REPORT: Predicting Taxi Demand in the Bronx Region

Introduction

- **Objective**: Predict the hourly demand for taxis in the Bronx for the first week of September 2024 and determine the optimal number of taxis required to meet this demand.
- **Datasets**:
    - **Taxi Trip Data**: Included details like trip distance, pickup/drop-off times, passenger count, and location IDs.
    - **Weather Data**: Included hourly weather observations such as temperature, precipitation, rain, and snowfall.
    - **Zone Lookup Data**: Provided mapping between location IDs and boroughs.

Data Preprocessing, Feature Engineering, and Data Visualization

- **Cleaning the Data**:
    - Removed outliers in trip distance to eliminate erroneous data.
    - Filled missing values with the median value or dropped the rows according to the number of missing values in the data.
    - Converted datetime columns to appropriate datetime types for time-based feature extraction.

- **Merging Datasets**:

    1. **Aligning Data by Hour**: We aligned weather data with taxi demand data using hourly timestamps, ensuring that weather conditions match the demand observed during each hour. Adjusted time zones for consistency.
    2. **Calculating Hourly Taxi Demand**: Aggregated taxi trip data to get the total number of trips for each hour. This serves as the target variable for predicting demand.
    3. **Merging Demand with Weather Data**: Combined hourly taxi demand data with corresponding hourly weather conditions. This allows us to analyze how weather factors like temperature, rain, and snowfall influence demand.
    4. **Zone-Specific Demand Calculation**: Calculated taxi demand for each zone in the Bronx to capture spatial variations in demand. Mapped each zone using its location ID to add geographical context.
    5. **Filtering for Bronx Area**: Focused the analysis specifically on the Bronx by filtering out data from other boroughs. This step ensures that we're targeting the area of interest.
    6. **Combining Bronx Demand with Weather**: Merged the Bronx-specific demand data with corresponding weather conditions. This created a comprehensive dataset with hourly demand data, weather conditions, and zone information.
    7. **Preparing the Final Dataset**: Selected relevant columns for modeling, including demand, temperature, precipitation, and zone names.

- **Feature Engineering**:

1. **Time-Based Features**: Extracted hour, day_of_week, and month to capture daily and seasonal variations in demand.
2. **Rush Hour and Weekend Indicators**: Added is_rush_hour_morning, is_rush_hour_evening, and is_weekend to identify peak times and weekend effects on demand.
3. **Interaction Terms**: Created precipitation_rush_hour to measure the impact of bad weather during rush hours.
4. **Lagged Demand Features**: Added lag_1h and lag_24h to capture recent demand patterns.
5. **Moving Averages**: Computed moving_avg_3h and moving_avg_24h to smooth short-term fluctuations.
6. **Weather Lagged Features**: Included temperature_1h_ago and temperature_24h_ago to track recent temperature changes for better weather predictions.

- **Encoding for Categorical Variables**:
    - **One-Hot Encoding**: Applied one-hot encoding for the Zone feature when training the Linear Regression model. This approach was used because Linear Regression can be sensitive to the ordinal implications of label encoding, which could introduce unintended biases.
    - **Label Encoding**: Used label encoding for the Zone feature in other models like XGBoost, Random Forest, and the Ensemble model. Label encoding was chosen for these models as they can handle numerical representations effectively without assuming ordinality, allowing for a more compact representation of categorical variables.

- **Data Visualization**:

1. **Correlation Heatmap:** To understand the relationship between different numerical variables like taxi_demand_per_zone, temperature_2m, precipitation, rain, snowfall, hour, and day_of_week. This helps identify which weather conditions or time-based features may influence taxi demand the most.
2. **Line Plot of Average Taxi Demand by Hour:** To observe how taxi demand varies throughout the day. This line plot shows the average taxi demand at each hour of the day, highlighting peak hours. Peaks might correspond to rush hours, while dips could represent low-demand periods like late nights.
3. **Line Plot of Average Taxi Demand by Day of the Week:** To analyze the variation in taxi demand across different days of the week. By plotting the average demand against the days of the week (0 = Monday), this visualization reveals which days have higher or lower demand, such as weekends versus weekdays. This helps in understanding weekly patterns and trends.
4. **Bar Plot of Average Taxi Demand by Zone:** To compare the average taxi demand across different zones within the Bronx, providing a clear view of which zones have the highest and lowest demand. This helps in identifying areas with consistent demand.

