Data Science test. Innovation and research

Arturo Leos Zamorategui

(Dated: December 26, 2017)

In this report I collect the datasets `demographic`, `transactions` and `products` using the Pandas library in Python. After pre-processing the data, I study the content of the three databases independently or by mixing the data of the three. I include different charts and histograms for this matter. For some of the data, I program multiclass classifiers using Random Forests and K-Nearest Neighbors models using the library scikit-learn in Python. Further, I build Deep Neural Networks using Keras to improve the predictive behaviour of those models. Even if the accuracy improves, it is not significantly higher. Also, by studying the average sales value per week I propose an ARIMA process as a forecast model with parameters $p = 1$, $d = 1$ and $q = 0$. I conclude the study of time series by discussing the possible applications of the DeepAR model proposed by Flunkert *et al* as one might desire to analyse all of the time series grouped by age, for instance.

Keywords: data visualization, random forests, KNN, DNN, time series, ARIMA model, forecasting

I. DATA DESCRIPTION

For this work I analyse three datasets using Pandas library in Python. The first dataset called `demographic` contains personal information regarding 801 customers such as their age, marital status, income and household composition.

The customer has a key associated. The second dataset contains the transactions performed by 2500 householders such as the day and the week the transaction was made, the product bought which has an identfication number, the quantity, the amount of the sale, the store and the discounts. The information of a transaction is associated to the customer key. The third dataset contains information of the products such as the department, the brand and the description of the product. The identification number of the product is also present in this dataset. By identifying the householder keys present in `demographic` and `transactions` we can associate the exact products bought by those customers during the time measured by identifying the product id present both in `transactions` and `products`. More importantly, we can look for buying patterns associated to the age, the householder composition or the gender (in the cases where this is known). Below I try to answer some of these questions.

By linking the datasets, we are able to obtain extreme value statistics such as the best week for stores overall, or the store that received the best revenue or the preferred product by customers. Also, we can answer more specific questions such as the maximum number of products offered in a certain week and what store offered those products.

Moreover, we can study the evolution of some variables in time such as the total or mean sales per day or per week. Or even the money paid by customer per day or per week. Can we detect some buying patterns in time of people from a certain age or a certain householder composition ? For example, if a certain person prefers to make small shopping more frequently or large shopping less often, is that person married or single ?

Further, I associate the money spent by customers of a certain age in each department, or the money spent in each department according to the householder composition. This information can be used to program a recommender system that detects the age and/or the householder composition of a person in order to propose certain products.

Finally, we can build multiclass classifiers in order to extract the buying patterns of different customers according to their age. These models can be used to predict the age or the houdseholder composition from the customers we ignore this information, i.e. from the customers we know the transactions made but that are not registered in the database `demographic`.

II. QUICK VALIDATIONS

To start with we can look at the raw data to visualize better what the datasets contain. We can use different techniques such as pie charts, histograms, heatmaps, or by making some statistics. Let us start by looking at the data contained in `demographic`.

A. Distribution of customers in dataset `demographic`

The dataset `demographic` allows us to visualize the population of customers sampled in the datasets. We can assume that the customers sampled in such dataset represent well the customers sampled in `transactions`. Figure 1 shows the distribution of the customers by minimum income. We observe that the largest group of customers (24%) earn between 50000 and 74000, followed by a 21.5% of customers who earn between 35000 and 40000. Fig. 2 shows the distribution of the population by age. We can identify that the largest proportion (36%) of customers are between 45 and 54 years old followed by a 24.2% between 35 and 44 years old.

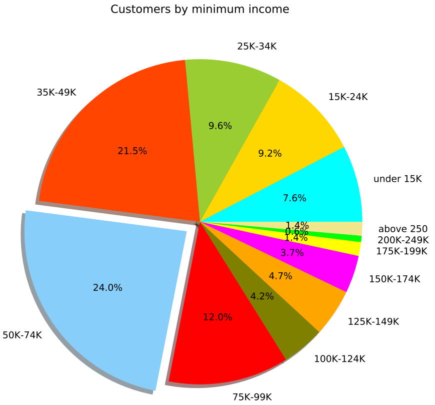

FIG. 1: Distribution of customers by minimum income. 24% of the customers earn between 50K and 74K.

