Babar Forecasting

2020

Chapter 1

Goal

1.1 The Problem

1.2 Our Solution

Chapter 2

Data Management

2.1 Selection and Gathering Sales Data

There are too many products to have a prediction for each one. So we decided to gather them into categories. We want to gather them into pertinent groups so that the forecast can really help the bar.

2.1.1 First Selection

The first step is to choose which product we are going to take account of. We choosed to focus our work on the drinks sales that had at least been sold ???. So we exclude any food or other kind of sales.

SQL Request too extract those product :

CREATE TABLE WantedProducts AS SELECT babar_server_purchase.* FROM babar_server_purchase JOIN babar_server_product ON product_id = babar_server_product.id WHERE NAME = 'Leffe' OR NAME = 'Hoegaaden blanche' OR NAME = 'Desperados' OR NAME = 'Smirnoff' OR NAME = 'Pastis' OR NAME = 'Hard' OR NAME = 'Grimbergen' OR NAME = 'Chimay Rouge' OR NAME = 'Chimay Bleue' OR NAME = 'Kro Demi' OR NAME = 'Cidre Demi' OR NAME = 'Pelforth' OR NAME = 'Kwak' OR NAME = 'Kir' OR NAME = 'Kro Pinte' OR NAME = 'Cidre Pinte' OR NAME = 'Cocktail Hard' OR NAME = 'Chimay Blanche' OR NAME = 'Shot' OR NAME = 'Blanche Demi' OR NAME = 'Blanche Pinte' OR NAME = 'Cidre Doux/Brut' OR NAME = 'Ambree demi' OR NAME = 'Ambree Pinte' OR NAME = 'Delirium' OR NAME = 'Rouge Pinte' OR NAME = 'Sangria' OR NAME = 'Karmeliet Triple' OR NAME = 'Duvel' OR NAME = 'Granita Hard' OR NAME = 'Skoll' OR NAME = 'Rouge Demi' OR NAME = '1664 Blanche' OR NAME = 'Chimay bleue' OR NAME = 'Pecheresse' OR NAME ='Cuvee des Trolls' OR NAME = 'Kriek' OR NAME = 'Elephant Pinte' OR NAME = 'Elephant Demi' OR NAME = 'Maredsous Triple' OR NAME = 'Hard Qualite' OR NAME = 'BrewDog Punk IPA' OR NAME = 'JagerBomb' OR NAME = 'Cubanisto' OR NAME = 'Chouffe Pinte' OR NAME = 'Chouffe Demi' OR NAME = 'Corona' OR NAME = 'Tigre Bock' OR NAME = 'Troll Pinte' OR NAME = 'Troll Demi' OR NAME = 'Triple Karmeliet Pinte' OR NAME = 'Triple Karmeliet Demi' OR NAME = 'Paix Dieu 33cL' OR NAME = 'Grim Triple Demi' OR NAME = 'Grim Triple Pinte' OR NAME = 'Cherry chouffe' OR NAME = 'Delirium rouge pinte' OR NAME = 'bush ambree' OR NAME = 'San Miguel';

List of considered product :
So now we have a data table gathering only the product we choose.

2.1.2 Other products that could be considered

2.1.3 Regrouping them into categories

A majority of this product has been coming and leaving from the bar menu so their sales aren't usable one by one. Therefor we decided to group them into pertinent categories. But how to choose this categories ? Here are our first thought.

High Degree Beers

- Chimay Bleu
- Kwak
- Karmeliet Triple
- Duvel
- Chimay bleue
- Maredsous Triple
- Chouffe Pinte
- Chouffe Demi
- Triple Karmeliet Pinte
- Triple Karmeliet Demi
- Grim Triple Pinte
- Grim Triple Demi
- bush ambree
- Delirium
- Elephant Pinte
- Elephant Demi

Normal Degree Beers

- Leffe
- Grimbergen
- Kro Demi
- Kro Pinte
- Pelforth
- Skoll
- BrewDog Punk IPA
- Tigre Bock
- Troll Pinte
- Troll Demi
- Cuvée des trolls
- Paix Dieu 33cL
- San Miguel
- Ambrée Pinte
- Ambrée Demi

Not Beer

- Smirnoff
- Pastis
- Hard
- Kir
- Cocktail Hard
- Shot
- Rouge Pinte
- Rouge Demi
- Sangria
- Granita Hard
- Hard Qualite
- JagerBomb

Special Beers

- Desperados
- Cidre Demi
- Cidre Pinte
- Cidre Doux/Brut
- Cubanisto
- Corona

Aromatized Beer

- Chimay Rouge
- Hoegaaden blance
- Chimay Blanche
- Blanche Demi
- Blanche Pinte
- 1664 Blanche
- Pecheresse
- Kriek
- Cherry chouffe
- Delirium rouge pinte

Other idea : Aromatize beer and Special together

Those are ideas of the categories, at every time at least one product of each category was being sold, assuring the continuity of sales in each categories. The goal is having sales that are homogene on a year scale.

