**Data Science Project Protocol**

*Author(s):*

# Introduction

*“AI technologies could eliminate many levels of manual activities in areas such as promotions, assortments, and supply chain. AI will enable retailers to increase both the number of customers and the average amount they spend by creating personal and convenient shopping experiences.” — McKinsey Global Institute (2017)*

Seeing the number of big retail chains that are closing stores (think Sears and Payless) may support a surface-level impression that retail is dying out. But that could not be further from the truth. The industry is booming in the digital space. The revolution started by companies like Amazon and eBay has led to huge challenges for the traditional retail business model, but also massive potential for retailers and consumers alike.

This also means that there is an increasing shift towards optimization and efficiency, and a shift away from excess and waste. Retail is well-placed to benefit from the intersection of Artificial Intelligence, machine learning and big data. There is a need to manage and track a large number of items across various categories, track consumers’ shopping habits and above all, maintain a compelling brand that keeps consumers coming back. Today’s consumer wants to keep up with the latest trends, but also craves convenience; hence, the popularity of subscription boxes and online shopping. A recent survey of retailers worldwide identified **cost savings**, **enhanced decision-making** and **process automation** as some of the main areas that AI has the potential to impact meaningfully[[1]](#footnote-1). (Chandran, Jul 2018)

In the presented project we have done, we looked in to the sales out data of a grocery retail chain that supports more the 100 franchises.

The franchises placing orders using the retailer web portal, later those orders are being sent to the manufactures or the retailer warehouse.

As can expected, the demand is varying from franchise to franchise and from time to time. A lot of causes effecting the demand, among others are: price and discounts. time of the year, holidays, weather, product availability and more.

A note that should be taking in to account are the delivery days, in the project below we measures all the parameters on a week based, since we are not dealing with the sale out of the franchises but with the sale out of the retailer to the franchises, the days of deliverables are, in most cases, fixed i.e. manufacture X deliver its goods to franchise Y every Monday, there warehouse delivers to franchises Z every Tuesday and so on.

In this project we are focusing on the purchasing orders form a few perspectives:

* We wish to identify patterns in the PO
* We wish to exam how "external causes" are affecting the franchises purchasing.
* We wish to see if franchises clustering is possible and based on which parameter
* We wish to build a forecasting model allowing the retailer (and its manufactures) to plan more accurately the availability of the goods

# Methodology (Project design)

## Data

The data we are using has been generated by the ERP system of the retailer and includes the sale out to the franchises in a period of 6 month.

The data includes the franchise ID, date of order, products ID (SKU), quantities and price.

Please note that the data is owned by the retailer and granted to be use in this exercise only.

In addition to it, we are using a wab public based information such as weather and week clustering (split the data to weeks).

Overall the data includes more then 600,000 order lines (before cleaning) covering a period begin at March 2018 until August 2018, as we refer to grocery retail, this period includes Peshach and the summer holidays, two periods that requires a special attention.

### Training and test

In this project we didn’t limit ourselves to a constant set of training and test data i.e. a few models have been operated,

For predictive we have used approximate 85% of the data as turning data and 15% as testing.

For clustering purposes, we have performed a few tests:

1. Full data set clustering, allowing us to see how the algorithm split the data (we are feeding the PO and price and looking to see of its finds similar franchises)
2. Train and tests data where 90% of the data is for training and we test the 10% to see if they aligned with section 1 above. (we wish to exam if a new franchise is added to the "right" group)

Note that originally the franchises are not clustered, and we are looking to find a method of clustering (in section 1) there are a few clusters we can think of and should be tested, for example: based of the quantities, based on the weekly purchases or the monthly purchases, based on number of different SKU's. (we don’t have the geographical data of the franchises and therefore it cannot be trained)

### Outcomes

We have two expected outcomes:

1. Quantities: the main purpose of the project is to forecast the demand. Since the manufacturing should (as much as possible) be aligned with the demand, our outcome should point to (1) the quantities that should be manufactured to meet the demand (2) allowing more accurate pre-orders that are performed by the retailer planners (in case that the product is delivered from its warehouse).
2. Clustering: we assume that franchises clustering can asset the retailer in products promotions, for example: we are estimating that franchises in the same cluster would purchase the same products.

Confounding: there are a few external confounders that can affect the results and reside in data that we don’t have. One example would be the franchise line of credit (LoC), the franchises are limited to a lined of credit that the retailer allow them (open credit) , and the franchises are not allowed to exceed this LoC. In most cases this LoC is manage properly, however it might affect the purchases in special periods such as Pesach where the pre-holidays purchase are dramatically increase.

