**Business problem:**

A given food chain with more than 400 stores in Brazil has difficulties in forecasting the production of each item each day. Calculating wrongly the number of items produced can generate two sorts of problems

1. More items are produced than sold: The kind of food produced can not be stored from one day to another and therefore what is not sold has to be discharged.
2. Less items are produced than the demand : In this scenario the company loses sales

Therefore we have a situation where predicting correctly how much you have to produce of each item each day is important.

Today this forecast is done based on experience but is far from accurate. Average loss (too much or too little) spins around 10%.

Here it is important to put the management perspective: It is their view that today they relay too much in the people experience and would like to have some tool which would free them from this human dependence. In addition of that they believe that the level of error is high and could be reduced.

Therefore, from the management perspective a system capable of predicting the sales would be worth having even if it doesn´t perform better than the current process. Equal performance would be acceptable. In addition of that there was an understanding that the system would be able to improve its performance overtime as more and more historical data is added to its reference database. (Machine learning effect)

**Analytical problem**

The problem is how to build a model which effectively predicts the demand with a level of assertiveness equal or superior to the current one and improves overtime. Listening the thoughts of the people currently in charge of making this forecast we were informed that they believe that there is correlations between the following factors:

1. Day of the month (payment days usually have bigger demand)
2. Day of the week (Fridays, Saturdays, Sundays and holidays usually have a big demand)
3. Month (Holidays months usually have bigger sales – In Brazil Dez-Jan-jun-Jul)
4. Weather (temperature, rain and sun have an impact in what people eat)

This particular company sales its food through several channels: 1) directly from its stores, 2) through a web deliver service and 3) through a call-center. In our study we are not going to differentiate these channels just counting the total volume sold of each item each day.

Here it is worth mentioning that the insigns provided by the people in charge of the process today should be seen with a grain of salt given the fact that they know that a system like that would be built to replace them. Therefore all these assumptions must be checked against the hard data.

**Datasets**

To be able to identify the patterns we select the city of São Paulo which alone responds for almost 40% of the sales. This is a simplifying strategy given the fact that if the process works for this city we easily deploy it in the others.

**Getting the data**

Initially we managed to get the weather stations measurements in São Paulo for the whole year of 2018 and January of 2019. It is public information available at the website:

<http://www.inmet.gov.br/projetos/rede/pesquisa/>

Secondly we managed to get from the company the sales by type by day. In sequence we prepared this data crossing these two files unifying them by date. Our database format was:

Data – Day of the weather measurements and sales

Precipitac – Volume of rain in millimetres per square meter during the day

Tempmax – Max temperature during the day

Tempmin – Minimum temperature during the day

Tempmed – Average temperature during the day

Umidade – Level of humidity in the air

Insolacao – Level of sun

Diasemana – day of the week abreviation

Diasem – weight of the day of the week considering the average volumes sold

Mes – Weight of the month considering the average sales volumes

Desserts – number of desserts sold in the day

Pizzas – Number of pizzas sold in the day

Beverage – Number of Beverage sold in the day

Cbmaker – Number of a special dish sold in the day

Combos – Number of combos sold in the day

Sfiha – Number of sfihas sold in the day

kit1 – Number of kits 01 sold in the day (Is a dish with a gift)

kit2 - Number of kits 02 sold in the day (Is a dish with a gift)

snack – Number of snacks sold in the day

pastas – Number of pastas sold in the day

dishes – Number of lunches sold in the day

promotion – Number of promotions sold in the day (This is episodic and may not be counted)

savory – Number of savory sold in the day

salads – Number of salads sold in the day

total – Total number of items sold in the day

The idea is to use 80% of the measurements as our training data and 20% as our test data.

