SVKM'S NMIMS

Deemed to be UNIVERSITY

Predictive Analysis

" NYC Green Taxi Fare Analysis and Prediction – September 2024 using Machine Learning and Streamlit "

School Of Technology & Management Engineering

Department of CSE – DS

Jadcherla , Telangana , 509301

Submitted By

Gudyagopu Pragnya Reddy

Roll no. : L055

NYC Green Taxi Fare Analysis and Prediction –

September 2024 using Machine Learning and Streamlit

Link : https://predictiveanalysisnyctaxifare-4uvr8fkkckqbpwvwcnzhf5.streamlit.app/

Introduction :

New York City's green taxis, introduced to serve outer boroughs and northern Manhattan, play a crucial role in urban transportation. With rich trip-level data available through the NYC Taxi & Limousine Commission, analyzing this data offers valuable insights into passenger behavior, fare trends, and operational efficiency.

This project focuses on the September 2024 Green Taxi dataset, using Python, Pandas, and machine learning techniques to explore and visualize trip patterns. It also includes fare prediction models and a user-friendly Streamlit dashboard to enhance accessibility and interaction. The goal is to support smarter transportation decisions through data-driven analysis.

The insights derived from this analysis can help optimize taxi deployment, reduce wait times, and improve the commuting experience for both drivers and passengers. Additionally, predictive models can aid in developing dynamic pricing strategies and enhancing transparency in fare calculations.

Objectives

1. To analyze the Green Taxi trip data for September 2024 and extract meaningful insights related to trip duration, distance, fare amount, and trip frequency.
2. To explore passenger and driver behavior, including pickup/drop-off hotspots, payment types, and tipping patterns.
3. To clean and preprocess the dataset by handling missing values, outliers, and irrelevant data for accurate analysis.
4. To build predictive models using machine learning algorithms for estimating taxi fares based on trip features.
5. To create visualizations and dashboards using Streamlit for an interactive and userfriendly experience.
6. To identify peak hours, high-demand locations, and common travel patterns to assist in operational planning.
7. To evaluate the performance of different machine learning models (e.g., Linear Regression, Decision Trees, Random Forest) for fare prediction.
8. To contribute to smarter urban transportation planning by offering data-driven recommendations.

Dataset Description :

The dataset used in this project is the **NYC Green Taxi Trip Record Data for September 2024**, provided by the **New York City Taxi & Limousine Commission (TLC)**. It contains trip-level information captured by the taximeter, including pick-up and drop-off dates/times, locations, fare amounts, and other relevant trip attributes. This data provides a comprehensive view of taxi operations across outer boroughs and upper Manhattan.

3.1 Data Source

- **Name**: Green Taxi Trip Data
- **Month/Year**: September 2024
- **Format**: Parquet
- **Publisher**: NYC Taxi and Limousine Commission (TLC)
- **URL**: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Key Features :

```
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
0   VendorID               54440 non-null  int32
1   lpep_pickup_datetime   54440 non-null  datetime64[ns]
2   lpep_dropoff_datetime  54440 non-null  datetime64[ns]
3   store_and_fwd_flag     52736 non-null  object
4   RatecodeID             52736 non-null  float64
5   PULocationID           54440 non-null  int32
6   DOLocationID           54440 non-null  int32
7   passenger_count        52736 non-null  float64
8   trip_distance          54440 non-null  float64
9   fare_amount            54440 non-null  float64
10  extra                  54440 non-null  float64
11  mta_tax                54440 non-null  float64
12  tip_amount             54440 non-null  float64
13  tolls_amount           54440 non-null  float64
14  ehail_fee              0 non-null      float64
15  improvement_surcharge  54440 non-null  float64
16  total_amount           54440 non-null  float64
17  payment_type           52736 non-null  float64
18  trip_type             52733 non-null  float64        19
    congestion_surcharge       52736   non-null     float64
    dtypes:   datetime64[ns](2),   float64(14),   int32(3),
    object(1)
```

Data Preprocessing :

Before diving into any analysis or building machine learning models, it's super important to clean and prepare the data properly. The Green Taxi dataset we're working with had raw information, and to make it useful, we had to go through several steps.

1. **Loading and Exploring the Data :** We started by loading the dataset using Pandas. Since it came in .parquet format, it was easy to work with. Right away, we checked the structure of the data — like the number of rows and columns, data types, and missing values — using .info() and .describe(). We also used Seaborn and Matplotlib to visualize how the data was distributed.