Source of bias: we don’t see any source of bias in the data we have.

### Data exploration strategy

Initially we have used Excel for data exploration and data unification, for data visualization we used Tableau.

### Enriching the data

We used data enrichment allowing the different models to examine the data

Data aggregation was the main enrichment technique we used

The data enrichment included, the total cost of an order, total cost of orders in a week and a month. Total amount of SKU's per franchise and total amount of pouched products (quantity).

Also, data enrichment allow us to re-order the data based on variety of categories, for example: top 10 in: price, SKU, quantities act.

In addition, we are using a holiday table (date of national holidays) to examine the effect of it on the purchased quantities

### Outliers

Generally, since we are dealing with real data outliers should not exist, even so, those will be handle based on the cause of each, errors in the data, miscalculation or any other cause i.e. it might be that two orders have been joined causing large than expected quantities, in that case we are splitting the order to two.

### Missing values and corrupted data:

We identified approximately 170 lines (out of 600K) with corrupted data or missing values, those lines are mostly in the boundaries of the data and therefore we decide to remove them from the dataset

## Models

### The data

Initially we sort the data based on quantities, in that manner we decide to use the top 30% SKU's (in quantity not price). This decision allowed us to have enough data on each SKU (app 10K per SKU with total of app 200K lines out of 600K).

### Train, validate and test

As mentioned above, we are analyzing the data on a week bases, we have operated the following models on the data:

1. Random: 80% of the data is randomly chosen, the other 20% are used for tests
2. Weekly based: the weeks of 5 month are used for training (week by week) while the weeks of the 6th month are tests.
3. Month based: the first 3 weeks of each month are for train data, the 4th week of each month is test.

### Data balance

Data balancing is not needed in the grocery market, a tolerance is acceptable in both direction

### Subsampling

The data we use is divided to franchises, however a sample data of each does not represent its purchasing history and trends, therefore stratified sampling is not applicable for our data subsampling

### Technique

The model technique is regression due to the data model,

In the data side we have, on a weekly basis:

1. The number of franchises who pouched this SKU
2. The 1st ,2nd ,3rd or 4th week of the month
3. Holiday (yes/no)
4. Average weather of this week (cold/ mid/ worm)
5. (The outcome is) the purchased quantity (the Y parameter)

In addition we are using unsupervised clustering in order to find (or not fine) groups of franchises who act a like.

### Cross validation

Since we have a limited amount of data, we are using K fold cross validation on it. It shall allow us to provide more accurate results. Regarding bootstrapping, we don’t see a point using it since we don’t have different models to apply on the data.

### Measures to evaluate the model

We are using root mean squared error or root mean squared deviation, using RMSD we measure the difference between the actual values and the estimated values.

Since we are calculating "large numbers" (i.e. the total quantity of a product that has been bought by all the franchises together) the estimation can have an error, however it is important to understand its magnitude and if it's positive or negative (over stock or out of stock).

We find this method as the most appropriate to use while dealing with finite numbers.

### Assembling

We intend to test the well-known model that fit to structured data. And to use the most accurate one. At this stage, since we are looking for a single result, we find it had to believe that a combination of a few models would provide better results, having said that, since we intend to cluster the database, it might be that different models are more suitable for different clusters. To conclude, at this stage we don’t know.

## Deployment of your model

### QA

TBD

### Final user

We have two final users (1) the manufactures and (2) the chain planner;

* The manufactures – they are planning the production based on the demand, a better forecasting of the demand shall optimize the production and supply chain operation.
* The chain planner – the chain planned place orders on behalf of the warehouse to the manufactures, a warehouse procurement is different from the end user, a purchasing order should be of high volume and for a long duration, an accurate planning of the warehouse reduce the overstock or our of stock of the warehouse in addition it assist in stock management in terms of expiration dates and such.

### Presentation:

The prediction will be presented in a table wherein it would see the forecast purchasing orders of each product for a specific duration of time (for the coming week, two weeks, one month)

### Training and interpretation

We don't think that it would be necessary, an expert in the field would understand the outcomes.

### Platform

At this stage of the project, as we are dealing with relatively small amount of data a standard computer with the needed environment installed will do the job. In the future, if necessary, it can be exported to a cloud services utilizing GPU's or TPU's.

### Model updates:

It is difficult to estimate since it related very much to the "real life" results it provides. As the amount of data increase and the model has more data to learn from, we believe that it would become more accurate. In case that it will not happen then a reevaluation of the model would be needed.

# 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.

1. <https://towardsdatascience.com/disruption-in-retail-ai-machine-learning-big-data-7e9687f69b8f> [↑](#footnote-ref-1)