The meaning of the numbers in the field diasem is:

1. Monday, Tuesday, Thursday
2. Friday
3. Saturday and Sunday
4. Holidays

The number represents the weight of the day regards sales. This weight reflects the view of the current planners regards sales and we need to check if the assumption holds. Analyzing the sales per day of the week we have the following graphic:



The current assumption of separating the days of week into four categories seems to be a bit wrong , the graphic shows three ranges as follows:

1. Monday – Tuesday – Wednesday-Thursday – Range 11% - 13% of the sales
2. Friday and Saturday – Range 17% and 19% of the sales
3. Sunday – Range 14% a 15% of the sales

The holidays match the volumes of FRI and SAT falling into the Range 2. The distribution by type of item goes as shown:



This analysis suggests that we change the days classification in the following manner:

1. Monday – Tuesday – Wednesday-Thursday – Range 11% - 13% of the sales
2. Sunday – Range 14% a 15% of the sales
3. Friday and Saturday – Range 17% and 19% of the sales

The current meaning of the numbers in the field mes is:

1. March, April, May, August, September, October, November
2. February and July
3. January ,June and December

The number represents the weight of the month regards sales, this weight reflects the view of the current planners regards sales and we need to check if the assumption holds. Analyzing the actual sales monthly by month classification we saw the following:



That suggests that there is a differentiation among the months as follows:

1. 1.500.000 items sold monthly (8,61% above the baseline)
2. 1.380.000 items sold monthly (Baseline)
3. 2.079.000 items sold (50% above the baseline)

That suggests the change 1 with 2 as a classification for the mes to keep it aligned with the sales volumes:

1. 1.380.000 items sold monthly (Baseline)
2. 1.500.000 items sold monthly (8,61% above the baseline)
3. 2.079.000 items sold (50% above the baseline)

Note that there is a mix between the climatic factors and what we could call seasonal effect however the inclusion of these factors were necessary given the fact that although it is assumed that the weather do influence the choice of the food the volumes are mostly influenced by the day of the week and the season.

Note that the objective is first to identify if the listed factors are in fact defining the demand and identify which ones are the most relevant.

Stablishing a correlation using the Kendall method between the demand for specific items. Here we are going to limit ourselves to the following items:

Desserts – number of desserts sold in the day

Pizzas – Number of pizzas sold in the day

Beverage – Number of Beverage sold in the day

Sfiha – Number of sfihas sold in the day

snack – Number of snacks sold in the day

pastas – Number of pastas sold in the day

dishes – Number of lunches sold in the day

savory – Number of savory sold in the day

salads – Number of salads sold in the day

The reason for that is the fact that the others are not regular items but some sort of promotion only made available for limited span of time. Items expurgated from the analysis:

Cbmaker – Number of a special dish sold in the day

Combos – Number of combos sold in the day

kit1 – Number of kits 01 sold in the day (includes a gift)

kit2 - Number of kits 02 sold in the day (Includes a gift)

promotion – Number of promotions sold in the day (This is episodic and may not be counted)

Seasonal parameters:

Diasem – weight of the day of the week considering the average volumes sold (1,2 or 3)

Mes – Weight of the month considering the average sales volumes (1,2 or 3)

Using the following code in R:

“setwd("/Curso-ML/Assignment-1/")

variables <- read.csv("data-06.csv")

summary(variables)

variables1 <- variables[17:32]

summary(variables1)

library(corrplot)

cor.test(variables1$diasem, variables1$pizzas)

cor.test(variables1$diasem, variables1$pizzas)

forcorrplot <-cor(variables1)

