**Data Science Project Protocol**

*Author:*

***Eliran Tzir***

# Introduction.

# **Recruit Restaurant Visitor Forecasting**

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

**Background about the restaurant industry:**

Running a thriving local restaurant isn't always as charming as first impressions appear. There are often all sorts of unexpected troubles popping up that could hurt business.

One common predicament is that restaurants need to know how many customers to expect each day to effectively purchase ingredients and schedule staff members. This forecast isn't easy to make because many unpredictable factors affect restaurant attendance, like weather and local competition. It's even harder for newer restaurants with little historical data.

## 1. <https://restaurant.org/research/restaurant-statistics/restaurant-industry-facts-at-a-glance> 2.<https://www.forbes.com/sites/aliciakelso/2021/12/30/what-does-the-restaurant-industry-look-like-in-2021-and-beyond/?sh=478b6c6926a4>

## **Restaurant Industry Facts at a Glance**

* $659 billion: Restaurant industry sales in 2020, down $240 billion from expected levels
* 12.5 million: Restaurant industry employees at the end of 2020, down 3.1 million from expected levels
* 110,000: Restaurant locations that are temporarily or permanently closed
* 9 in 10 restaurants have fewer than 50 employees
* 7 in 10 restaurants are single-unit operations
* 8 in 10 restaurant owners started their industry careers in entry-level positions
* 9 in 10 restaurant managers started in entry-level positions
* Restaurants employ more minority managers than any other industry

The question that I would love to get answers on are:  
How many customers will visit us next week?

The outcomes will help us to predict the number of restaurant visitors so we can prepare the logistic better and as a result, try to maximize the revenue.

The outcomes might increase the revenue of the restaurant.

The outcomes will help us to predict the number of restaurant visitors so we can prepare the logistic better and as a result, try to maximize the revenue.

As we know the day of the week, time of the day, holidays, and location are influencing the outcomes.

Today we have technological tools through which we can define the factors that affect the number of visitors. By using them, we can predict what is going to happen in the future and prepare ourselves to earn the highest potential profit.

# 

# Methodology (Project design)

## Data

* I used **Recruit Restaurant Visitor Forecasting data.**
  + The data shows the number of visitors by days, times, dates, holidays etc.  
    In addition, we can find the genre of the restaurant and the location.
  + possible external data sources that may enrich our data are the number of the population in the locations, restaurant trends (for genres) and national holiday ( Italian, French, Japanese etc … )
  + Data for external validation that I used were academic researches, professional magazines (Forbes, Bloomberg, INC., THE ECONOMIST, CONSUMER REPORTS, etc .. )
* The time frames of the project are:
  + Time-frame for training = 80% of 2016-01-01 until 2017-04-22
  + Time-frame for the test = 20% of all the data. The test set is split based on time and covers a chosen subset of the air restaurants. Note that the test set intentionally spans a holiday week in Japan called the "Golden Week."
  + There are days in the test set where the restaurant were closed and had no visitors. These are ignored in scoring. The training set omits days where the restaurants were closed.
* How do you define your subjects? Every criterion that can affect restaurant visitors
* The outcome of my project will be sum\_visitors of the following week.
* Are there confounding variables that may affect the outcome ? Yes, we will drop them using regression models in the feature selection stage.
* Describe your data exploration strategy.
  + Data Preparation
  + Flat-File generation
  + Outlier detection and treatment
  + Missing values and imputation
  + Variable selection
  + Train - DEv - Test preparation
  + Model Selection

Model Fine Tuning

* Which techniques will be applied to enrich the data ? Create new columns from the data we have that can add new ways for us to explore the data
* How will you deal with outliers? - Will use IQR analysis method and remove them if needed.
* How you will deal with missing values - Replace the missing column values with new variables (categorical)
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Test 20 of All the data, Train 80% of (80\* from all the data), Dev 20% of (80\* from all the data)
* Do you need to stratify/subsample your data? How?
  + I used the sample the following cod to sort and split the data.
  + X\_train, X\_test, y\_train, y\_test = train\_test\_split(df, y, test\_size=0.2, random\_state=42)
* What techniques will you apply to model your outcome?
  + Unsupervised
  + **Regression**
  + Classification
* Will you use **cross-validation** and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
  + Max\_visitors | avg\_visitors | count\_visitors | weekend\_visitors

I used Lasso, XGboost, SVM and Random Forest regression models for choosing the best features

* Do you plan to use ensemble or will use your best model? Use the best (elastic-net)