5. **Histogram of Temperature:** To understand the distribution of temperature values. It helps identify common temperature ranges, and whether the distribution is skewed towards warmer or cooler temperatures.
6. **Scatter Plot of Temperature vs. Taxi Demand:** To examine the relationship between temperature and taxi demand. This scatter plot helps identify patterns or trends, such as whether higher or lower temperatures correlate with increased or decreased demand.
7. **Scatter Plot of Precipitation vs. Taxi Demand:** To explore the impact of precipitation on taxi demand. It helps assess if rainfall or snow significantly affects the demand, indicating if more taxis are needed during bad weather conditions.

3. Baseline Model

- **Model Choice**: Linear Regression
    - **Reasoning**: It serves as a simple, interpretable model that can capture basic linear relationships between weather conditions, time-based features, and taxi demand. It directly models the relationship between the input features (such as time of day, day of the week, and weather conditions) and the output variable (taxi demand). The simplicity of Linear Regression makes it an ideal starting point, allowing us to establish a benchmark for more complex models. Additionally, using Linear Regression provides insights into the influence of each predictor, offering transparency in understanding how features like temperature or precipitation impact taxi demand.
    - **Key Considerations**: Linear Regression assumes a linear relationship between the features and the target, which might be a limiting factor for capturing complex patterns. However, it is a great starting point to quickly assess the predictive power of our basic features before adding complexity.
- **Results**:
    - MSE: 0.3796118
    - MAE: 0.4060604
    - **Interpretation**: The linear regression model provided a quick estimate of demand but struggled with capturing non-linear relationships and interactions, leading to relatively high error values.

4. Improved Model

- **Model Choice**: XGBoost Regressor + Feature Engineering + Data Scaling + Hyperparameter Tuning
    - **Reasoning**: XGBoost is a robust gradient-boosting model capable of capturing complex interactions between features. It is known for handling large datasets and capturing non-linear relationships between features and the target variable. Given the complexity of taxi demand, which depends on both temporal patterns and weather conditions, XGBoost can capture intricate interactions and variations more effectively than a linear model. The model's ability to perform feature importance analysis also helps in identifying which weather variables or time-based features are most predictive of demand.

    - **Key Considerations**: XGBoost's flexibility comes with a risk of overfitting, especially with smaller datasets. Therefore, careful hyperparameter tuning (e.g., number of trees, learning rate, and max depth) is essential to balance model complexity and performance.
    - **Feature Enhancements**:
        - Added lagged demand features and moving averages to account for temporal dependencies.
        - Improved feature set with binary indicators for rush hours and weekends.
- **Hyperparameter Tuning:** For hyperparameter tuning, we used GridSearchCV to optimize the XGBoost model. The process involved searching over a range of key parameters like n_estimators, learning_rate, max_depth, and min_child_weight to find the best combination that minimized the Mean Absolute Error (MAE). We employed 3-fold cross-validation to ensure robust evaluation across different subsets of the training data. The best parameters were then used to train a final model, which improved prediction accuracy by balancing model complexity and preventing overfitting, leading to more accurate demand forecasts.
- **Results**:
    - MSE: 0.362298
    - MAE: 0.38698
    - **Comparison with Baseline**: The XGBoost Regressor outperformed the baseline Linear Regression model in predicting taxi demand. While Linear Regression provided a straightforward approach, it struggled with capturing complex, non-linear relationships in the data, such as the effects of temperature, rush hours, and lagged demand. In contrast, XGBoost, with its gradient boosting framework, effectively modeled these interactions, leading to lower Mean Squared Error (MSE) and Mean Absolute Error (MAE). The improvement in performance demonstrates XGBoost's ability to better capture the temporal patterns and interactions between weather conditions and demand, making it a more suitable choice for this task.

