**Data Science Project Protocol for**

**Restaurant Visitor Prediction**

*Author:*

Yael Coletti

March 2023

**Introduction**

Data Science is an interdisciplinary field that combines statistical analysis, machine learning, and domain expertise to extract insights and knowledge from data. It is a rapidly growing field that has significant implications for businesses, organizations, and individuals alike. The primary objective of data science is to discover patterns, relationships, and insights in data that can be used to make informed decisions.

In this project, we will be focusing on predicting the number of customers for a selected group of restaurants in the Air Regi system. This is a common problem in the restaurant industry, where accurate predictions of customer demand are critical for efficient restaurant management and financial performance.

To achieve this goal, we will be using various statistical and machine learning techniques to analyze the data and build predictive models. We will be answering questions such as:

* What factors influence the number of visitors to a restaurant?
* How can we predict the number of customers accurately?
* What are the most significant predictors of customer demand?

We will be exploring new and innovative ways of analyzing the data, such as feature engineering and data visualization techniques.

The ultimate goal of this project is to provide restaurant owners with accurate predictions of customer demand that can be used to optimize their operations and improve their financial performance. By leveraging the power of data science, we can help businesses make informed decisions that can lead to increased profitability and long-term success.

**Methodology (Project design)**

The data will be obtained from Kaggle challenge “Recruit Restaurant Visitor Forecasting”.

You can find the full information in the link below:

<https://www.kaggle.com/competitions/recruit-restaurant-visitor-forecasting>

This dataset contains information from two restaurant services:

* AirREGI (air) - a reservation control and cash register system.
* Hot Pepper Gourmet (hpg) - here users can search restaurants and also make a reservation online.

The same restaurant can appear in both of the systems under unique ID under air\_ or hpg\_. The relation between the two ID’s is described in the `store\_id\_relation.csv`. This file allows to join selected restaurants that have both the air and hpg system.

Latitudes and Longitudes are not exact to discourage de-identification of restaurants.

### **Info from Kaggle:**

### **air\_reserve.csv**

This file contains reservations made in the air system. Note that the reserve\_datetime indicates the time when the reservation was created, whereas the visit\_datetime is the time in the future where the visit will occur.

* air\_store\_id - the restaurant's id in the air system
* visit\_datetime - the time of the reservation
* reserve\_datetime - the time the reservation was made
* reserve\_visitors - the number of visitors for that reservation

### **hpg\_reserve.csv**

This file contains reservations made in the hpg system.

* hpg\_store\_id - the restaurant's id in the hpg system
* visit\_datetime - the time of the reservation
* reserve\_datetime - the time the reservation was made
* reserve\_visitors - the number of visitors for that reservation

### **air\_store\_info.csv**

This file contains information about select air restaurants. Column names and contents are self-explanatory.

* air\_store\_id
* air\_genre\_name
* air\_area\_name
* latitude
* longitude

### **hpg\_store\_info.csv**

This file contains information about select hpg restaurants. Column names and contents are self-explanatory.

* hpg\_store\_id
* hpg\_genre\_name
* hpg\_area\_name
* latitude
* longitude

### **store\_id\_relation.csv**

This file allows you to join select restaurants that have both the air and hpg system.

* hpg\_store\_id
* air\_store\_id

### **air\_visit\_data.csv**

This file contains historical visit data for the air restaurants.

* air\_store\_id
* visit\_date - the date
* visitors - the number of visitors to the restaurant on the date

### **sample\_submission.csv**

This file shows a submission in the correct format, including the days for which you must forecast.

* id - the id is formed by concatenating the air\_store\_id and visit\_date with an underscore
* visitors- the number of visitors forecasted for the store and date combination

### **date\_info.csv**

This file gives basic information about the calendar dates in the dataset.

* calendar\_date
* day\_of\_week
* holiday\_flg - is the day a holiday in Japan

Project Design:

1. The first part - DataBase and flat file generation will be done SQLite. Since this program doesn’t support many functions to format dates, I will continue this part to R.
2. The second part – EDA and Feature Engineering will be executed in R language.
3. The third part – Feature Selection and Modelling will be done in Python.

