count
    - Pickup Hour
    - Day of Week
- Predicts taxi fare instantly.
- User-friendly layout with sidebar and result display.

Live App: https://nyc-taxi-fare-prediction-using-machine-learning---april-2024-b.streamlit.app/


localhost:8501/#ef9a2f81
Deploy
Filter Options
Select Date
All
Select Weekday
All
Hour Range
0
23
0
23
NYC Green Taxi Dashboard – April 2024
Upload April 2024 Parquet File
Drag and drop file here
Limit 200MB per file • PARQUET
Browse files
green_tripdata_2024-04.parquet 1.3MB
Key Metrics
Total Trips
56,471
Avg Fare
$17.52
Avg Trip Distance
12.09 mi
Avg Tip
$2.52
Dataset
Maps
Analytics
Predictions
Dataset Preview
VendorID
lpep_pickup_datetime
lpep_dropoff_datetime
store_and_fwd_flag
RatecodeID
PULocationID
DOLocationID
passenger_count
trip_distance
0
2
2024-04-01 00:18:50
2024-04-01 00:19:48
N
1
146
146
1
0.15
1
2
2024-04-01 00:56:16
2024-04-01 01:12:56
N
1
65
225
1
3.06
36°C
Rain showers
Search
ENG
IN
15:14
06-04-2025


localhost:8501/#ef9a2f81
Deploy
Filter Options
Select Date
All
Select Weekday
All
Hour Range
0
23
0
23
Dataset
Maps
Analytics
Predictions
Train Model & Predict Fare
Choose Model
Linear Regression
Train Model
Predict Fare from Your Input
Trip Distance
12.09
Fare Amount
17.52
Extra
0.93
Mta Tax
0.57
Tip Amount
2.52
Tolls Amount
0.21
Congestion Surcharge
0.79
Trip Duration
18.41
Passenger Count
1.29
Predict Fare
36°C
Rain showers
Search
ENG
IN
15:15
06-04-2025

10. Learnings

- Worked with real-world data in Parquet format.
- Implemented preprocessing and feature engineering.
- Built a machine learning model and visualized results.
- Deployed a live interactive ML app using Streamlit.

11. Conclusion

- This project demonstrates the power of machine learning in making fare predictions in a transparent and interpretable way. Streamlit makes the model accessible, easy to interact with, and useful in real-world application.
- The deployment of the model via a Streamlit application makes it usable for business users, students, and city planners alike.

12. References

- NYC TLC Trip Record Data
- Streamlit Documentation
- Scikit-learn API
- Pandas, Matplotlib, Seaborn libraries