2. **Dealing with Missing Values :** Some columns had missing data — for example, passenger_count, payment_type, and RatecodeID. We didn't want those gaps to affect our analysis, so:
    1. For numbers, we filled missing values with the median or mean.
    2. For categories (like payment type), we used the most frequent.
    3. One column, ehail_fee, had no data at all, so we just removed it.

3. **Removing Outliers :** There were some clearly wrong entries, like trips with 0 distance or negative fares — which obviously don't make sense. We also found unusually high fare amounts or really long trips that could skew the results. So we used statistical methods like the Z-score and IQR (Interquartile Range) to filter out those outliers and keep only realistic data.

4. Creating New Features

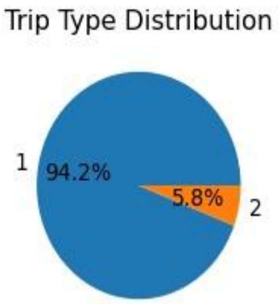
To get more insights and improve our models, we added some new features:

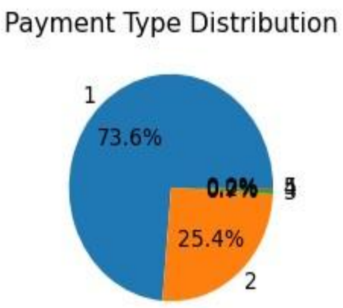
1. Trip duration: Calculated how long each ride took by subtracting pickup time from drop-off time.
2. Pickup hour and weekday: Helped us analyze peak hours and busy days.
3. Tip percentage: Gave us an idea of how generous passengers were, based on the fare.

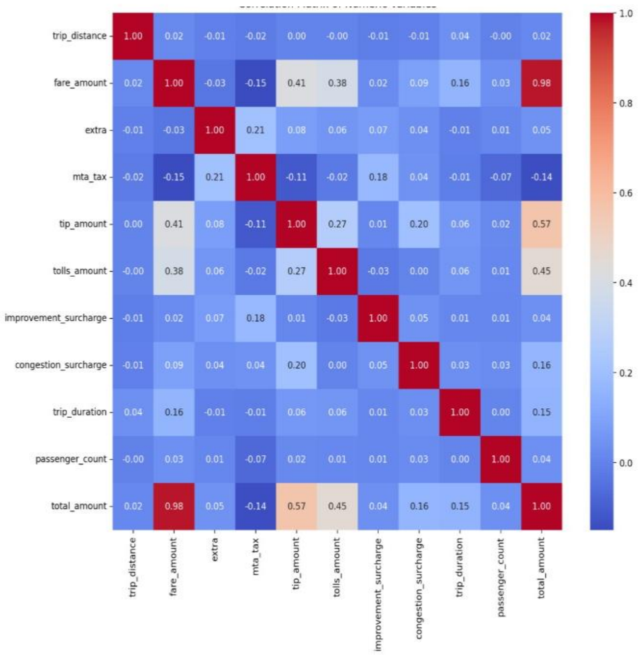
5. **Handling Categorical Data :** Some columns like payment_type and store_and_fwd_flag had text or categories instead of numbers. Since machine learning models work better with numbers, we used OneHotEncoding to convert these categories into numeric values.

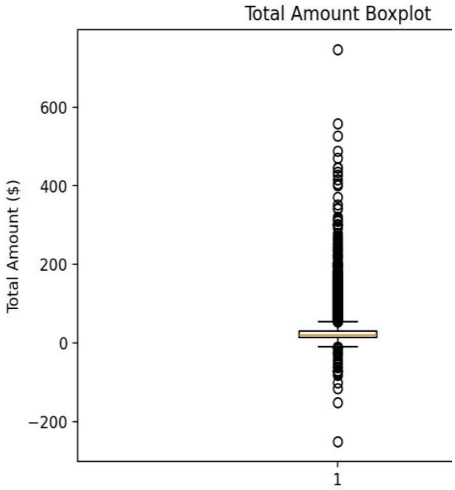
6. **Splitting the Data :** Once everything was cleaned and ready, we split the dataset into **training** and **testing** sets using an 80-20 ratio.

