Predictive Analysis

NYC Green Taxi Fare Prediction - Feb 2022 - with Streamlit App

Submitted By:

Ananya P: [70572200052]

Submitted To: Prof. Rajesh Prabhakar

SVKM'S NMIMS HYDERABAD

Project Overview

This project focuses on predicting fare amounts for NYC Green Taxi rides using real-world data from February 2022. It employs data cleaning, feature engineering, exploratory data analysis, statistical testing, and machine learning techniques. The end product is a deployed Streamlit web application that offers real-time fare predictions and insights into ride behavior across New York City.

Objective

- Analyze taxi ride patterns using features like time, distance, and trip type
- Perform Exploratory Data Analysis (EDA)
- Apply hypothesis testing to validate assumptions
- Develop regression models to predict fare amounts
- Deploy an interactive Streamlit web application for users

Data Description

The dataset is from the **NYC Taxi and Limousine Commission (TLC)**, specifically the **Green Taxi trip data for February 2022**. It contains millions of ride records, with features such as:

- Pickup & drop-off timestamps
- Trip distance
- Fare components (base fare, surcharges, tips, tolls)
- Payment and trip types
- Passenger count

Data Preprocessing

- Dropped columns with missing or constant values (`ehail_fee`, etc.)
- Converted timestamps to derive trip duration, hour of day, and weekday
- Imputed missing numerical values with medians; categorical with "Unknown"
- Removed or adjusted outliers in fields like `trip_distance` and `total_amount`

Feature Engineering

- **Trip Duration**: Calculated in minutes
- **Weekday and Hour**: Extracted from drop-off time
- **One-Hot Encoding**: Applied to categorical features such as `payment_type`, `trip_type`, `weekday`, etc.

These features improved model accuracy and revealed user behavior patterns across different timeframes and ride conditions.

Exploratory Data Analysis (EDA)

EDA was performed to investigate underlying patterns in the dataset and support hypothesis generation. A variety of visualizations and summary statistics were used to better understand data distribution, identify outliers, and observe feature interactions.

Key Findings:

- **Peak travel hours** occur primarily during **morning (7–10 AM)** and **evening (4–7 PM)** rush hours.
- **Dispatched trips** tend to have **higher average fares** compared to street-hail rides.
- **Credit card payments** result in **higher average tips**, suggesting a behavioral trend in tipping habits.
- **Weekend trips** show different fare and tip structures than weekdays, likely influenced by leisure travel.
- Most trips are taken by **solo travelers or small groups (1–2 passengers)**.

Hypothesis Testing

To validate some of the patterns observed during EDA and to ensure they were not the result of random chance, two key statistical tests were performed: **ANOVA (Analysis of Variance)** and the **Chi-Square Test of Independence**. These tests helped establish whether certain categorical features had a statistically significant impact on fare-related variables.

1. ANOVA (Analysis of Variance)

ANOVA was used to determine whether the **average total fare amount** differed significantly across various groups—specifically:

- **Days of the Week (Monday through Sunday)**
- **Trip Type (Street-hail vs. Dispatched)**

Hypotheses:

- **Null Hypothesis (H_0):** The mean fare amount is the same across all categories (i.e., day of the week or trip type has no impact).
- **Alternative Hypothesis (H_1):** At least one category has a significantly different mean fare.

Results and Interpretation:

The ANOVA test yielded a **p-value less than 0.05**, which led to the rejection of the null hypothesis. This implies that:

- There are significant differences in **average fare amounts on different days of the week**.
- The **type of trip** (dispatched vs. street-hail) significantly affects the total fare.

These findings confirm that both day and trip type are important factors influencing taxi fares and should therefore be included in the predictive modeling process.

---

2. Chi-Square Test of Independence

This test was employed to analyze the **relationship between trip type and payment method**—two categorical variables.

Hypotheses:

- **Null Hypothesis (H_0):** Trip type and payment method are independent of each other (i.e., the mode of payment does not vary with trip type).
- **Alternative Hypothesis (H_1):** There is a significant association between trip type and payment method.

