**CAPSTONE PROJECT**

**Title: NYC Taxi Trip Data Analysis: A Machine Learning Perspective**

*Bharath Reddy Emani - 811295816 - bemani@kent.edu*

*Deekshith venepally - 811252891 - dvenepal@kent.edu*

*Indira Devi Galla - 811288674 -* [*igalla@kent.edu*](mailto:igalla@kent.edu)

*05/08/2024*

**3. An exact Description of the research question**

How can we leverage historical taxi trip data to forecast the demand for taxis at different locations and times of day using machine learning models? This could help optimize taxi fleet management and resource allocation.

**4. A discussion about why the solution to that problem would be valuable**

Leveraging historical taxi trip data for forecasting taxi demand using time series analysis and machine learning can be valuable for optimizing taxi fleet management and resource allocation. By accurately predicting demand at different locations and times, taxi companies can efficiently deploy their vehicles, reducing wait times for customers, optimizing driver schedules, and potentially reducing overall operational costs. This data-driven approach can lead to improved customer satisfaction and increased profitability for taxi businesses.

**5. A discussion about previous attempts to solve the problem and what you learned from them**

In the previous attempts to solve the problem, They experimented with different machine learning models to predict the number of passengers in a taxi ride based on various features. Initially, a linear model, likely a Linear Regression model, but encountered low accuracy. Linear models are effective when the relationship between the features and the target variable is linear. However, if the relationship is more complex or nonlinear, linear models may not capture it accurately.

So instead of doing that I have done three models and later on I have finalized one,

**Random Forest Regressor , ADA Boost Classifier , Gradient Boosting**

**6. A discussion about the methods your team used to solve the problem.**

1. **Data Exploration and Preprocessing**:
   * We began by loading the dataset and performing exploratory data analysis (EDA) to understand the dataset's structure, distributions, and relationships between variables. This included examining statistical summaries, visualizing distributions of variables like trip distance and fare amount, and identifying any anomalies or missing values.
   * Preprocessing steps such as handling missing values, filtering out erroneous data points (e.g., negative trip durations), and extracting relevant features from datetime variables were performed to ensure data quality and suitability for modeling.
2. **Feature Engineering**:
   * We engineered additional features such as pickup hour, pickup day, and pickup month from datetime variables to capture temporal patterns in taxi demand.
   * Additionally, we calculated aggregate statistics such as average trip counts per hour and day of the week to gain insights into temporal trends in taxi demand.
3. **Model Selection and Training**:
   * We experimented with multiple machine learning models, including RandomForestClassifier, AdaBoostClassifier, and GradientBoostingClassifier, to predict the number of passengers based on features such as pickup location, time of day, and day of the week.
   * Each model was trained on the training dataset and evaluated using performance metrics such as accuracy and mean absolute error (MAE) on the test dataset.
4. **Evaluation and Comparison**:
   * We evaluated the performance of each model using metrics such as accuracy and MAE and compared their performance to identify the most effective approach.
   * Visualizations such as bar plots and heatmaps were used to illustrate predicted passenger counts by pickup hour, pickup day, and pickup location, providing insights into temporal and spatial patterns of taxi demand.
5. **Discussion and Decision Making**:
   * Finally, we discussed the strengths and limitations of each modeling approach and made informed decisions based on the performance metrics and insights gained from the analysis.
   * The team considered factors such as model accuracy, computational efficiency, and interpretability to select the most suitable model for predicting taxi passenger counts.

**7. A discussion of the results your team obtained from your work.**

1. **Model Performance**:
   * Overall, our team achieved promising results with all three machine learning models: Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier.
   * The Random Forest and Gradient Boosting Classifiers performed particularly well, both achieving a test accuracy of 0.750. This indicates that they were able to correctly predict the number of passengers in 75% of the taxi rides in the test dataset.
   * The AdaBoost Classifier, while slightly lower in test accuracy at 0.709, still demonstrated respectable performance.
2. **Mean Absolute Error (MAE)**:
   * The MAE values for the Random Forest and Gradient Boosting Classifiers were quite close, both around 0.433. This suggests that, on average, the predicted number of passengers deviated from the actual number by approximately 0.433 passengers.
   * The MAE for the AdaBoost Classifier was not provided, but it seems to be slightly higher than that of the other two models.
3. **Model Comparison**:
   * When comparing the models, we observed similar performance in terms of test accuracy between the Random Forest and Gradient Boosting Classifiers. This indicates that both models were able to capture the underlying patterns in the data effectively.
   * The AdaBoost Classifier, while performing slightly lower in terms of test accuracy, still provided competitive results. However, it may require further tuning or adjustments to improve its performance.

**8. A discussion about how we can determine if you have successfully solved the problem.**

Model Performance Evaluation:

The success of the solution can be determined by evaluating the performance metrics of the machine learning models used. In the provided code, metrics such as accuracy, mean absolute error (MAE), and model scores (e.g., test accuracy) are computed to assess how well the models predict taxi demand.

Comparison of Different Models:

By comparing the performance of different models like Random Forest, AdaBoost, and Gradient Boosting, we can identify which approach yields the highest accuracy and reliability in forecasting taxi demand. This comparison helps in selecting the most effective solution for operational deployment.

**9. A discussion about how you might deploy the solution  in the real world to create value for someone.**

Optimizing Resource Allocation:

Utilizing forecasted demand insights helps taxi companies optimize resource allocation. This includes strategically deploying vehicles and drivers to areas with high anticipated demand, minimizing idle time, and improving overall fleet efficiency.

Decision Support for Managers:

Providing decision support tools based on predictive analytics empowers fleet managers to make informed decisions. They can adjust staffing levels and schedule shifts based on anticipated demand fluctuations to meet service demands efficiently.

Enhanced Customer Experience:

Implementing accurate demand forecasting contributes to a better customer experience. Reduced waiting times and improved availability of taxis based on predicted demand patterns enhance passenger satisfaction and loyalty.