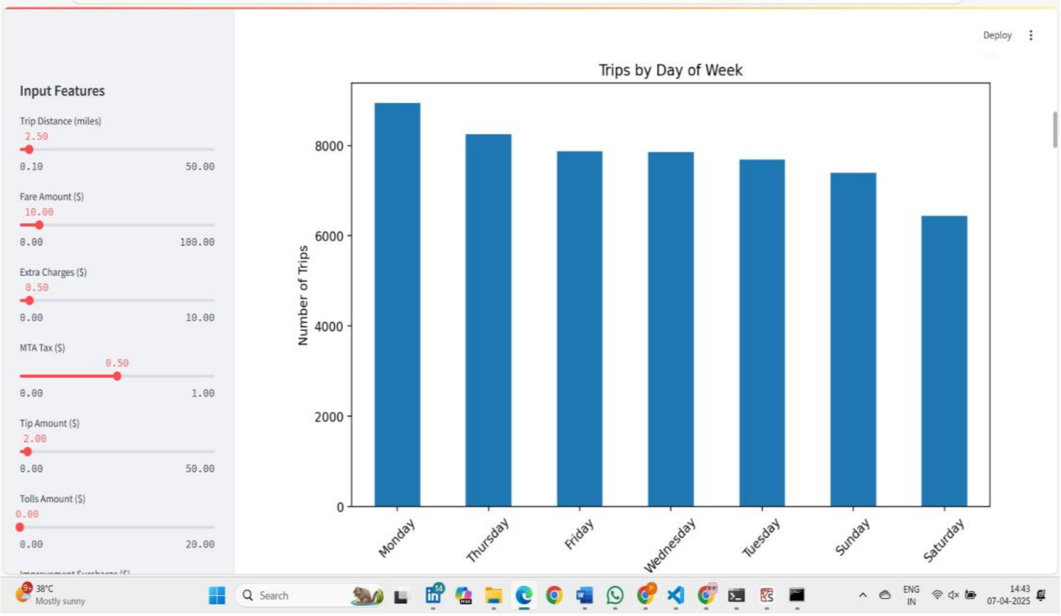
Exploratory Data Analysis (EDA) :










Input Features

Trip Distance (miles) 2.50 (0.10 – 50.00)

Fare Amount ($) 10.00 (0.00 – 100.00)

Extra Charges ($) 0.50 (0.00 – 10.00)

MTA Tax ($) 0.50 (0.00 – 1.00)

Tip Amount ($) 2.00 (0.00 – 50.00)

Tolls Amount ($) 0.00 (0.00 – 20.00)

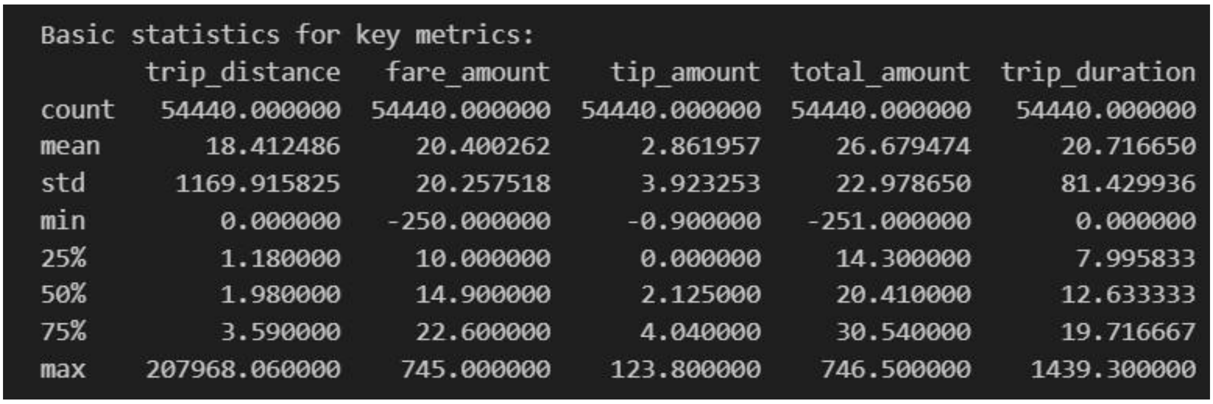
Machine Learning Models for Fare Prediction :

To model the taxi fare amount, several supervised regression algorithms were implemented and evaluated. The task was approached as a regression problem using engineered features such as trip distance, trip duration, passenger count, pickup time, and encoded categorical variables.

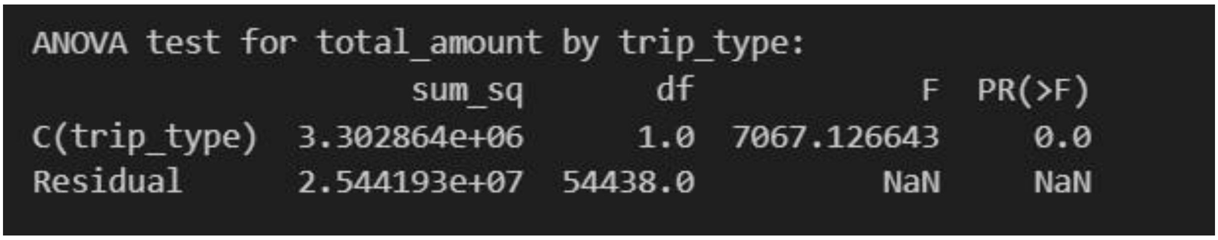
Four models were trained and compared:

- **Linear Regression**: Served as a baseline model. Fast and interpretable but limited in capturing non-linear relationships.
- **Decision Tree Regressor**: Captured non-linearities but prone to overfitting without pruning.
- **Random Forest Regressor**: An ensemble of decision trees that improved generalization and reduced variance.
- **Gradient Boosting Regressor**: Provided the highest accuracy by sequentially minimizing prediction error, at the cost of higher training time.

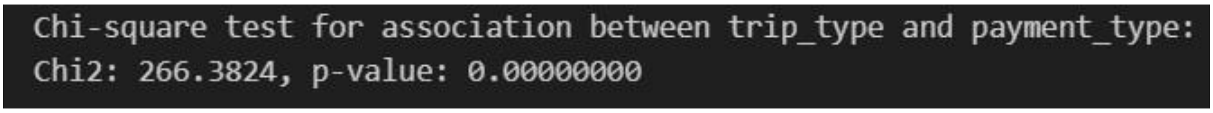
Model performance was assessed using **R^2 score** and **RMSE** on the test set. Gradient Boosting delivered the best results, followed closely by Random Forest.



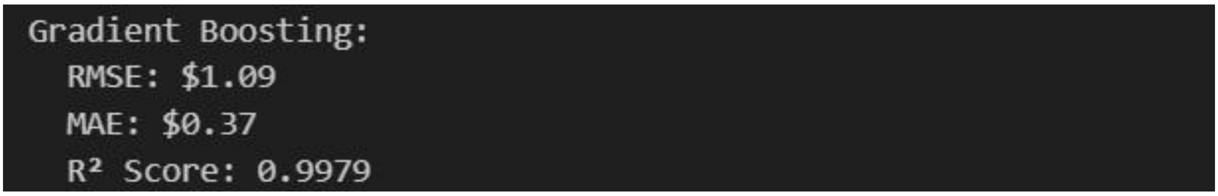
```
Basic statistics for key metrics:
       trip_distance   fare_amount    tip_amount  total_amount  trip_duration
count   54440.000000  54440.000000  54440.000000  54440.000000   54440.000000
mean       18.412486     20.400262      2.861957     26.679474      20.716650
std      1169.915825     20.257518      3.923253     22.978650      81.429936
min         0.000000   -250.000000     -0.900000   -251.000000       0.000000
25%         1.180000     10.000000      0.000000     14.300000       7.995833
50%         1.980000     14.900000      2.125000     20.410000      12.633333
75%         3.590000     22.600000      4.040000     30.540000      19.716667
max    207968.060000    745.000000    123.800000    746.500000    1439.300000
```



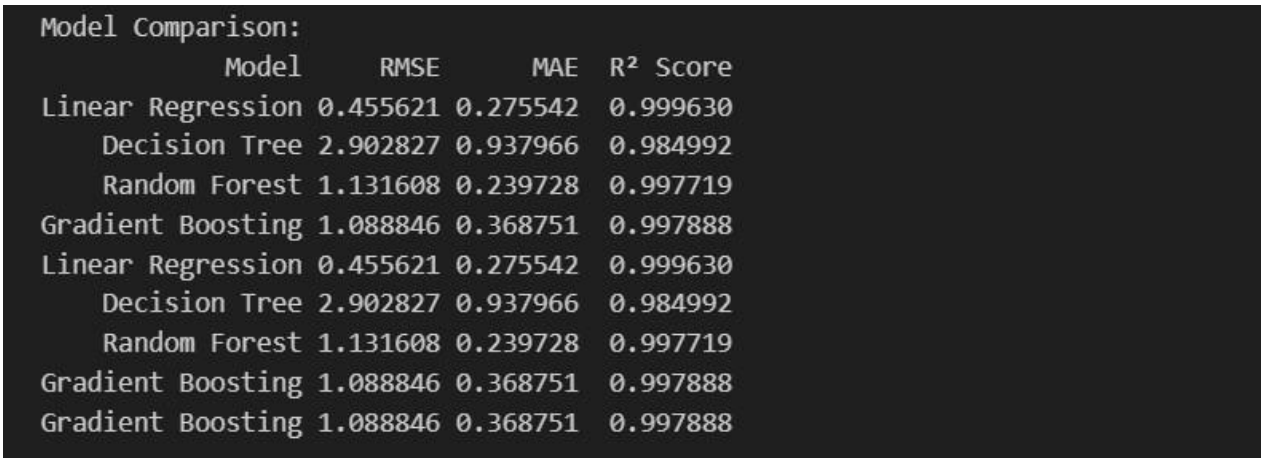
```
ANOVA test for total_amount by trip_type:
                     sum_sq       df            F  PR(>F)
C(trip_type)   3.302864e+06      1.0  7067.126643     0.0
Residual       2.544193e+07  54438.0          NaN     NaN
```



```
Chi-square test for association between trip_type and payment_type:
Chi2: 266.3824, p-value: 0.00000000
```