A comprehensive workflow description is as follows (initially, the SQLite-generated flat file was not utilized):

The first step is to import all the CSV files and combine them into a comprehensive dataset that includes the dates, restaurant IDs, genre and area names, and the number of reserved visitors.

The following step involves carrying out an Exploratory Data Analysis (EDA) using R packages. This step will entail analyzing the original variables, as well as the engineered variables, through univariate, bivariate, and multivariate analyses, alongside visualization techniques that employ graphs and statistical tests.

Thereafter, an assessment for missing values and outliers will be conducted. Next, we will utilize the Jupyter Notebook in Python to undertake Data Preprocessing, Feature Selection, and Data Partitioning for test, train, and validation purposes. Eventually, we will test and assess a few regression models to determine their efficacy.

**Deployment the Model: Theoretically**

The successful deployment of a model relies on a comprehensive Quality Assurance (QA) process to ensure that the model operates at optimal levels, produces accurate results, and meets the expectations of end-users. The QA process should cover all units that make up the project, including data preparation, data modeling, and data visualization.

To conduct QA for the project, a team of experienced professionals who understand the domain of the project will be assembled. The team will assess each unit of the project, including the data preparation, model training, and model validation stages.

For each step of the project, a QA protocol will be developed to ensure that the project meets the required standards. The QA protocol will cover areas such as data quality, data accuracy, and model accuracy. This protocol will enable the team to identify any issues that may arise during the model development process and ensure that they are addressed promptly.

The final user of the predictions will be individuals or organizations who require information on restaurant reservations, such as restaurant owners or managers. The predictions will be presented through an intuitive and user-friendly interface that will allow the end-user to access the information they need quickly and easily.

End-users will receive training on how to interpret and use the predictions. The training will cover the basics of the model and the information that it provides. It will also demonstrate how to access and navigate the interface, how to interpret the results, and how to use the information to make informed decisions.

The predictions will be deployed on a web-based platform that is accessible to end-users. The platform will provide a user-friendly interface that allows end-users to access the predictions quickly and easily.

The model will be updated on a regular basis to ensure that it remains accurate and up-to-date. The frequency of updates will depend on the nature of the data being used and the rate at which the data changes.

In cases where the model returns a null prediction due to incomplete data, the end-user will be alerted, and the system will prompt the user to provide additional data. The end-user will also be given suggestions on how to improve the data quality to improve the accuracy of the predictions.

Several models were used during the development of the project, including regression models and decision tree models. The regression model was selected for the final prediction based on its accuracy and efficiency.

Various metrics were used to evaluate the accuracy of the predictions, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The regression model produced a final prediction accuracy of 90%, indicating that it was highly effective in predicting restaurant reservations.

**Results**

1. The final amount of data used:

The total number of rows of the full data: 108,439 and 60 columns (variables)

The test set the last 10% of the full data: 10,843 rows

For the validation set the last 20% of the remaining data: (108,439 - 10,843)\*0.2 = 19520 rows

The Train size : 78076 rows

2. Outliers were not found in the data.

3. The full data did not include missing data.

4. Methods used to transform the data and generate new features

* bin-counting (categorical variables)
* one-hot encoding/dummy variables for small categories.
* geospatial data geocoding
* extraction of prefixes (`prefecture`)
* date and time info extraction (day, month, week, year, hours etc.)

**Conclusions**

In conclusion, this project was challenging and required a considerable amount of time and effort. The most challenging parts were gathering all the data required for the model, as well as dealing with missing data. Even though I attempted to include weather data from another database, it did not work due to the high amount of missing data. Despite these challenges, the final model performed well and provided reasonable predictions.

It is important to note that the model's limitations are due to the data's constraints. Since the model was only trained on a specific set of restaurants in the Air Regi system, it may not generalize well to other restaurant types or locations. Additionally, the model's accuracy may be affected by external factors such as unexpected weather patterns or local events. Therefore, caution should be exercised when using this model to make predictions outside of the data's scope.

Overall, this project demonstrates the importance of carefully selecting and preparing data for machine learning models. Additionally, it highlights the potential benefits of accurately predicting the number of customers for restaurant owners to ensure efficient restaurant management, enhance financial performance, and promote global sustainability.