**PREDICTION ON TAXI TRAVEL TIMES IN NEW YORK CITY**

**Snehal Gaddime**

**Data Science Trainer**

**Edureka, Hyderabad**

**ABSTRACT**

Predicting taxi travel times in New York City (NYC) is essential for optimizing urban transportation. This abstract provides a concise overview of the data and methods used for NYC taxi time prediction.

Data sources include GPS data, historical trip records, and real-time traffic updates, offering comprehensive insights into NYC's dynamic transportation landscape. Key predictive factors encompass distance, time of day, weather, day of the week, and special events.

Machine learning algorithms, real-time updates, and historical averages constitute the prediction methods. Challenges include data quality assurance, traffic variability, and unexpected events. Accurate predictions empower passengers and service providers to make informed decisions, enhancing the efficiency and reliability of taxi services in NYC's bustling streets.

***Keywords : Exploratory Data Analysis, Feature Engineering, Model Training, Model Evaluation***

1. **PROBLEM STATEMENT**

To Build a machine learning model that predicts the duration of NYC taxi trip using the dataset which includes pickup time, geo-coordinates, the number of passengers, and several other variables.

1. **INTRODUCTION**

A regression model was developed to predict the duration of the taxi trip. The model was trained on a large dataset of over 1.5 million taxi trips, which were randomly split into training and testing sets. The features used in the regression model included distance, pickup and dropoff coordinates, pickup datetime, day of the week, and weather conditions such as temperature, precipitation, and wind speed.

The model was evaluated using various metrics such as Mean Square Error (MSE) and Root Mean Squared Error (RMSE), R2 Score, Adjusted R2-Score and was compared to other machine learning algorithms such as Linear Regression, Decision Tree Random Forest, Gradient Boosting and XGboost. The regression model outperformed the other algorithms in terms of accuracy, with an R2 score of 67%. Overall, the NYC Taxi Time Prediction project demonstrates the potential for regression models to accurately predict the duration of taxi trips in New York City, using a combination of various features such as location, time, and distance.

1. **DATASET CHALLENGES**

Dataset challenges encompass issues like data quality, with missing values, outliers, and inconsistencies, potentially skewing analyses. Data volume, whether too small or excessively large, can limit the depth and accuracy of insights. Data imbalance, noise, and ambiguity can hinder model training and interpretation, while privacy and security concerns demand responsible handling of sensitive information. Heterogeneous data from diverse sources and temporal variations require specialized pre-processing. Bias and fairness must be addressed to prevent unfair outcomes. Data preparation, access, and ethical considerations, along with domain-specific knowledge, play pivotal roles in overcoming these challenges. Proper data storage, management, scalability, and version control ensure the effective utilization of datasets for meaningful analyses and informed decision-making.

1. **FACTORS AFFECTING TAXI TIMINGS**
2. NYC taxi time predictions are influenced by multiple dynamic factors due to the city's complex urban environment.
3. Traffic congestion is a significant factor, especially during rush hours, affecting travel times.
4. Time of day plays a critical role, with distinct traffic patterns during peak and off-peak hours.
5. Weather conditions, such as rain, snow, or ice, can lead to slower traffic and longer travel times.
6. The day of the week can affect traffic, with weekdays often having heavier congestion than weekends.
7. Special events like parades or road closures can disrupt traffic flow and impact predictions.
8. Road infrastructure, including construction and road closures, can cause delays.
9. The specific pickup and drop-off locations within NYC can influence travel times.
10. The distance between locations is a fundamental factor in estimating taxi travel times.
11. The chosen route by the taxi driver may vary in efficiency, affecting travel duration.
12. Traffic signals and their timing at intersections can impact traffic flow and travel times.
13. The availability and reliability of public transportation options can affect taxi demand and travel times.
14. **STEPS INVOLVED**

**Step1: Exploratory Data Analysis**

In this project, the exploratory data analysis (EDA) was conducted on the NYC taxi time prediction dataset with no missing values present. Here's a summary of the key findings in 7 lines:

The dataset was clean, with no missing values, ensuring data integrity.

The distribution of passenger counts revealed that most trips had a low passenger count.

The pickup and dropoff times showed variation across different days of the week and hours of the day.

Trip duration had a wide range, with some outliers indicating unusually long trips.

The distance travelled by taxis exhibited variation, reflecting differences in trip lengths.

Vendor IDs and store-and-forward flags were relatively balanced, indicating a diverse dataset.

Further analysis will focus on relationships between features, exploring factors affecting trip durations and potentially building predictive models for NYC taxi travel times.

**Step2: Model Training**

We have successfully completed model training on the NYC taxi time prediction dataset. After thorough data pre-processing, including feature selection and data splitting.

The model was trained on a substantial portion of the dataset, allowing it to learn the underlying patterns and relationships.

**Step 3: Model Evaluation**

1. **Linear Regression**

With a low R2 score of 48% and a high Mean Squared Error (MSE), Linear Regression does not perform well for our NYC taxi time prediction model. To improve accuracy, we should explore more advanced regression techniques like Random Forest or Gradient Boosting Feature engineering, including incorporating additional relevant features, may enhance model performance. Addressing outliers and data transformations can also lead to better results. Ultimately, a different algorithm combined with feature enhancements is essential to achieve more accurate predictions for taxi travel times in NYC.