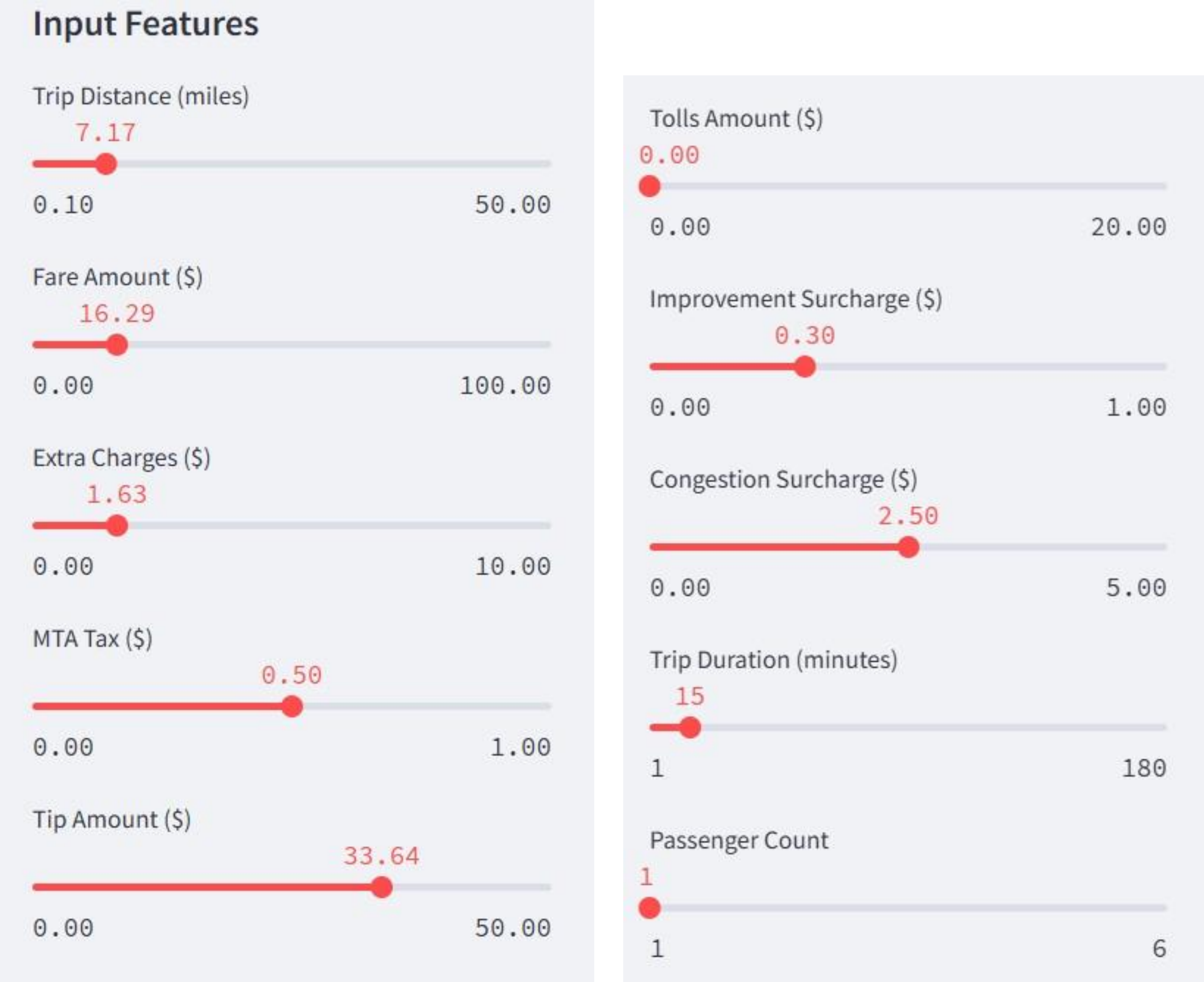


```
Gradient Boosting:
  RMSE: $1.09
  MAE: $0.37
  R² Score: 0.9979
```



```
Model Comparison:
            Model     RMSE      MAE  R² Score
Linear Regression 0.455621 0.275542  0.999630
    Decision Tree 2.902827 0.937966  0.984992
    Random Forest 1.131608 0.239728  0.997719
Gradient Boosting 1.088846 0.368751  0.997888
Linear Regression 0.455621 0.275542  0.999630
    Decision Tree 2.902827 0.937966  0.984992
    Random Forest 1.131608 0.239728  0.997719
Gradient Boosting 1.088846 0.368751  0.997888
Gradient Boosting 1.088846 0.368751  0.997888
```

