**Data Science Project Protocol –**

**Favorita Sales Prediction**

*Author(s):*

Alaa Kabaha

Yuval Erez

# Introduction

Favorita is a chain of grocery stores in Ecuador with multiple locations, each location is selling mutliple products. In this project we will predict sales of different products for the stores across Ecuador, on different dates.

The outcome (“y”) is the number of sold units for an item, per store and per date.

Known factors that might affect sales are:

* Promotions – if a product is being promoted might affect how many units will be sold
* Is item perishable
* Oil prices – Ecuador is an oil dependent country and it’s economics is sensitive to changes in oil prices, therefore grocery shopping might be also affected by oil price changes
* Holidays
* Date when salaries are paid in the public sector – might cause increased sales
* Significant earthquakes (such as on April 16, 2016) – which causes people to stock supplies and increases sales

The Favorita stores are located on different cities and states throughout Ecuador, where similar stores are grouped into clusters. Some items were not sold on specific stores on specific dates.

The training data is very big (4GB) and containing multiple locations/items/years, therefore to overcome technical difficulties we might choose to narrow it down by using specific years and/or specific stores (e.t stores with the highest amount of sales).

Training data has 16% of NAN values for *onpromotion* field.

# 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.