Predictive Analysis

NYC Green Taxi Fare Prediction - Dec 2020 - with Streamlit App

Submitted By:

Malde Saicharan: [70572200033]

Submitted To: Prof. Rajesh Prabhakar

SVKM’S NMIMS HYDERABAD

Project Overview

This project is centered around analyzing and predicting taxi fare amounts for **New York City's Green Taxi services Dec 2020**. Using real-world trip data, the aim is to explore patterns, derive insights, and build accurate machine learning models that can estimate the fare based on trip-related variables.

The project also includes building a **Streamlit-based web application** to make the predictive model and visualizations accessible and interactive for users.

Objective

- To **analyze patterns** in NYC taxi rides using various features like time, day, distance, and trip type.
- To conduct **exploratory data analysis** (EDA) and derive insights.
- To perform **hypothesis testing** to verify assumptions.
- To apply **machine learning algorithms** for fare prediction.
- To build an **interactive web app** for visualizing data and predicting taxi fares.

Data Description

The dataset used comes from the **New York City Taxi and Limousine Commission (TLC)** and includes millions of records collected from NYC Green Taxis.

Each row represents a single taxi trip and contains attributes such as:

- **Pickup and drop-off times**
- **Trip distance**
- **Passenger count**
- **Payment type**
- **Trip type (e.g., street-hail or dispatched)**
- **Fare details**: base fare, tips, taxes, surcharges, total amount

Data Preprocessing

Before analysis and modeling, the dataset was cleaned and prepared:

- **Dropped irrelevant or unused columns** (e.g., columns with constant or null values).
- **Converted datetime columns** to extract meaningful features such as trip duration, day of the week, and hour of the day.
- **Handled missing values** using appropriate imputation methods.
- **Detected and handled outliers** in features like `trip_distance` and `total_amount`.

Feature Engineering

Exploratory Data Analysis (EDA)

EDA was conducted to visualize patterns and understand the relationships between features.

Key findings include:

- **Peak travel hours** typically occur during mornings and evenings.
- **Trip type** affects fare amounts – dispatched trips tend to have higher fares.
- **Payment methods** influence tips – card payments result in higher tipping.
- **Weekends** show different fare and tip patterns compared to weekdays.
- Most passengers are **solo travelers** or groups of two.

These insights helped validate the assumptions and guide feature selection for modeling.

Hypothesis Testing

Statistical tests were conducted to validate patterns in the dataset and ensure that observed trends were not random. Two key tests were performed:

1. ANOVA (Analysis of Variance)

ANOVA was used to check if the **average fare amount** differed significantly across:

- **Days of the week** - **Trip types**

Hypotheses:

- **Null Hypothesis (H_0):** No significant difference in average fare among groups.
- **Alternative Hypothesis (H_1):** At least one group has a different average fare.

Findings:

- The p-value was below 0.05, indicating a significant difference in fares across both trip types and weekdays.
- This suggests that these variables influence fare amounts and should be included in predictive modeling.

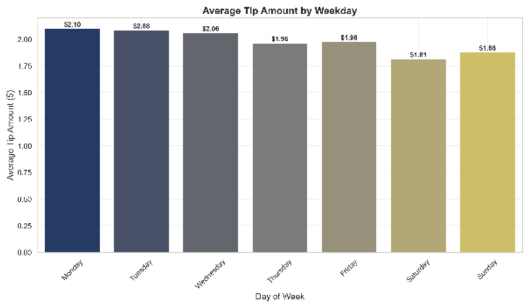
2. Chi-Square Test of Independence

This test was applied to examine the relationship between **trip type** and **payment type**.

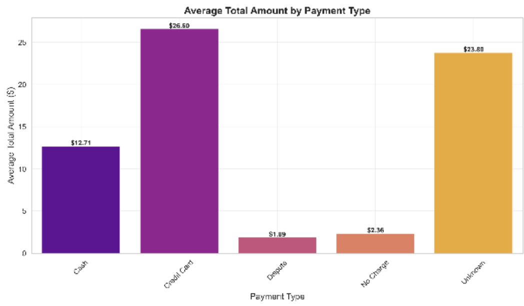
Hypotheses:

- **Null Hypothesis (H_0):** Trip type and payment type are independent.
- **Alternative Hypothesis (H_1):** There is an association between trip type and payment type.