1. **Ridge Regression**

Both Ridge and Linear Regression models achieved a similar R-squared (48%) in the analysis. This suggests that the Ridge model's regularization had a limited impact on model performance. Careful feature selection and tuning of hyperparameters may be required for improved results. Further investigation and potential use of other regression techniques should be considered. Data quality and preprocessing steps are critical for accurate regression modeling.

1. **Lasso Regression**

With a low R2 score of 48% and a high Mean Squared Error (MSE), Lasso Regression also does not perform well for our NYC taxi time prediction model. To improve accuracy, we should explore more advanced regression techniques like Random Forest or Gradient BoostingFeature engineering, including incorporating additional relevant features, may enhance model performance. Addressing outliers and data transformations can also lead to better results. Ultimately, a different algorithm combined with feature enhancements is essential to achieve more accurate predictions for taxi travel times in NYC.

1. **Decision Tree**

In the analysis of the NYC dataset, a decision tree model achieved an impressive 99% accuracy. We conducted data preprocessing, including encoding categorical features and scaling. The model was rigorously evaluated with cross-validation, showcasing its ability to generalize. Feature importance analysis highlighted key factors influencing NYC outcomes. Future steps include further optimization and real-world applications, considering data quality and model interpretability.

1. **Random Forest**

Achieving a 99% accuracy with a Random Forest model indicates a highly accurate prediction. Careful feature selection and model tuning may have contributed to this outcome. Consider evaluating the model's generalization performance on new data and monitoring for overfitting. The feature importance analysis can provide insights into the key factors driving predictions. Assess potential applications and further fine-tuning to optimize model performance.

1. **XGBoost**

The XGBoost model displays strong predictive power with 99% accuracy on the training data. Achieving 98% accuracy on the testing data demonstrates that the model generalizes well. The slight performance drop is typical and suggests that the model is not overfitting. Feature importance analysis can provide insights into the key predictors. Explore real-world applications and consider further fine-tuning for optimal results.

1. **K-Nearest Kneighbors**

The K-Nearest Neighbors (KNN) model exhibited a training accuracy of 72%, indicating reasonable predictive performance. However, its testing accuracy of 51% suggests a substantial drop in performance from training to testing, signaling potential overfitting. Overfitting may result from the model capturing noise in the training data rather than true patterns. To improve generalization, it is crucial to revisit the model's hyperparameters, consider feature selection, and enhance data preprocessing. A comprehensive evaluation and potential adjustments are essential to boost model performance.

1. **AdaBoost**

The AdaBoost model demonstrates a strong ability to predict with a 98% accuracy rate, indicating a robust fit to the data. Feature importance analysis can uncover critical predictors used by the model. The high accuracy suggests it generalizes well, but cross-validation and evaluation on new data are recommended to ensure reliable performance. We can explore the potential applications and further fine-tuning to optimize results in real-world scenarios.

1. **Support Vector Machine**

The initial SVM model displayed inadequate performance with a negative R-squared score of -1.6, indicating a poor fit to the data. However, through rigorous hyperparameter tuning using GridSearchCV, substantial improvements were achieved. The revised model attained a training accuracy of 93% and a testing accuracy of 89%, demonstrating the effectiveness of hyperparameter optimization in enhancing predictive capabilities. To ensure the model's reliability and generalization, further evaluation, including cross-validation and testing on new data, is recommended. This successful transition underscores the importance of parameter fine-tuning in optimizing SVM models for accurate predictions in real-world scenarios.

1. **CONCLUSION**

**EDA:**

Vendor id distribution shows Vendor 2 receives more number of bookings

Store\_and\_fwd\_flag Count shows that majority of the time the taxi driver hasn't logged onto the vendor's systems.

Distribution of pickups and dropoffs on daily basis interprets that we can see that compared to other days, taxi booking rates are higher on the weekends (4- Friday and 5-Saturday). This suggests that individuals used to go out on weekends for their celebrations, parties, or even other personnel work.

Distribution of pickups and dropoffs on monthly basis shows that taxi reservations were more in the month of March and April.

Monthly trend for vendors tells us that both vendors' trips are at their maximum in the month of March and their lowest in the month of January, February, and after June.

Distribution of pickups and dropoffs on hourly basis gives us the insight that people often use taxi services to get to their workplaces in the mornings after 10:00. Additionally, the demand for taxis tends to surge in the late evening after six o'clock.

Passenger count distribution shows that most of the bookings are made by solo travellers, which means less number of people prefer car pool be less number of groups book car...people prefer to ride solo

**Model Evaluation:**

Ensemble methods, including Decision Tree, Random Forest, Xgboost, and Adaboost, outperform in taxi time prediction with near-perfect R2 scores around 0.99, showcasing their robust pattern capturing abilities. Linear regression models exhibit moderate performance, achieving an R2 of approximately 0.47. KNN holds its own with an R2 of 0.54, suggesting its effectiveness in this context. Conversely, SVM struggles with a negative R2, giving signal to suboptimal fit for the data. In conclusion, opt for ensemble methods like Random Forest or Adaboost for precise predictions, consider linear regression models for simplicity, and avoid SVM due to its performance drawbacks in this particular dataset.

1. Top of Form