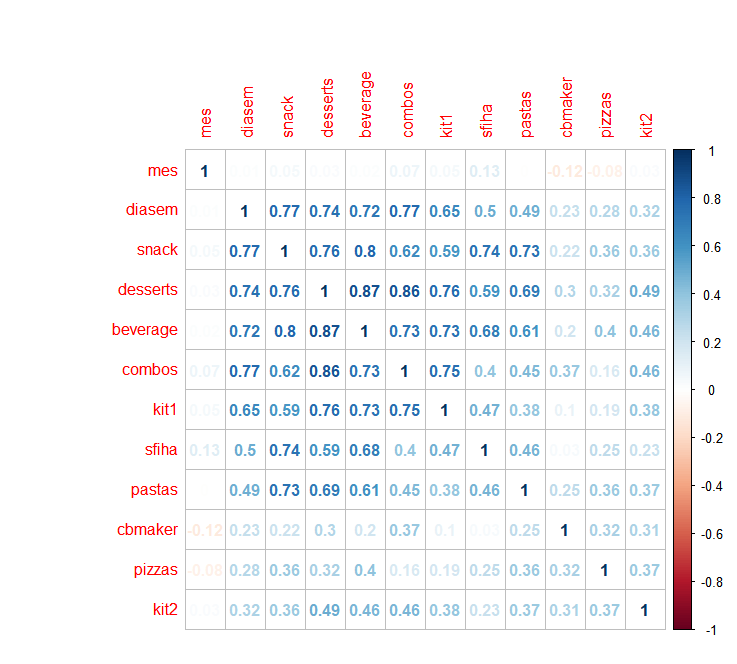
corrplot(forcorrplot)

corrplot(forcorrplot, upper="number", lower="color", order="hclust")

corrplot(forcorrplot, method="number",order="hclust" )

cormat(variables)”

We managed to see the correlation between the day of the week and month and the consumption of the several items:



As we can see the consumption of some items vary more them others as the days of the week change.

Now we have to see if there is correlation between the weather factors and the consumption by item:

Precipitac – Volume of rain in millimetres per square meter during the day

Tempmax – Max temperature during the day

Tempmin – Minimum temperature during the day

Tempmed – Average temperature during the day

Umidade – Level of humidity in the air

Insolacao – Level of sun

Remembering that doing the correlation we are going to identify the variable R which can varies from +1 to -1 indicating that a relationship exists between the two variables from absolute direct correlation (+1), no correlation (0) to inverse correlation (-1). Using R studio and the command “corrplot” we managed to see the following:

“setwd("/Curso-ML/Assignment-1/")

variables <- read.csv("data-06.csv")

summary(variables)

variables$diasemana <- NULL

variables1 <- variables[10:31]

summary(variables1)

library(corrplot)

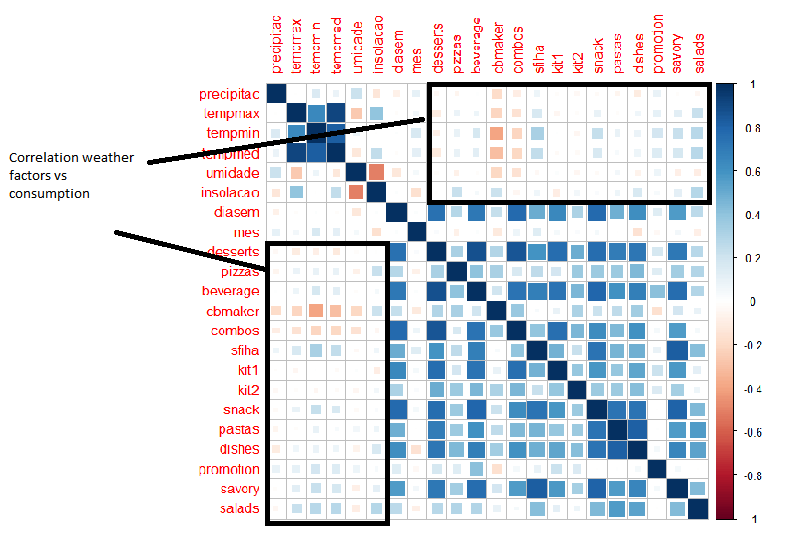
cor.test(variables1$diasem, variables1$pizzas)

cor.test(variables1$diasem, variables1$pizzas)

forcorrplot <-cor(variables1)

corrplot(forcorrplot)

corrplot(forcorrplot, method = "square")”



As we can see in the graphics there are some correlation between weather parameters and the consumption. However the weather conditions are not so determining as the day of the week and the month of the year.

