**Restaurant Visitor Forecasting:** This project focuses on forecasting the number of visitors to restaurants in Japan. The goal is to build a model that can accurately predict the daily number of visitors for various restaurants, helping them with resource planning and management.

**Table of Contents**

1. Problem Statement

2. Data

3. Methodology

Data Preprocessing

Exploratory Data Analysis (EDA)

Feature Engineering

Modelling

4. Results

5. Conclusion

**Problem Statement:** The objective of this project is to predict the daily number of visitors for a given restaurant. This is a time-series forecasting problem that also involves relational data from two different systems (Air and HPG).

**Data:** The dataset is comprised of multiple CSV files, each containing specific information about the restaurants, reservations, and visitor data. The main files used are:

`air\_store\_info.csv`: Information about restaurants in the Air system.

`hpg\_store\_info.csv`: Information about restaurants in the HPG system.

`air\_visit\_data.csv`: Historical visitor data for Air restaurants.

`air\_reserve.csv`: Reservations made in the Air system.

`hpg\_reserve.csv`: Reservations made in the HPG system.

`date\_info.csv`: Information about calendar dates, including holidays.

`store\_id\_relation.csv`: A mapping between Air and HPG store IDs.

`sample\_submission.csv`: The format for the submission file.

**Methodology**

**Data Preprocessing**

**1.Data Loading:** All the provided CSV files were loaded into pandas DataFrames.

**2.Data Merging:** The different data sources were merged to create a comprehensive dataset.

`air\_store\_info` and `hpg\_store\_info` were merged using the `store\_id\_relation` file.

`air\_reserve` and `hpg\_reserve` were combined to create a unified reservations dataset.

The visitor data (`air\_visit\_data`) was combined with the submission file to create a single `visits\_df` for training and prediction.

**3.** **Handling Missing Values:** After merging the datasets, some missing values were introduced. For instance, not all Air restaurants had a corresponding HPG ID. These missing values were handled by filling them with 0, as they represent the absence of a corresponding entry in the other system. This approach was chosen to maintain the integrity of the data while allowing the models to process it.

**Exploratory Data Analysis (EDA):** EDA was performed to understand the data and identify patterns. Key findings from the EDA include:

**Visitor Trends:** A clear weekly seasonality was observed, with more visitors on Fridays and weekends. A long-term step structure was also noticed, likely due to new restaurants being added to the dataset.

**Visitor Distribution:** The number of visitors was found to be right-skewed. A log transformation (`log1p`) was applied to the target variable to make its distribution closer to normal.

**Holiday Impact:** Holidays and weekends generally saw a higher number of visitors.

**Genre Analysis:** Different restaurant genres showed varying visitor patterns. "Izakaya" was the most popular genre.

**Location Analysis:** The distribution of restaurants across different areas of Japan was visualized, with "Fukuoka-ken Fukuoka-shi Daimyo" having the most restaurants.

**Feature Engineering:** Several new features were created to improve model performance:

**Date-based Features**: `year`, `month`, `day\_of\_year`, `week\_of\_year`, `day\_of\_week`, `is\_month\_end`.

**Holiday Features:**`weekend`,`day\_off\_flg`, `tomorrow\_is\_holiday`, `yesterday\_is\_holiday`.

**Location-based Features:** The `air\_area\_name` was split to extract `Todofuken`, `city`, and `street`. Features like the number of stores in the same street, city, and prefecture were also created.

**Reservation-based Features:** Aggregated reservation data to get features like reserve\_visitors\_count ,mean\_visit\_hour, and hours\_ahead.

**Interaction Features:** Combined categorical features to create new interaction features like area\_genre and store\_weekday.

**Target-derived Features:** Calculated statistics like mean, median, min, and max visitors for each store-weekday-holiday combination.

**Modelling:** Several models were trained and evaluated:

**1.ARIMA/SARIMA:** Time-series models were built at both the genre and restaurant level. While they captured seasonality, their performance was not optimal due to the complexity of the data.

**2.Prophet:** Facebook's Prophet library was used for time-series forecasting, which showed better performance than traditional ARIMA models, especially when external regressors were included.