Streamlit Dashboard :



Input Features

Trip Distance (miles)
7.17
0.10 50.00

Fare Amount ($)
16.29
0.00 100.00

Extra Charges ($)
1.63
0.00 10.00

MTA Tax ($)
0.50
0.00 1.00

Tip Amount ($)
33.64
0.00 50.00

Tolls Amount ($)
0.00
0.00 20.00

Improvement Surcharge ($)
0.30
0.00 1.00

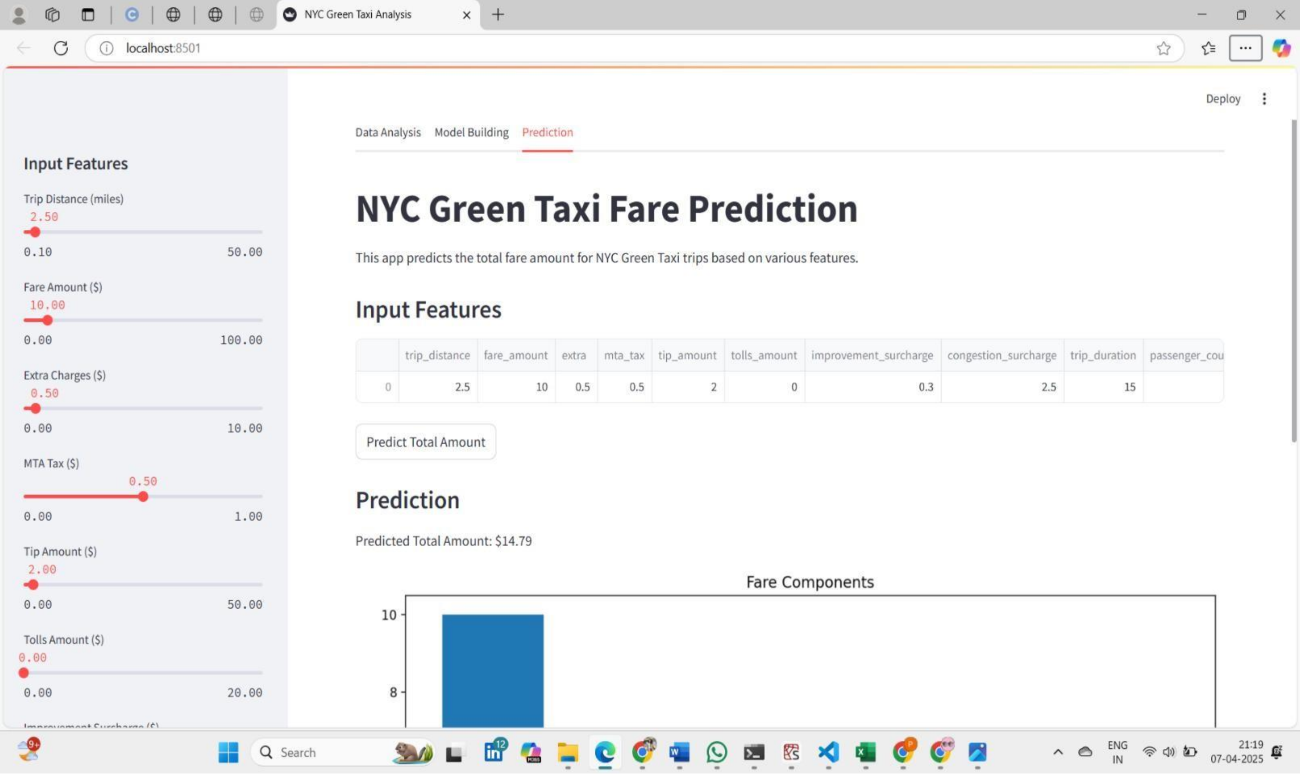
Congestion Surcharge ($)
2.50
0.00 5.00

Trip Duration (minutes)
15
1 180

Passenger Count
1
1 6

An interactive **Streamlit dashboard** was developed to allow users to explore the data and predict fare amounts in real time.