5. Best Model

- **Model Choice**: Ensemble Model (Voting Regressor) + Advanced Feature Engineering + Data Scaling
    - **Components**: XGBoost, Random Forest Regressor, LightGBM
    - **Reasoning**: Combining the predictions of multiple models through an ensemble approach helps to reduce the variance and improve the overall accuracy of predictions. The ensemble model, specifically the Voting Regressor, combines predictions from XGBoost, Random Forest, and LightGBM. XGBoost is excellent for capturing complex interactions between features. Random Forest provides robustness and reduces overfitting by averaging multiple decision trees. LightGBM offers fast training times and efficient handling of large datasets.
- **Feature Set**: Included engineered features like lag_1h, lag_24h, and moving averages.
- **Hyperparameter Tuning:** Tried to employ hyperparameter tuning, but it did not improve the results.
- **Key Considerations:** While the ensemble model typically improves predictive performance, it comes at the cost of increased computational complexity. Training and predicting with multiple

models require more resources, making it less ideal for real-time applications but suitable for forecasting tasks like this one.

- **Results**:
    - MSE: 0.168335
    - MAE: 0.210238
    - **Improvement**: The ensemble model outperformed individual models by leveraging the strengths of each, further reducing the prediction error. The Voting Regressor leverages the strengths of these diverse models, resulting in more accurate and generalized predictions than any individual model alone. By taking a weighted average of predictions, the ensemble model can better adapt to varying conditions such as different times of day or extreme weather events.

6. Forecasting the First Week of September 2024

- **Approach**: For predicting weather conditions such as temperature, precipitation, rainfall, and snowfall, a separate model was developed using XGBoost. The model used time-based features like hour of the day, day of the week, and month alongside lagged temperature features (temperature_1h_ago and temperature_24h_ago) to capture temporal trends. The model was trained on historical weather data, and the resulting predictions for the first week of September 2024 served as inputs for the final demand forecasting model.

7. Challenges

- **Data Quality and Cleaning:** I encountered inconsistencies between the different datasets, such as time zone mismatches. This required careful alignment of timestamps to ensure accurate merging of weather and trip data.
- **Creating Lagged Features**: Incorporating lagged values (e.g., taxi demand 1 hour ago or 24 hours ago) added complexity, as I needed to backfill missing values and ensure that the rolling averages were calculated correctly.
- **Balancing Simplicity and Complexity**: The baseline model using Linear Regression was easy to implement but struggled with non-linear relationships in the data. On the other hand, advanced models like XGBoost and ensemble methods required more tuning to achieve optimal performance.
- **Accuracy of Weather Models**: Predicting weather variables like temperature and precipitation required a high degree of accuracy, as errors in weather predictions could compound when used to predict taxi demand.
- **Generalizing Predictions to Future Dates**: Using the weather predictions for September 2024 as inputs to the demand model posed a challenge because there was no historical data for this specific period. The model had to be robust enough to handle this extrapolation.
- **Potential for Overfitting**: Given the high number of features and complex interactions, there was a risk of overfitting the demand models to the training data. This required careful validation and testing to ensure that the models generalized well.
- **Model Training Time**: Advanced models like XGBoost and ensemble methods took longer to train, especially when combined with hyperparameter tuning. This required balancing between model complexity and computational feasibility.

9. Conclusion and Future Work

- **Summary**: This project demonstrated the importance of feature engineering and model optimization in accurately predicting taxi demand. By carefully selecting time-based and weather-related features, we were able to capture the underlying patterns that influence taxi demand in the Bronx. The addition of lagged features and interaction terms further improved the model's ability to understand temporal trends and sudden demand spikes, such as during rush hours or adverse weather conditions.
- **Model Performance**: The transition from a simple Linear Regression baseline model to a more sophisticated ensemble approach, including XGBoost, Random Forest, and LightGBM, resulted in a significant reduction in error metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE). This improvement underscores the value of using non-linear models that better capture the complexities in taxi demand data, such as non-linear relationships between weather conditions and demand.
- **Future Work**: Future enhancements could include integrating data from external sources like local events, holidays, or major public transport disruptions, which could have a significant impact on taxi demand. This would help the model adjust predictions based on factors beyond weather and time-based trends, improving its applicability for operational planning.

- SHARANYA AKKENAPALLY