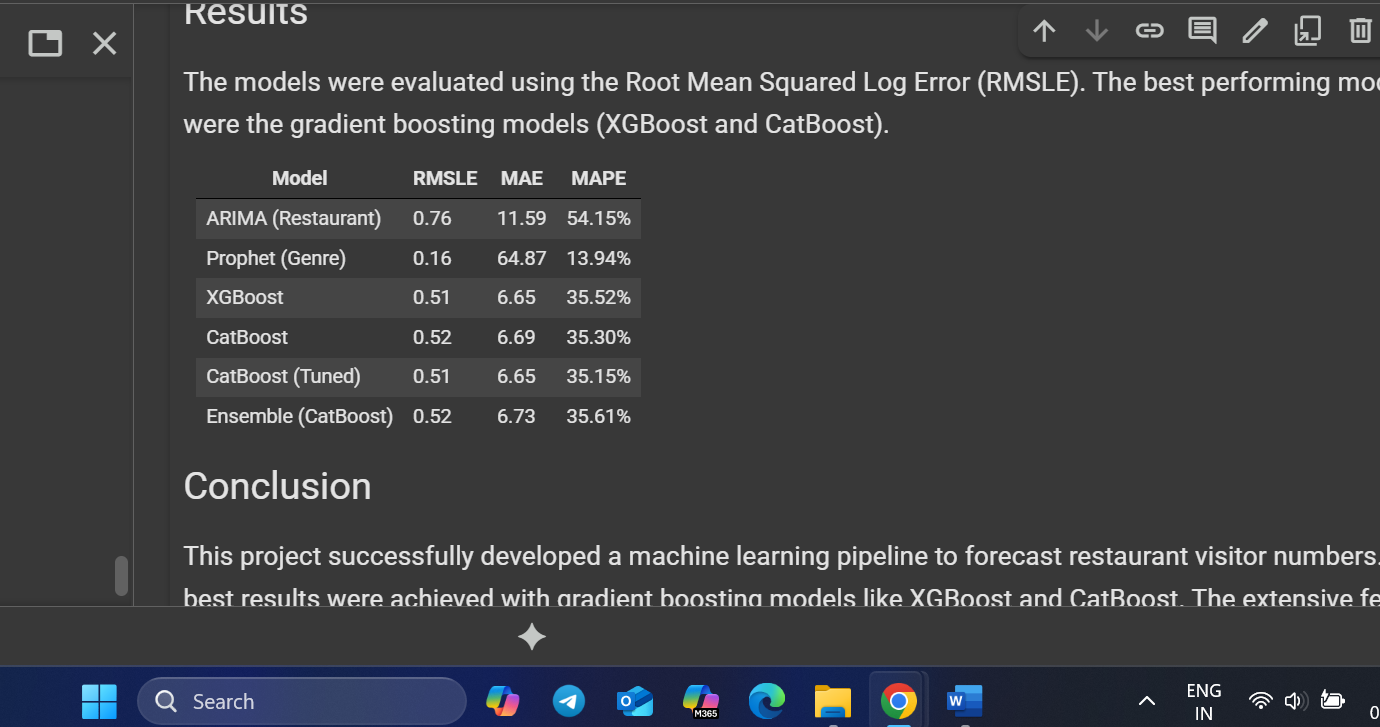
**3.XGBoost:** A gradient boosting model that performed well. Feature importance plots showed that mean\_visitors was the most important feature.

**4.CatBoost:** Another gradient boosting model that is particularly effective with categorical features. It yielded results similar to XGBoost.

**5.Hyperparameter Tuning:** GridSearchCV was used to tune the hyperparameters of the CatBoost model, which resulted in a slight improvement in performance.

**6. Ensemble Models:** A cohorted ensemble approach was tested, where separate models were trained for different categories of features (e.g., weekday, holiday). This did not significantly improve the results.

**Results:** The models were evaluated using the Root Mean Squared Log Error (RMSLE). The best performing models were the gradient boosting models (XGBoost and CatBoost).



**Conclusion:** This project successfully developed a machine learning pipeline to forecast restaurant visitor numbers. The best results were achieved with gradient boosting models like XGBoost and CatBoost. The extensive feature engineering played a crucial role in the model's performance. The final model can be used to provide valuable insights to restaurant owners for better planning and management.