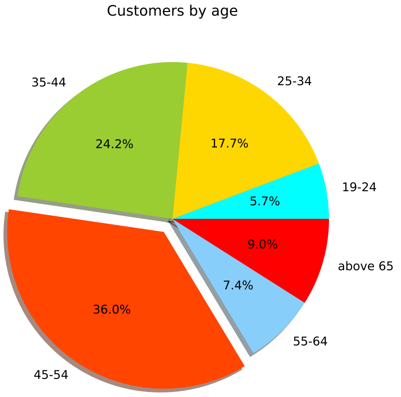

FIG. 2: Distribution of customers by age. 36% of the customers are between 45 and 54 years old.

B. Distribution of money paid per week

The dataset `transactions` can also give us some information about the money spent by customers (which is not part of the datasets) computed from the sales value, the coupon offered by the manufacturer who reimburses the retailer and the discount offered by the retailer in the case the customer has a loyalty card. Therefore, the money spent by the customers with retailer's loyalty card:

$$\frac{\text{sales value} - (\text{retailer discount} + \text{coupon match})}{\text{quantity}}, \quad (1)$$

and without retailer's loyalty card:

$$\frac{\text{sales value} - \text{coupon match}}{\text{quantity}} \quad (2)$$

Figure 3 shows the distribution $P(m)$ of the money spent by customers in a week. We observe that m follows an exponential behaviour meaning that $P(m) \sim \exp(-m/m_0$ with $m_0 = 75.8$ the average amount spent in a week.

In the discussion below we observe the time series of the money spent by customers throughout all the period.

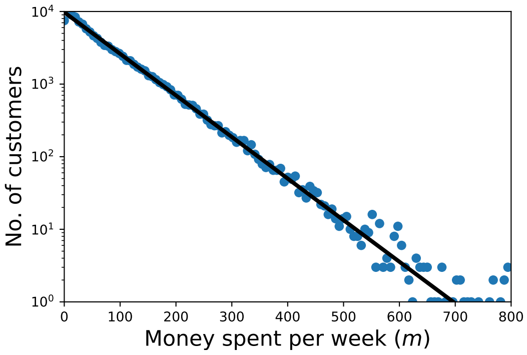

FIG. 3: Distribution of money spent per week (denoted m) by customers in log-linear scale. The distribution of the money paid by customers per week follows an exponential law as $P(m) \sim \exp(-m/75.8)$.

As we discuss below, we might be interested in identifying buying patterns for differents customers belonging to a certain group. A simple idea might be to classify series in two groups: those with a loyalty card and those without loyalty card. Are their buying patterns different with out without a loyalty card ?

C. Distribution of sales per product during the whole period

Further, we can look at the histogram $P(n)$ of the money received by retailers (sales value) during all the period from the different products offered. In this case, such quantity seems to follow a power-law as shown by the solid line in Fig. 4. Such scale-free behaviour is given by the following expression $P(n) \sim n^{-\zeta}$ with $\zeta = 2.6$. A power-law behaviour observed for the amount of money received by product in a given period means that cheap products are sold much more times than expensive ones. This might seem trivial at first sight. However, it can tell us how demand works and the money a retailer can lose by the lack in stock of a given product. Many man-made processes follow a scale-free behaviour such as the frequency of words in a text[1] or the population size in cities[2].