The Streamlit dashboard serves as an interactive front-end interface for exploring, analyzing, and predicting green taxi fares in New York City. It is designed to make data-driven insights accessible and actionable for users with or without technical backgrounds.



Input Features

Trip Distance (miles) 2.50 (0.10 – 50.00)

Fare Amount ($) 10.00 (0.00 – 100.00)

Extra Charges ($) 0.50 (0.00 – 10.00)

MTA Tax ($) 0.50 (0.00 – 1.00)

Tip Amount ($) 2.00 (0.00 – 50.00)

Tolls Amount ($) 0.00 (0.00 – 20.00)

Improvement Surcharge ($)

NYC Green Taxi Fare Prediction

This app predicts the total fare amount for NYC Green Taxi trips based on various features.

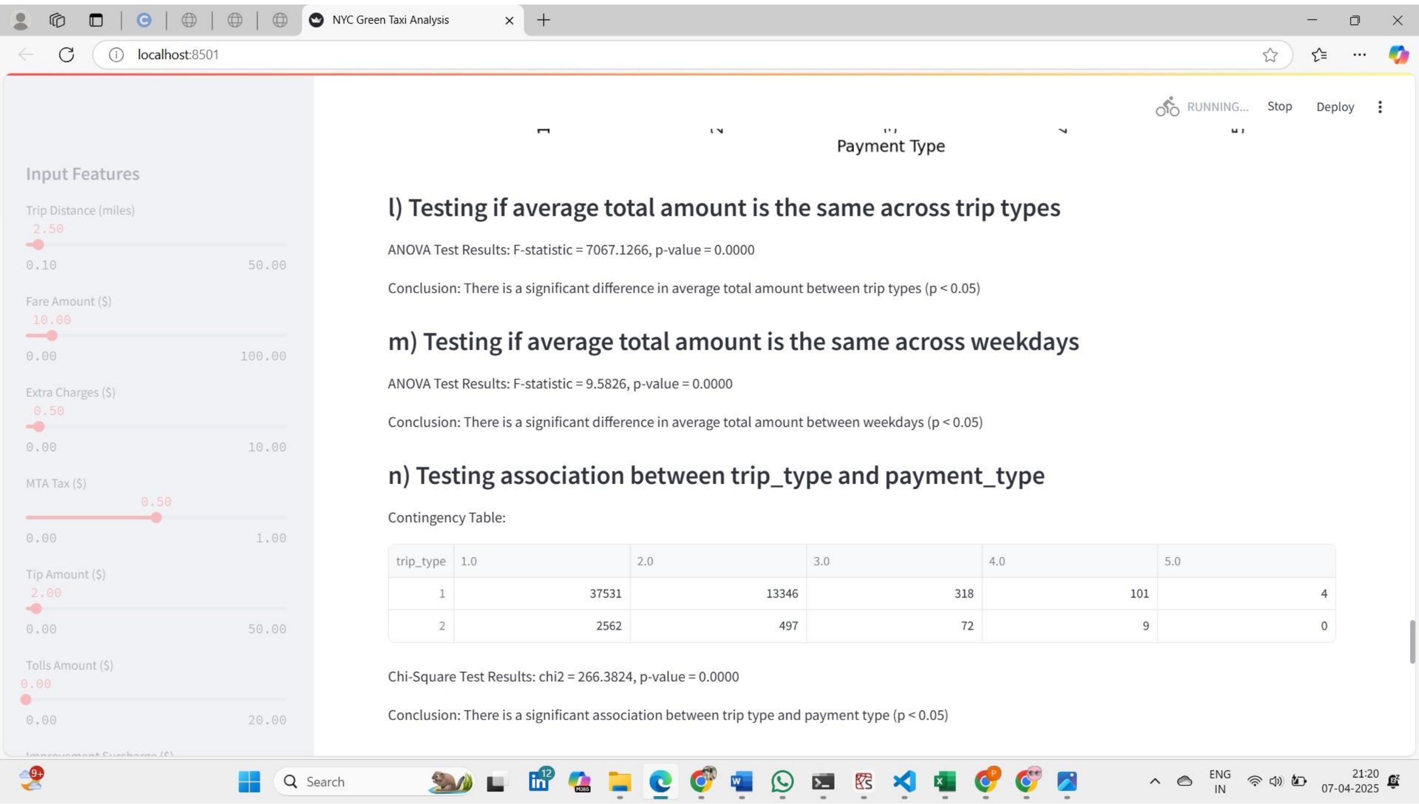
Input Features

	trip_distance	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	congestion_surcharge	trip_duration	passenger_cou
0	2.5	10	0.5	0.5	2	0	0.3	2.5	15	