## Deployment of your model

* Who will make the QA of the project? Restaurant waiters
  + Which units will be assessed - Easy to use and the gap between the prediction and the reality
  + Write a QA protocol for each step of the project
  + 1. Data Preparation
    - Validate that all the tables were loaded as expected to SQL with the correct data types.
  + 2. Flat File Generation
    - Run several queries to test the results (No duplicates)
  + 3. Exploratory Data Analysis (EDA)
    - Check that the file was uploaded as needed and run the correlation metrics
  + 4. Missing Values imputation
    - Run a missing values report and compare the results
  + 5. Outlier detection and treatment
    - Run outliers report and compare the results
  + 6. Variable selection
    - Run the Regressions and compare the results
  + 7. Train-Dev-Test Prepration
    - Calculate the amount of rows that should be in each of the fles
  + 8. Model Selection
    - Run the regressions to check if the results are the same
  + 9. Model Fine-Tuning
    - Run the model before and after the fine tuning to check if the results are better and run the results on the test data.
* Who is the final user of the predictions? - managers of restaurants
* How the prediction will be presented to the final user? As a weekly report to their emails and obviously they can get this information at the website as well.
* How will the final user be trained to use and interpret the prediction? A support team that will give to the end user a technical support
* On which platform the predictions will be deployed? Web and App
* How frequently will the model be updated? Once a week
* What will happen in cases where the model return a null prediction (eg. incomplete data)? Should use the most closest prediction
* Which models were used and which were selected for the final prediction.
  + The mthat were tested are : Decision Tree, GBM, LinearRegression, BayesianRidge, Elastic-Net, RandomForest, ADABoost, SVR, SGD and Ridge.
  + The model that were chosen using MAE factor was Elastic-Net
* Which measurements were used to evaluate the prediction?
  + Max\_visitors | avg\_visitors | count\_visitors | weekend\_visitors
* Which results we got from those models.
  + We got a model with 0.120211 mean\_absolute\_error result which is a good place to be especially at the beginning of the process of collecting more data.

# Results

Here you will present the main results of the process. We will describe:

* The final amount of data used (total, train, test, etc)
  + Total = 42281 rows × 37 columns
  + Train = 27059 rows × 5 columns
  + Test = 8457 rows × 5 columns
  + Dev = 6765 rows × 5 columns
* The number of outliers and the way of treating them
  + visit\_year 1
  + visit\_week 30
  + restaurant\_id 412
  + sum\_visitors 120
  + max\_visitors 28
  + min\_visitors 10
  + avg\_visitors 17
  + count\_visitors 1
  + sum\_reserve\_visitors 0
  + max\_reserve\_visitors 0
  + min\_reserve\_visitors 0
  + avg\_reserve\_visitors 0
  + count\_reserve\_visitors 0
  + sum\_diff\_days\_between\_reserve\_visit 70
  + max\_diff\_days\_between\_reserve\_visit 19
  + min\_diff\_days\_between\_reserve\_visit 1
  + avg\_diff\_days\_between\_reserve\_visit 6
  + count\_diff\_days\_between\_reserve\_visit 0
  + number\_of\_holiday\_days\_visit 1
  + area\_id 3
  + genre\_id 9
  + station\_id 24
  + station\_latitude 1.010
  + station\_longitude 4.415
  + station\_vincenty 3.380954
  + station\_great\_circle 3.382398
  + restaurant\_latitude 1.001666
  + restaurant\_longitude 4.410035
  + number\_of\_restaurants\_in\_area 370
  + weekend\_visitors 240
  + After testing the outliers my conclusion was not to remove them
* The amount of missing values and the methods used for imputing them
  + visit\_year 0
  + visit\_week 0
  + week\_year\_visit 0
  + restaurant\_id 0
  + air\_store\_id 0
  + sum\_visitors 0
  + max\_visitors 0
  + min\_visitors 0
  + avg\_visitors 0
  + count\_visitors 0
  + sum\_reserve\_visitors 0
  + max\_reserve\_visitors 0
  + min\_reserve\_visitors 0
  + avg\_reserve\_visitors 0
  + count\_reserve\_visitors 0
  + sum\_diff\_days\_between\_reserve\_visit 39136
  + max\_diff\_days\_between\_reserve\_visit 39136
  + min\_diff\_days\_between\_reserve\_visit 39136
  + avg\_diff\_days\_between\_reserve\_visit 39136
  + count\_diff\_days\_between\_reserve\_visit 0
  + holiday\_flg\_visit 0
  + holiday\_week\_visit 0
  + number\_of\_holiday\_days\_visit 0
  + area 0
  + area\_id 0
  + genre 0
  + genre\_id 0
  + station\_name 0
  + station\_id 0
  + station\_latitude 0
  + station\_longitude 0
  + station\_vincenty 0
  + station\_great\_circle 0
  + restaurant\_latitude 0
  + restaurant\_longitude 0
  + number\_of\_restaurants\_in\_area 0
  + weekend\_visitors 0
    - I replaced the variables with missing values with categorical variables that I created.
* The distribution of the data (timeframes)
  + From 1\_2016 to 16\_2017 week\_year
* The methods used to transform the data and to generate new features.
  + Transfer the columns with Null values to new categorical features

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how you got the final prediction. You have to discuss the limitations of the model, when it can be used and when not.