**Introduction**

The objective of this project is to predict the duration of taxi trips in New York City using machine learning models. The dataset used is the nyc-taxi-trip-duration.zip, which contains taxi trip records including pickup and dropoff locations, timestamps, and trip duration. Two models, Linear Regression and Random Forest, were trained and evaluated based on their performance in predicting trip duration.

**Dataset**

The dataset nyc-taxi-trip-duration.zip contains three files:

1. train.zip
2. test.zip
3. sample\_submission.zip

For this project, we focused on the train.zip file, which includes the training data. The training data contains 1,458,644 records with the following features:

* id: Unique identifier for each trip
* vendor\_id: Identifier for the provider
* pickup\_datetime: Timestamp when the trip started
* dropoff\_datetime: Timestamp when the trip ended
* passenger\_count: Number of passengers in the trip
* pickup\_longitude: Longitude of the pickup location
* pickup\_latitude: Latitude of the pickup location
* dropoff\_longitude: Longitude of the dropoff location
* dropoff\_latitude: Latitude of the dropoff location
* store\_and\_fwd\_flag: Flag indicating if the trip data was stored and forwarded due to connectivity issues
* trip\_duration: Duration of the trip in seconds

**Methodology**

1. **Data Extraction and Loading:** The dataset was extracted from the nested ZIP files and loaded into a Pandas DataFrame for analysis and preprocessing. The initial shape of the data was (1458644, 11).

Train data shape: (1458644, 11)

pickup\_date pickup\_hour pickup\_day\_of\_week

0 2016-03-14 17 0

1 2016-06-12 0 6

2 2016-01-19 11 1

3 2016-04-06 19 2

4 2016-03-26 13 5

Quantity of trips daily: 8015.0

1. **Data Preprocessing:**
   * **Datetime Features:** The pickup\_datetime and dropoff\_datetime columns were converted to datetime objects. Additional features such as pickup\_date, pickup\_hour, and pickup\_day\_of\_week were extracted from the pickup\_datetime.
   * **Geographical Features:** The Haversine formula was used to calculate the distance between pickup and dropoff locations. The angle of direction was also calculated based on the geographical coordinates.

haversine\_distance direction

0 1.498521 99.970196

1 1.805507 -117.153768

2 6.385098 -159.680165

3 1.485498 -172.737700

4 1.188588 179.473585

* + **Encoding Categorical Variables:** The store\_and\_fwd\_flag was encoded as a binary feature, mapping 'N' to 0 and 'Y' to 1.
  + **Log Transformation:** The target variable trip\_duration was log-transformed to trip\_duration\_log to normalize its distribution.

Shape of data after dropping columns: (1458644, 13)

1. **Model Training:** The dataset was split into training and validation sets. The features were normalized using Min-Max scaling. Two models were trained:
   * **Linear Regression:** A simple linear model that predicts the log-transformed trip duration.
   * **Random Forest:** An ensemble model that uses multiple decision trees to improve prediction accuracy.
2. **Evaluation Metrics:** The models were evaluated using the following metrics:
   * **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.
   * **Root Mean Squared Error (RMSE):** Square root of MSE, providing an error metric in the same units as the target variable.
   * **Mean Absolute Error (MAE):** Measures the average absolute difference between actual and predicted values.
   * **R² (R-squared):** Indicates the proportion of variance in the target variable explained by the model.
   * **Custom Accuracy:** Defined as the proportion of predictions within 10% of the actual value.

**Results**

1. **Linear Regression:**
   * Train RMSLE: 0.65
   * Valid RMSLE: 0.64
   * MSE: 9720032.95
   * RMSE: 3117.70
   * MAE: 317.26
   * R²: 0.03
   * Custom Accuracy: 0.26

Linear Regression Train RMSLE: 0.65

Linear Regression Valid RMSLE: 0.64

Linear Regression MSE: 9720032.95, RMSE: 3117.70, MAE: 317.26, R2: 0.03

Linear Regression Accuracy: 0.26

1. **Random Forest:**
   * Train RMSLE: 0.41
   * Valid RMSLE: 0.43
   * MSE: 9720032.95
   * RMSE: 3117.70
   * MAE: 317.26
   * R²: 0.03
   * Custom Accuracy: 0.26

Random Forest Train RMSLE: 0.41

Random Forest Valid RMSLE: 0.43

Random Forest MSE: 9720032.95, RMSE: 3117.70, MAE: 317.26, R2: 0.03

Random Forest Accuracy: 0.26

Despite the Random Forest model typically outperforming Linear Regression in many scenarios, both models showed similar performance metrics. The RMSLE values indicate that while the Random Forest model had a lower training error, both models had similar validation errors. The custom accuracy metric shows that only about 26% of the predictions were within 10% of the actual values, indicating room for improvement.

**Conclusion**

The project successfully demonstrated the use of machine learning models to predict the duration of taxi trips in New York City. Both Linear Regression and Random Forest models showed comparable performance in this task, with similar RMSLE and custom accuracy metrics. Future work could explore additional feature engineering, tuning of model hyperparameters, and the use of other advanced models to further improve prediction accuracy.