D. Statistics of the features in datasets

In Table I we summarize some of the main statistics of the features in the datasets including the number of products offered in week 50 and the best selling product. These properties can be extracted straighforwardly from the datasets by grouping the features in groups. For instance, if we group the sales value per store per week we can look at the evolution of the revenues for a given retailer in time. In this section we are interested in average properties so we content ourselves with the average sales

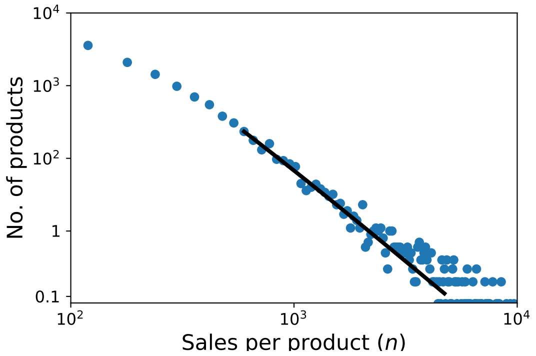

FIG. 4: Distribution of sales or revenues per product (denoted by n) during all the period considered in log-log scale. The distribution of n follows a power law $P(n) \sim n^{-2.6}$ as shown by the black solid line.

Description	Value
Average sales value per store per week	564.14
Best week for retailers	92
Largest quantity of products offered in week 50	736
Store that offered the largest quantity of products in week 50	367
Best selling product	6534178 (GASOLINE-REG UNLEADED)
Best customer in a day	Customer 1609 in day 339
Money spent by customer 1609 in day 339	1008.08
Best customer in a week	Customer 1609 in week 49
Money spent by customer 1609 in week 49	1312.77
Age of customer 1609	$45-54$
Marital status of customer 1609	Married
Number of kids of customer 1609	> 3
Best customer in all the period	Customer 1023
Money spent by customer 1023[a] during all period	$37,416.72$

[a] The demographic details of this customer are ignored

TABLE I: Aggregations filtered out from the three databases.

value per week as shown in I. Even if average values do not give us precise information about the distribution of the revenue by store, we get some insight about the order of magnitudes in the variables of interest.

Table I also summarizes other information such as the week where the total revenue by all the stores was highest, the preferred products, or even the customer that spent more money in a week, its age and its marital status.

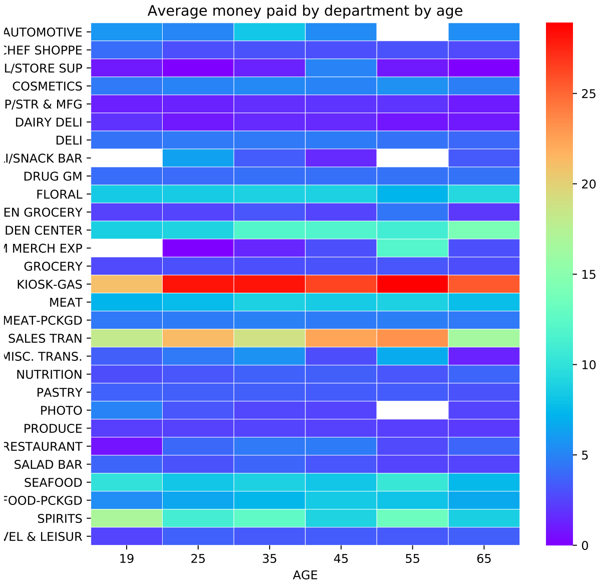

FIG. 5: Average money paid by customers belonging to different age groups in different departments. Gasoline that belongs to the department KIOSK-GAS is the best selling product overall in all the groups with ages larger than 25.