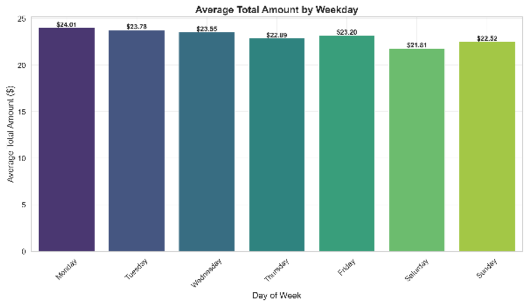
Observation's & Graphs



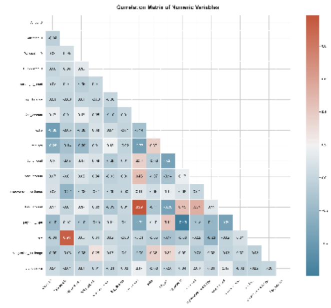
avg_tip_by_weekday.png



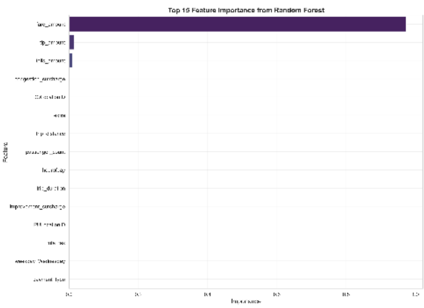
avg_total_by_payment.png



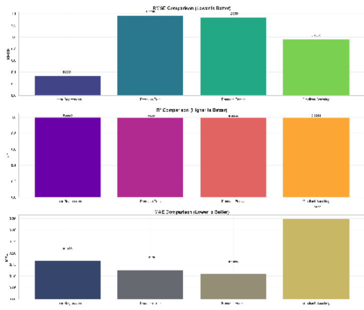
avg_total_by_weekday.png



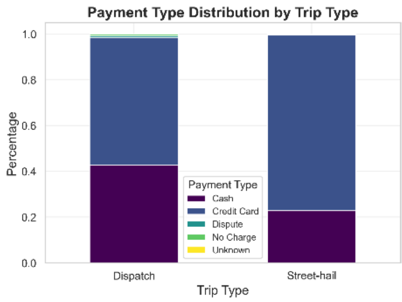
correlation_matrix.png



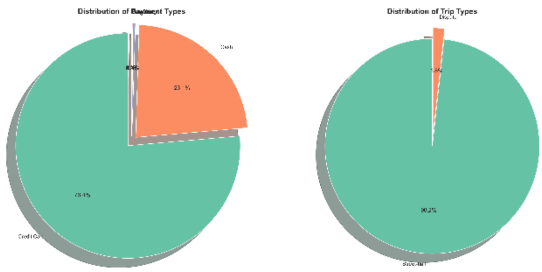
feature_importance.png



model_comparison.png



payment_by_trip_type.png



payment_trip_type_distribution.png

Correlation Analysis

Correlation between numeric features was analysed to identify:

- Strong positive or negative relationships
- Multicollinearity that may affect model accuracy

It was observed that:

- **Fare amount** strongly correlates with **trip distance**
- **Tip amount** correlates with both fare and payment method

These helped in refining feature selection and improving the interpretability of the model.

Machine Learning Models

Various regression algorithms were applied to predict the **total fare amount**:

Models Used:

- **Multiple Linear Regression**: Assumes linear relationship between features and target.
- **Decision Tree Regression**: Non-linear model that splits data into branches for prediction.
- **Random Forest Regression**: Ensemble of decision trees to reduce overfitting and improve accuracy.
- **Gradient Boosting Regression**: Advanced boosting technique for high performance.