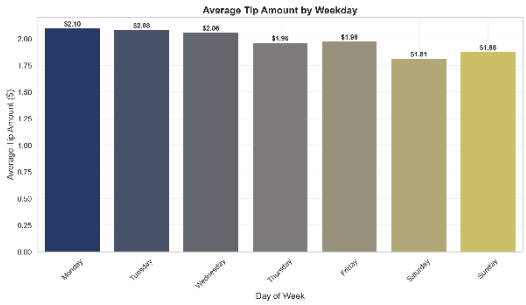
Results and Interpretation:

The Chi-Square test produced a statistically significant result, indicating a **dependence** between trip type and payment method. In other words:

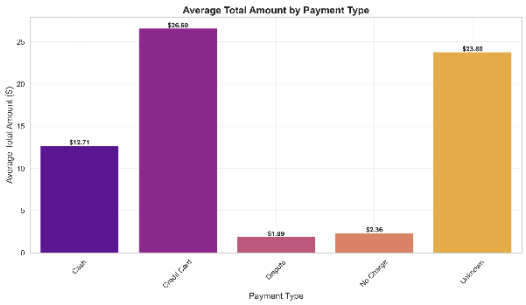
- **Certain payment methods** (like credit cards) are used more frequently with **dispatched trips** compared to street-hails.
- This association implies user behavior varies based on how the taxi was booked, reinforcing the relevance of both variables in modeling and analysis.

These statistical validations reinforced assumptions made during EDA and provided confidence in using these features in model development

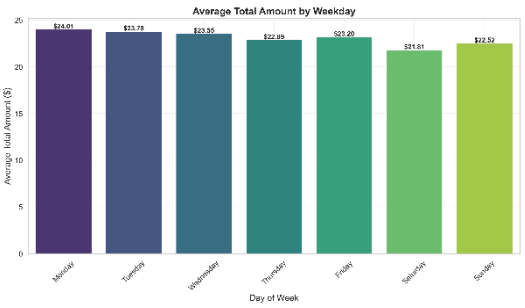
Observation's & Graphs



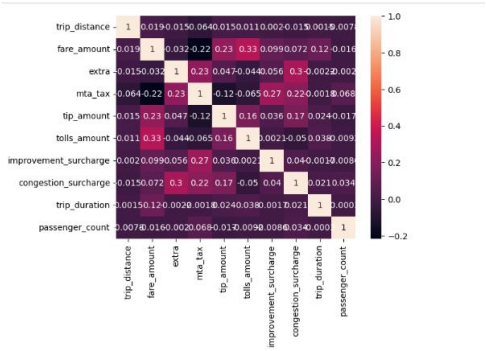
avg_tip_by_weekday.png



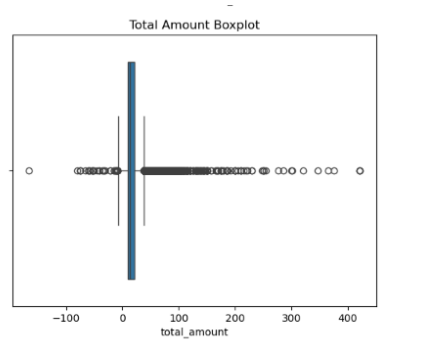
avg_total_by_payment.png



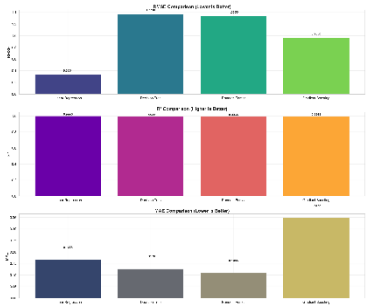
avg_total_by_weekday.png



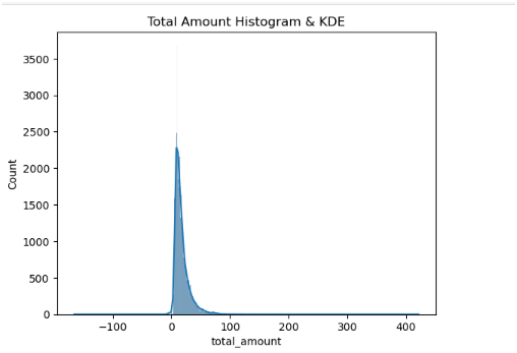
correlation_matrix.png



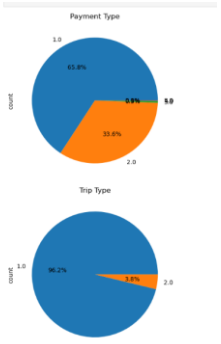
Total_amount.png



model_comparison.png



Total_amount.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:

After preprocessing and analysis, various regression models were developed to predict the **total fare amount** of taxi trips. These models aimed to learn the complex relationships between features such as trip distance, surcharges, time of day, and trip type, and the final amount paid by the customer.

Models Implemented:

- **Multiple Linear Regression:** Served as a baseline, assuming a linear relationship between predictors and the target.
- **Decision Tree Regression:** Captures non-linear relationships by splitting data into decision nodes based on feature thresholds.
- **Random Forest Regression:** An ensemble model that combines multiple decision trees to reduce overfitting and improve generalization.
- **Gradient Boosting Regression:** A sequential ensemble model that focuses on correcting errors made by previous models, offering strong predictive accuracy.

Model Evaluation:

To assess the models' performance, the following evaluation metrics were used:

- **Root Mean Squared Error (RMSE):** Measures the model's prediction error magnitude.
- **Mean Absolute Error (MAE):** Indicates the average absolute difference between predicted and actual values.
- **R^2 Score:** Represents the proportion of variance in the target variable that is predictable from the features.

Outcome:

Among all the models, the **Random Forest Regressor** delivered the best performance in terms of accuracy and stability. It effectively handled the feature complexity and interactions present in the dataset.

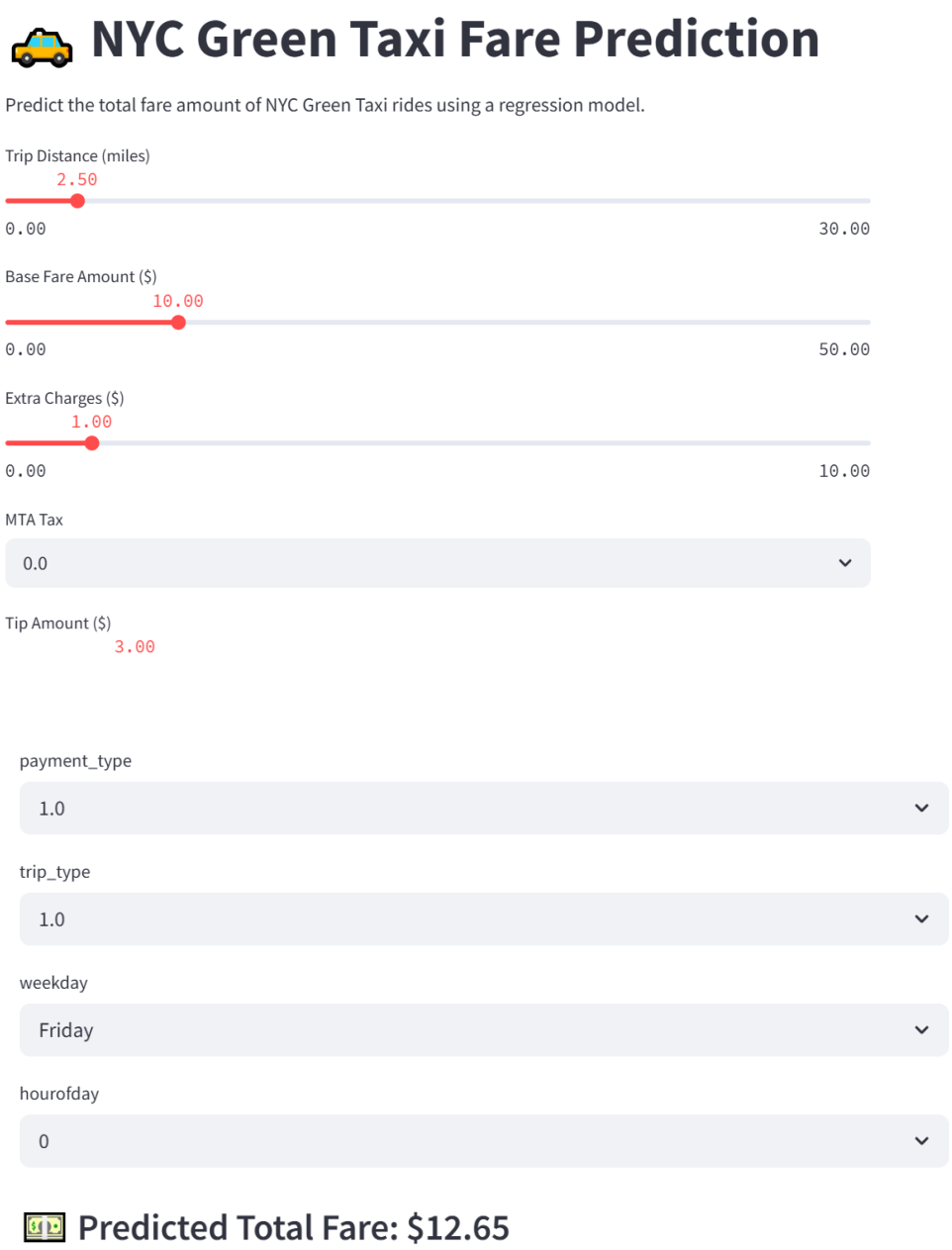
Streamlit Web Application

To ensure the project was not only technically sound but also usable and interactive, a **Streamlit web application** was developed. This app allows users to input trip details—such as distance, tip amount, surcharges, time, and trip type—and instantly get a fare prediction based on the trained machine learning model.

Features of the App:

- User-friendly sliders and dropdowns for inputting custom trip data
- Real-time prediction of total fare
- Visualization of key insights and feature importance
- Ability to explore how various inputs affect fare outcomes

- 🔗 **Live Application**: predictive-analysis-nyc-green-taxi-trips-data-analysis.streamlit.app



Key Learnings