2.1.4 SQL code for product gathering

DROP TABLE IF EXISTS ‘High_Degree_Beer‘;

CREATE TABLE ‘High_Degree_Beer‘ AS SELECT babar_server_purchase.* FROM babar_server_purchase JOIN babar_server_product ON babar_server_purchase.product_id = babar_server_product.id WHERE NAME = ’Chimay Bleue’ OR NAME = ’Chimay Bleu’ OR NAME = ’Kwak’ OR NAME = ’Karmeliet Triple’ OR NAME= ’Duvel’ OR NAME = ’Maredsous Triple’ OR NAME = ’Chouffe Pinte’ OR NAME = ’Chouffe Demi’ OR NAME = ’Triple Karmeliet Pinte’ OR NAME = ’Triple Karmeliet Demi’ OR NAME = ’Grim Triple Pinte’ OR NAME = ’Grim Triple Demi’ OR NAME = ’bush ambrée’ OR NAME = ’Delirium’ OR NAME = ’Elephant Pinte’ OR NAME = ’Elephant Demi’;

DROP TABLE IF EXISTS ‘Normal_Degree_Beer‘;

CREATE TABLE ‘High_Degree_Beer‘ AS SELECT babar_server_purchase.* FROM babar_server_purchase JOIN babar_server_product ON babar_server_purchase.product_id = babar_server_product.id WHERE NAME = ’Leffe’ OR NAME = ’Grimbergen’ OR NAME = ’Kro Demi’ OR NAME = ’Kro Pinte’ OR NAME = ’Pelforth’ OR NAME = ’Skoll’ OR NAME = ’BrewDog Punk IPA’ OR NAME = ’Tigre Bock’ OR NAME = ’Troll Pinte’ OR NAME = ’Troll Demi’ OR NAME =’Cuvée des trolls’ OR NAME = ’Paix Dieu 33cL’ OR NAME = ’San Miguel’ OR NAME = ’Ambrée pinte’ OR NAME = ’Ambrée demi’;

DROP TABLE IF EXISTS ‘Not_Beer‘;

CREATE TABLE ‘High_Degree_Beer‘ AS SELECT babar_server_purchase.* FROM babar_server_purchase JOIN babar_server_product ON babar_server_purchase.product_id = babar_server_product.id WHERE NAME = ’Smirnoff’ OR NAME = ’Pastis’ OR NAME = ’Hard’ OR NAME = ’Kir’ OR NAME

= 'Cocktail Hard' OR NAME = 'Shot' OR NAME = 'Rouge Pinte' OR NAME = 'Rouge Demi' OR NAME = 'Sangria' OR NAME = 'Granita Hard' OR NAME = 'Hard Qualité' OR NAME = 'JagerBomb';

2.2 Time Scale

We think that a Weekly prediction can be done, so we'll consider the sales week by week (Time series approach). If we have difficulties doing it we will consider doing it monthly but then the forecast would be much less usable. (Time series approach)

We will also maybe consider a year index. Indeed promotion aren't all the sames therefor they don't consume the same quantities. Adding a bias for each promotion (which would be compute thanks to the first month of data ? september ?) could maybe increase the accuracy of our estimations.
When to start ? Beginning of 2011 or rentree 2011 ?

2.2.1 Sorting By week

For each week we will sum the sales for each categories. The weeks will be register by a index of week from 1 to 52, and a year. So 2 keys.

I don't think that I can sort the data by week by using SQL, so I'll do a python code that do that. The results will be a table with the sales for each category for each week.

Week	HighDegree	NormalDegree	SpecialBeer	NotBeer
15/03/2014 – 22/03/2014	60	82	31	15
23/03/2014 – 30/03/2014	51	74	29	25
31/03/2014 – 6/04/2014	74	90	50	38
07/04/2014 – 14/04/2014	56	77	23	20

2.3 Other used Data

In order to do our prediction we thing about using the following features :

- Day before next holiday
- Day before next exam
- Day after exam or holiday ?
- Day before next large event ?

We think that those are the main influences on whether people buy more or less drinks in a bar.