Evaluation Metrics:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **R^2 Score**

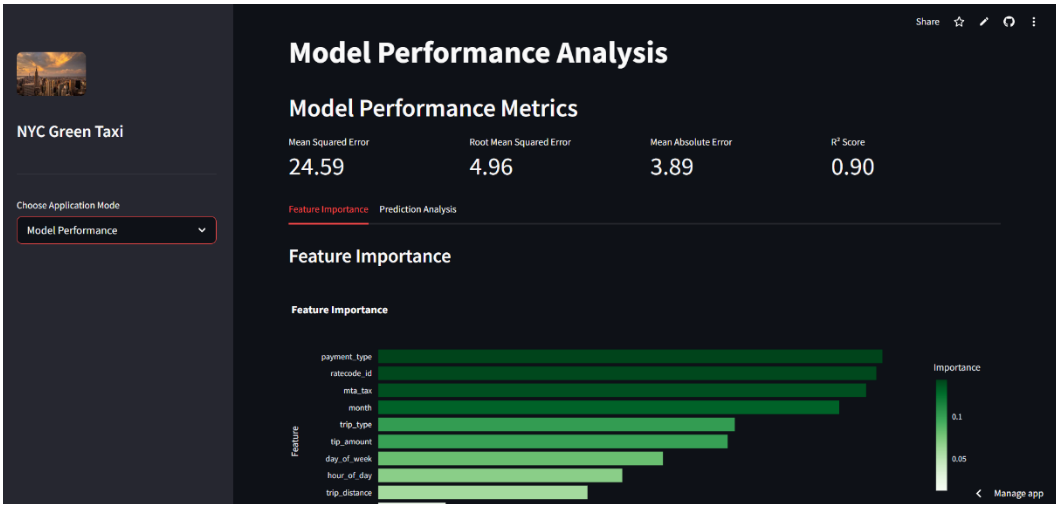
The best-performing model was the **Random Forest Regressor**, offering the most accurate and robust fare predictions.

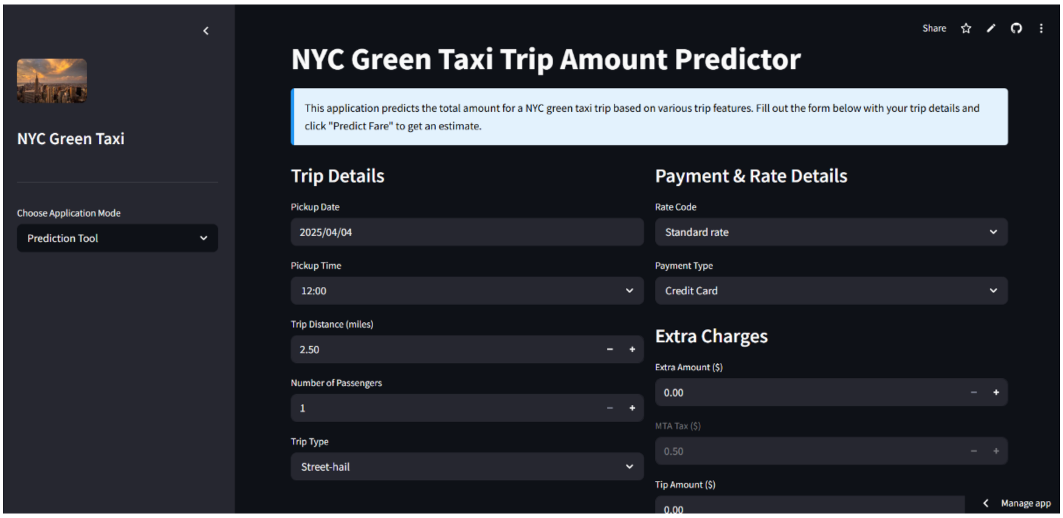
Streamlit Web Application

To make the project interactive and accessible, a **Streamlit web application** was developed. The app allows users to:

- Visualize key data insights (e.g., trip distribution by hour/day)
- Select features and predict taxi fares in real-time
- Explore feature importance and model evaluation results

Live App: https://nyc-greentaxi-fare-prediction.streamlit.app





Key Learnings

- Understanding how **time, trip type, and distance** affect taxi fares.
- Gaining hands-on experience in **EDA, hypothesis testing, and regression modeling**.
- Building a **complete data science pipeline** from raw data to deployed application.
- Learning the importance of **data cleaning** and **feature engineering** in real-world datasets.

Conclusion

This project successfully demonstrates how machine learning can be used to predict taxi fares using real-world transportation data. It combines **data preprocessing, visualization, statistical analysis**, and **regression modelling** to generate actionable insights and accurate predictions.

The deployment of the model via a Streamlit application makes it usable for business users, students, and city planners alike.