Therefore building a model to guess the demand will imply the definition of the average demand of the day of the week in a given month and in sequence calibrate this demand by the weather conditions.

To archive that we grouped the demand by the combination of type of day vs type of month with nine categories:



Using this code in R we manage to get the average volumes of sales per type of product per range:

“setwd("/Curso-ML/Assignment-1/")

data <- read.csv("data-06.csv")

data1 <- read.csv("dia-sem-me.csv")

summary(data)

summary(data1)

print(data1)

len3=nrow(data1)

print(len3)

print(data1[1,4])

print(data[1,17])

library(stringr)

len2=nrow(data)

print(len2)

contador2 = 1

while (contador2<=len2){

print(contador2)

print(contador2)

print(contador2,17)

len3=nrow(data1)

print(len3)

contador1 = 1

while (contador1<=len3){

print(contador1)

print(data1[contador1,2])

print(data[contador2,17])

if (data1[contador1,2] == data[contador2,17] && data1[contador1,3] == data[contador2,18]){

print(data1[contador1,1])

data1[contador1,4]<-data1[contador1,4]+data[contador2,19]

data1[contador1,5]<-data1[contador1,5]+data[contador2,20]

data1[contador1,6]<-data1[contador1,6]+data[contador2,21]

data1[contador1,7]<-data1[contador1,7]+data[contador2,22]

data1[contador1,8]<-data1[contador1,8]+data[contador2,23]

data1[contador1,9]<-data1[contador1,9]+data[contador2,24]

data1[contador1,10]<-data1[contador1,10]+data[contador2,25]

data1[contador1,11]<-data1[contador1,11]+data[contador2,26]

data1[contador1,12]<-data1[contador1,12]+data[contador2,27]

data1[contador1,13]<-data1[contador1,13]+data[contador2,28]

data1[contador1,14]<-data1[contador1,14]+data[contador2,29]

data1[contador1,15]<-data1[contador1,15]+data[contador2,30]

data1[contador1,16]<-data1[contador1,16]+data[contador2,31]

data1[contador1,17]<-data1[contador1,17]+data[contador2,32]

data1[contador1,19]<-data1[contador1,19]+1

}

contador1 = contador1 + 1

}

contador2 = contador2 + 1

}

len4 = nrow(data1)

contador3 = 1

while (contador3<=len4){

print(contador3)

print(data1[contador3,2])

print(data1[contador3,1])

data1[contador3,4]<-data1[contador3,4]/data1[contador3,19]

data1[contador3,5]<-data1[contador3,5]/data1[contador3,19]

data1[contador3,6]<-data1[contador3,6]/data1[contador3,19]

data1[contador3,7]<-data1[contador3,7]/data1[contador3,19]

data1[contador3,8]<-data1[contador3,8]/data1[contador3,19]

data1[contador3,9]<-data1[contador3,9]/data1[contador3,19]

data1[contador3,10]<-data1[contador3,10]/data1[contador3,19]

data1[contador3,11]<-data1[contador3,11]/data1[contador3,19]

data1[contador3,12]<-data1[contador3,12]/data1[contador3,19]

data1[contador3,13]<-data1[contador3,13]/data1[contador3,19]

data1[contador3,14]<-data1[contador3,14]/data1[contador3,19]

data1[contador3,15]<-data1[contador3,15]/data1[contador3,19]

data1[contador3,16]<-data1[contador3,16]/data1[contador3,19]

data1[contador3,17]<-data1[contador3,17]/data1[contador3,19]

contador3 = contador3 + 1

}

print(data1)

##write.csv(data1,'/Curso-ML/Assignment-1/dia-sem-me.csv')

class(data$diasem)

print(nrow(data))”

This code generates the following table:



This would be the demand if only these two factors were influencing the actual sales, therefore the challenge is to identify how the weather conditions make the average demand deviate (up and down).