**Data Science Project Protocol**

*Author(s):*

Ariel Shafir & Tzvika Sandrov

# Introduction

Here you have to give some known facts about the field you will work on.

Try to focus on the problems that are most common and then state the goals of the project.

* You can try to answer to the following question:
* Which questions do we want to answer ?
* What is known about the problem?
* How we define the outcome(s)?
* What is known to influence the outcome?
* Does we have any possible new knowledge that has not been in use before?

This part must be between half to one and half page.

Brick-and-mortar grocery stores are always in a delicate dance with purchasing and sales forecasting. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leaving money on the table and customers fuming.

The problem becomes more complex as retailers add new locations with unique needs, new products, ever transitioning seasonal tastes, and unpredictable product marketing. [Corporación Favorita](http://www.corporacionfavorita.com/), a large Ecuadorian-based grocery retailer, knows this all too well. They operate hundreds of supermarkets, with over 200,000 different products on their shelves.

[Corporación Favorita](http://www.corporacionfavorita.com/) has challenged the Kaggle community to build a model that more accurately forecasts product sales. They currently rely on subjective forecasting methods with very little data to back them up and very little automation to execute plans. They’re excited to see how machine learning could better ensure they please customers by having just enough of the right products at the right time.

In this competition, you will be predicting the unit sales for thousands of items sold at different Favorita stores located in Ecuador. The training data includes dates, store and item information, whether that item was being promoted, as well as the unit sales. Additional files include supplementary information that may be useful in building your models.

## **File Descriptions and Data Field Information**

### **train.csv**

* Training data, which includes the target unit\_sales by date, store\_nbr, and item\_nbr and a unique id to label rows.
* The target unit\_sales can be integer (e.g., a bag of chips) or float (e.g., 1.5 kg of cheese).
* Negative values of unit\_sales represent returns of that particular item.
* The onpromotion column tells whether that item\_nbr was on promotion for a specified date and store\_nbr.
* Approximately 16% of the onpromotion values in this file are NaN.
* **NOTE:** The training data does not include rows for items that had zero unit\_sales for a store/date combination. There is no information as to whether or not the item was in stock for the store on the date, and teams will need to decide the best way to handle that situation. Also, there are a small number of items seen in the training data that aren't seen in the test data.

### **test.csv**

* Test data, with the date, store\_nbr, item\_nbr combinations that are to be predicted, along with the onpromotioninformation.
* **NOTE:** The test data has a small number of items that are not contained in the training data. Part of the exercise will be to predict a new item sales based on similar products..
* The public / private leaderboard split is based on time. All items in the public split are also included in the private split.

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

* A sample submission file in the correct format.
* It is highly recommend you zip your submission file before uploading!

### **stores.csv**

* Store metadata, including city, state, type, and cluster.
* cluster is a grouping of similar stores.

### **items.csv**

* Item metadata, including family, class, and perishable.
* **NOTE:** Items marked as perishable have a score weight of 1.25; otherwise, the weight is 1.0.

### **transactions.csv**

* The count of sales transactions for each date, store\_nbr combination. Only included for the training data timeframe.

### **oil.csv**

* Daily oil price. Includes values during both the train and test data timeframe. (Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices.)

### **holidays\_events.csv**

* Holidays and Events, with metadata
* **NOTE:** Pay special attention to the transferred column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.
* Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

## **Additional Notes**

* Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
* A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.
* This supermarker is the bigger of Ecuador
* Ecuador is divide in 3 principal regions: Costa, Sierra and Oriente. This supermarket have more store in the Sierra. The principal competitor have more store in the Costa. The principal city in the Sierra is Quito, the capital and the principal city in the Costa is Guayaquil. You can see this predominance in the Sierra in the sales in Quito and the next in Guayaquil . If think that is convenient include integrate information about population of city, if you see this information you can see for example that Guayaquil is the bigger of Ecuador. I think that is posible depend too of the population and this information can help the forecast.
* This division affect in some forms, for example the types of the food between Costa y Sierra is sligthly different, the school clases and the weather too.
* The vacation in the Sierra its between July and August and if possible that the less sales is in August, the other side the vacation in the Costa its between February and March, and if possible that is cause of less sales is in February. It's is other interesting information for the forecast.
* It's possilbe that vacation affect too the type of product purcharse, because this have relation with the school clases.
* Normally is more interesating information about the City more than the State.
* The state Manabi is where had more effect the earthquake and the citys in this state like Pedernales, Manta, Portoviejo and other cities.
* Other possible effect in the grocery and beverage is the weather, normally between june and august in the Sierra is warm and in the Costa is between december and march.

# 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.