Babar Forecasting

2020

Chapter 1

Goal

1.1 Forecasting

1.2 The Problem

For a buisness to be optimized it seems necessary to have an idea about how much is going to be sold, it enable the buisness house to produce or gather the right quantities at the right time. Furthermore it enables to make arrangement in advance for material, equiment or labour.

In a bar or restaurant it is very important to have an approximation of the futur sales as it allow the owner to order the right quantity of food and avoid any problem with waste or expiration (which can quickly become expensive for those kind of buisness).

1.3 Our Solution : Forecasting

Any forecast can be termed as an indicator of what is likely to happen in a specified future time frame in a particular field. Therefore, the sales forecast indicates as to how much of a particular product is likely to be sold in a specified future period in a specified market at specified price.

Our solution is to offer a trustfull estimation of the futur sales for a bar. This would accompany the manager in his/her choices about orders and organisation. We would use multiple machine learning models to have the most precise forecast on a weekly scale, and this by only using a few information like the holiday or exam dates.

Furthermore this forecast would help the management in determining as to how much revenue can be expected to be realised and what shall be the requirement of men, machine and money.

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 have chosen to focus our work on the drinks sales that had at least been sold ???. So we exclude any food or other kind of sales.

So now we have a data table gathering only the product we have chosen to consider.

2.1.2 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

- 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

- 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
- 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.

2.2 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 Features

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

- Day before next holiday
- Day before next exam
- Promotion Index
- 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

3.2 Random Forest

3.3 Neural Network

See Regularization ? ML course 24 november.

Chapter 4

Python Code

4.1 Data Manipulation

The data is stored in Données.

4.1.1 Sales DataFrame

This file gather the function necessary to create a data frame gathering all sales information for each week. There are (by order of apparition) the functions.

- textbfManageNullWeek : It add rows of zeros for week with no sales.
- textbfProductSales : This function takes the sale of a product dataframe in argument and gives an array with the usefull categories, ie for each product it gives a array of the form : $Year|Weeknumber|Numberofsales$ (And the null row are filled by ManageNull Week).
- textbfCreateSalesArray : opens thee CSV files of the sales and made a dataFrame for each table. Then it collect the number of sales for each week by using the two previous functions. It returns a array containing the each array of each product.
- **CreateSalesData** : It take the precedent array in argument and creates an array for each year of the wanted form, ie for each week the sales of each product.
- **CreateSalesFrame** : This function is the gathering of all this file, it returns the previously described data frame by using all the functions.

4.1.2 Holidays and Exams

According to us, one of the most important parameter is the holidays. We will consider the number of weeks (or days ?) to the next holiday (and after ?). The data is gathered in a csv file with for every week if there was a holiday (indicated by a 1) or if there wasn't (indicated by a 0). Our goal is to treat this data in order to have for each week the distance to the next holiday with the convention that the holiday week has a distance of 0. For the exam we are considering the option of pondering the distance to the next exam depending on the number of exams that occurs in a week. But how to do that ?

- openHandEData() : This function just opens the csv file and make it a Dataframe
- distanceToNextHoliday() : This function creates the counter for holidays and add it to the previously made dataframe
- distanceToNextExam() : This function creates a coefficient for each week for exams (as presented before, ponderated) and add it to the previously made dataframe.
- CreateHandE(): DataFrame with holiday and exams.

4.1.3 Promotion

Another important feature is the school year !

Chapter 5

Finding the best Forecasting Method

5.1 Results

5.1.1 KNN Regressor

The first results are very disappointing, by using the KNN regressor with different number of neighbors we find the following results :

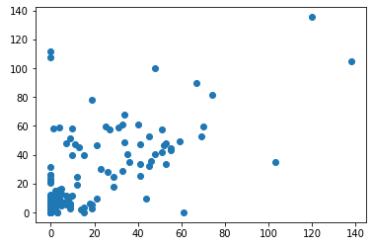

Figure 5.1: 5 Neighbors : A = -0.19

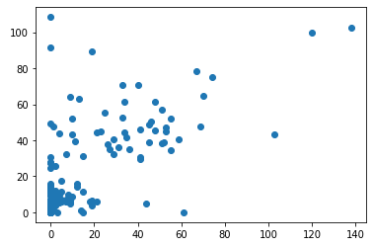

Figure 5.2: 10 Neighbors : A = -0.10

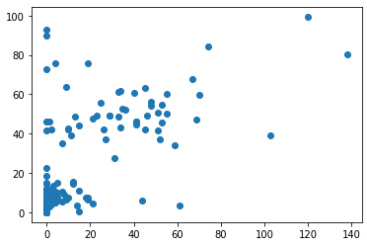

Figure 5.3: 20 Neighbors : A = -18

5.2 Random Forest

5.3 Comparing Models

5.4 Comparing the importance of features

Chapter 6

Next Step

Adding a coefficient considering the promotion ! Every promotion doesn't drink the same, a coefficient should be schoolyear based.

Chapter 7

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

Chapter 8

Individual Repartition

8.1 Maxime

8.2 Mathieu