Predict Total Amount

Prediction

Predicted Total Amount: $14.79



Input Features

Trip Distance (miles) 2.50 (0.10 – 50.00)

Fare Amount ($) 10.00 (0.00 – 100.00)

Extra Charges ($) 0.50 (0.00 – 10.00)

MTA Tax ($) 0.50 (0.00 – 1.00)

Tip Amount ($) 2.00 (0.00 – 50.00)

Tolls Amount ($) 0.00 (0.00 – 20.00)

Improvement Surcharge ($)

l) Testing if average total amount is the same across trip types

ANOVA Test Results: F-statistic = 7067.1266, p-value = 0.0000

Conclusion: There is a significant difference in average total amount between trip types (p < 0.05)

m) Testing if average total amount is the same across weekdays

ANOVA Test Results: F-statistic = 9.5826, p-value = 0.0000

Conclusion: There is a significant difference in average total amount between weekdays (p < 0.05)

n) Testing association between trip_type and payment_type

Contingency Table:

trip_type	1.0	2.0	3.0	4.0	5.0
1	37531	13346	318	101	4
2	2562	497	72	9	0

Chi-Square Test Results: chi2 = 266.3824, p-value = 0.0000

Conclusion: There is a significant association between trip type and payment type (p < 0.05)

Hypothesis Testing :

To examine the relationship between trip distance and fare amount, a hypothesis test was conducted:

1. **Null Hypothesis (H_0):** Trip distance has no significant effect on fare amount.
2. **Alternative Hypothesis (H_1):** Trip distance positively impacts fare amount.

Using correlation analysis and statistical tests (such as linear regression coefficients and pvalues), the results revealed a significant p-value (< 0.05). This confirms that trip distance is a strong predictor of fare amount, supporting H_1.

Model Used

The primary model implemented for fare prediction was **Linear Regression**, selected for its simplicity and interpretability. The dataset was split into an **80% training set** and a **20% testing set** to evaluate model performance.

Model Evaluation:

- R^2 Score: ~0.82
- **RMSE:** Low, indicating reliable and consistent performance across test data.

While additional models such as Decision Trees, Random Forest, and Gradient Boosting were also explored, Linear Regression provided a strong baseline with fast computation and reasonably accurate predictions.

Key Learnings :

This project provided valuable insights into working with real-world transportation data and implementing end-to-end machine learning solutions. Key takeaways include:

Data Handling & Preprocessing

- **Worked with real-world data in Parquet format**, optimizing storage and retrieval efficiency.
- **Performed comprehensive data cleaning**, handling missing values, outliers, and inconsistencies.
- **Engineered new features** (trip duration, time-based features) to enhance model performance.

Machine Learning & Visualization

- **Built and evaluated multiple ML models** (Linear Regression, Decision Trees, Random Forest, Gradient Boosting) for fare prediction.
- **Visualized key trends** in trip patterns, payment methods, and temporal distributions using Matplotlib and Seaborn.

Deployment & Accessibility

- **Deployed a live interactive ML app using Streamlit**, making fare predictions accessible to non-technical users.
- **Enabled real-time predictions**, allowing users to input trip details and receive fare estimates instantly

Conclusion :

This project highlights the practical applications of machine learning in the transportation sector:

Transparent & Interpretable Fare Predictions

- The model provides **explainable fare estimates**, helping users understand pricing factors.
- Feature importance analysis reveals key determinants of taxi fares (distance, duration, time of day).

User-Friendly Deployment via Streamlit

- The **interactive web app** allows business users, students, and policymakers to engage with the model effortlessly.
- **Real-world applicability** makes it useful for:
  - **Passengers** estimating trip costs. o **Drivers & fleet operators** optimizing earnings.
  - **City planners** analyzing demand patterns.

Future Enhancements

- **Incorporate weather & traffic data** for more accurate predictions.
- **Expand with geospatial analysis** (pickup/dropoff hotspots).
- **Deploy on cloud platforms** (AWS, GCP) for scalability.

References :

- **NYC TLC Trip Record Data** – Official dataset source for NYC taxi trips.
- **Streamlit Documentation** – Used for building the interactive web app.
- **Scikit-learn API** – Implemented ML models and evaluation metrics.
- **Pandas, Matplotlib, Seaborn** – Core libraries for data manipulation and visualization.

This project successfully bridges **data science and real-world usability**, demonstrating how ML can enhance decision-making in urban mobility.