Working on this project offered valuable insights into the end-to-end lifecycle of a real-world data science solution, from raw data handling to building an interactive application. Some of the most important learnings include:

- **Importance of Data Cleaning and Preparation:** Real-world datasets are often messy and inconsistent. Identifying and handling missing values, outliers, and irrelevant fields is critical to ensuring high-quality results.
- **Feature Engineering Makes a Difference:** Creating meaningful variables such as trip duration, weekday, and hour significantly improved model performance and interpretability.
- **EDA Reveals Actionable Insights:** Exploring patterns in taxi usage helped not only in understanding user behavior but also in guiding the selection of features for modeling. For example, card payments leading to higher tips and fare variations across weekdays were insightful findings.
- **Statistical Validation Strengthens Analysis:** Using ANOVA and Chi-Square tests provided empirical backing to trends observed in EDA, enhancing the credibility of the analysis.
- **Model Evaluation and Comparison Is Crucial:** Comparing different machine learning models highlighted the strengths of ensemble methods like Random Forest in handling complex datasets.
- **Deployment Brings Real Value:** The use of Streamlit enabled the transformation of a static analysis into a dynamic, user-facing tool. This made the project not only technically complete but also functionally impactful.

Conclusion

This project successfully demonstrates how data science techniques can be applied to solve practical problems in urban transportation. By leveraging NYC Green Taxi trip data, it was possible to uncover key patterns in ride behavior and develop an effective predictive model for estimating fare amounts.

From data acquisition and preprocessing to modeling and deployment, each phase contributed to building a robust system. The integration of a user-friendly Streamlit application further enhances the utility of the model by allowing users to explore insights and make predictions interactively.

Overall, the project reflects the power of combining exploratory analysis, statistical rigor, and machine learning to generate actionable solutions from complex real-world datasets. It stands as a comprehensive example of applied data science with tangible, real-world relevance