**Recruit Restaurant Visitor Forecasting**

*Author(s):*

Ayala Shvarzman

Arnon Kleinman

# Introduction

The running of a successful and profitable restaurant is not easy to say the least...

Even if you hired the right cooks and staff, the food is delicious, your location is prime, and your service level is excellent, you could still go under.

Many successful restaurants go belly up, it's a difficult to succeed in this business, and you need to save your money wherever you can.

We decided to make a predictive model that will help cut expenses and save money.

Our focus is the on number of customers a restaurant should expect each day, and by so to save a lot of money on buying excess raw materials that will be dumped at the end of the day, or by reducing the use of unnecessary labor, and by planning campaigns in advanced to attract customers on weaker days.

This paper examines the number of customers restaurants in japan have each day.

We were provided with data from two sites, "Hot Pepper Gourmet" (a restaurant review service) and "AirREGI" (a restaurant point of sales service).

"Hot Pepper Gourmet" (hpg): similar to Yelp, here users can search restaurants and also make a reservation online.

"AirREGI / Restaurant Board" (air): similar to Square, a reservation control and cash register system

The data includes number of visitors per restaurant per day, number of reservations throw the two sites, information about the restaurants, location, area, what kind of food it serves, we were also provided with information about the date, is it a holiday or not.

We added data from the web about weather and statistical data per prefecture (socioeconomic state, population density, etc...)

Our target optimization metric is the Root Mean Squared Logarithmic Error.

The RMSLE is calculated as:

Where:

**n** - is the number of observations

**pi** - is our predicted count

**ai** - is the actual count

**log(x)** - is the natural logarithm of **x**

We seek to identify the models that result in predictions which minimize this error.

# Methodology (Project design)

## Data

Here you have to describe how do you plan to manipulate the data. For this you have to answer to the following questions:

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How you will deal with outliers?
* How you will deal with missing values
* 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
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* 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?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

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

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new 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 did you get the final prediction. You have to discuss also the limitations of the model, when it can be used and when not.