2.3.1 Holidays

The holidays data is gather in a CSV file with 3 columns, Year, WeekNo and Holiday. If the week is a Holiday week then the holiday column contains a 1. If it's a work week then there is a 0. After thinking about it we decided to not only consider if the week is or isn't a holidays week but also count the number of week before the next holidays. Indeed we forsee that people tend to buy more in a Bar when a Holiday is near. So the data given to the machine learning model will be the counter to the next holiday.

2.3.2 Exams

The Exams data is also gathered in a CSV file with 3 columns, Year, WeekNo and Exam. If there is a Exam this week the Exam column contains a 1 and a 0 if not. As we did for the holidays what would interest us is a counter to the next exam. But in the case of exams we also want to transcript the possibility of several week of exams following each other. So our idea is to examine the 3 following weeks and ponderate them (the first is more impactfull than the third one, with coefficient 3,2,1).

! HOWEVER ! : We think about being more precise and adding a value '2' that will transcript the importance of some exams.

2.3.3 Promotion

Adding a coefficient considering the promotion ! Every promotion doesn't drink the same, a coefficient should be schoolyear based. SO we just add a feature indicating the "promotion". We hope that the ML model will be able to consider the feature and balance the sales data of each year.

WeekNo	Normal Beer	High Degree Beer	Special Beer	Not Beer	Holiday Counter	Exam Counter
14	254	145	58	24	5	2
15	312	98	75	12	4	1

Chapter 3

Model Selection

Now the core of our work is to choose the best method to predict from our data. Or goal is to look in the dataset for features such as trends, cyclical fluctuations, seasonality, and behavioral patterns.

Here are the algorithm that we read about and could be used to forecast sales :

- KNN Regressor
- Random Forest
- Neural network

3.1 KNN Regressor

The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

3.2 Random Forest

The Decision Tree algorithm has a major disadvantage in that it causes over-fitting. This problem can be limited by implementing the Random Forest Regression in place of the Decision Tree Regression. Additionally, the Random Forest algorithm is also very fast and robust than other regression models. To summarize in short, The Random Forest Algorithm merges the output of multiple Decision Trees to generate the final output.

3.3 Neural Network

See Regularization ? ML course 24 november.

3.3.1 Promotion

Another important feature is the school year !

Chapter 4

Finding the best Forecasting Method

4.1 Comparing Models

In order to compare models we need to choose a metric.

4.1.1 Choosing a Metric

There are the three possible metric that we consider : score (R^2) / Absolute Error (MAE) / Squared Error (MSE). We think that the best metric to use is the mean squared error (RMSE), but in a concrete way the best one is actually the absolute error that is the most indicative as it shows the actual difference that the bar will encounter. We are working with a bar so we think that the best way to evaluate the best one is to consider the MAE only.

4.1.2 Choosing a model

Difference between shuffle and not shuffle is very notable ! So our model is missing something that considerate the year maybe ? We thought adding the promotion as a feature would be a solution but apparently that's not enough. THe results are not greate when not shuffling

4.1.3 Separating the y

We tried separating the type of beer so that each one has its own model but it doesn't improve the results (it worsen them actually)

4.2 Comparing the importance of features

In order to see the importance of features we are going to try different combination of them.

- Promotion : It seems to be a very important feature, when we take it out the results drops
- Exams : It doesn't affect the results as much as promotion but it still taking them up, when dropping this feature we observe a small drop in the results

- Week Number : Very important, as promotion dropping it really worsen our results
- Week to holiday : Like exam

It seems that some other features would be needed, like the schedule of events ! Or the number of students of each year (we could have access to the number of clients ! and not only the sales).... IDEAS OF OTHER FEATURES

4.3 Conclusion on the results

Seeing the graphs shows that the considerable irregularities in the sales of the establishment creates real difficulties when trying to forecast. Indeed our models are not able to forecast the drops or pics in sales that are due to some random particular event that takes place often in an university environment. However the random forest model still propose some notable results with a MAE of 28, with means that in average the model differs of 28 sales each week, which isn't that bad ! Knowing that those product have long expired dates then it isn't a real problem to have +-28 beers at the end of the week ! We could even see if it equalize over time (maybe the week after it compensates).

Chapter 5

Inspirations

https://towardsdatascience.com/sales-forecasting-from-time-series-to-deep-learning-5d115514bfac : Forecasting principles and basis

https://medium.com/analytics-vidhya/walmart-sales-forecasting-d6bd537e4904 : Forecasting at Walmart