E. Heatmap of money spend in departments by customers grouped by age

Another interesting tool to visualize the data is to compare different features. For instance, we might be interested in knowing the departments that customers of a certain age prefer. Figure 5 shows the average money spent in each department by customers grouped by age. On average, customers spend more in gasoline belonging to the department KIOSK-GAS, this was found before when we looked for the most sold product. In second place, the department MISC SALES TRAN sells more on average, such department includes the following products: GASOLINE-REG UNLEADED, MISC SALES TRANS, MISCELLANEOUS H and B AIDS, FLORAL DEPT KEY RING, DEA SCHEDULE C II, TICKETS, ELECTRONIC GIFT CARDS ACTIVATI, PRODUCE DEPT KEY RING, MEAT SUPPLIES, ELECTRONIC GIFT CARDS REFRESH, OUTSIDE VENDORS GIFT CARDS, AMERICAN EXPRESS GIFT CARD, MASTERCARD GIFT CARD. Regarding the department AUTOMOTIVE, people between 35 and 45 years old spend more money on average. In the department FLORAL, customers between 55 and 65 years old spend less overall but spend more in FROZEN GROCERY overall. The youngest group between 19 and 24 spend less money overall in RESTAURANT but more in PHOTO and SPIRITS than older customers.

Actual/predicted	1	2	3	4	5
1	0	0	0	10	4
2	0	0	0	7	3
3	0	0	0	1	0
4	2	0	0	26	7
5	1	0	0	27	11

TABLE II: Confusion matrix from a random forest classifier by household composition using the money spent by department as input

1. *Multiclass classifiers*

Since there are customers whose transactions are known, thus the products they bought, but whose demographic information is ignored, we can build a multiclass classifier as an attempt to classify them in a group of a certain household composition or of a certain age, for instance. To build a model to classify people by their household composition we start with simple models such as random forests and K-nearest neighbors. As an example, we show the results of a random forest classifier applied to the classification of the customers by their householder composition (1:'single female', 2:'single male', 3:'Single parent', 4:'Parents without kids', 5:'Parents with kids') from the average money spent by department. For this random forest we obtain the confusion matrix in Table II

Since the groups 1, 2, and 3 are under-represented and fewer examples of these groups are shown to the classifier during the training process, it is more difficult for the model to classify correctly those cases. On the contrary, for groups 4 and 5 where there are more examples to train the model, the classification performs better. Hence, for group 4 the precision and the recall are 0.37 and 0.74, respectively. Similarly, for group 5 the precision and the recall are 0.44 and 0.28. The accuracy for this model is 0.37. A K-Nearest neighbors (KNN) classifier performs slightly better that the random forest with an accuracy of 0.42 even if none of the customers belonging to groups 1, 2 and 3 were correctly identified.

Now let us build a KNN classifier to identify customers by their ages. The supervised learning consists as before, in showing the average money spent in each department as input, and the target is the age group the customer belongs to as the output. In this case the number in the target corresponds to (1:19-24 2:25-34 3:35-44 4:45-54 5:55-64 6:>65). For a model with 30 nearest neighbors the accuracy of the model is 0.35.The confusion matrix for this model is shown in Table III.

In this case, the groups that are better respresented are 2, 3 and 4. The precision measured for these groups is 0.5, 0.29 and 0.34, respectively. And the recall values is 0.13, 0.08 and 0.89, respectively.

Finally, I build a deep neural network (DNN) using keras to classify the customers by age. This model can be used to classify the customers whose demographic information is ignored. The architecture of the DNN has one hidden layers with 12 units, I use dropout regularization in the hidden layer to avoid overfitting. The visible layer in the input has 43 units corresponding to the 43 departments considered as input. The output layer has 6 units as there are 6 age groups as outputs. The accuracy of this DNN has an accuracy of 0.424. Fig. 6 shows the evolution of the loss and the accuracy through 200 epochs.

Actual/predicted	1	2	3	4	5	6
1	0	0	0	7	0	0
2	0	3	1	19	0	0
3	0	1	2	21	0	0
4	0	1	3	31	0	0
5	0	0	0	3	0	0
6	0	1	1	9	0	0

TABLE III: Confusion matrix from a K-nearest neighbors classifier by age using the money spent by department as input and the age group as given by the following dictionary (1:19-24 2:25-34 3:35-44 4:45-54 5:55-64 6:>65).

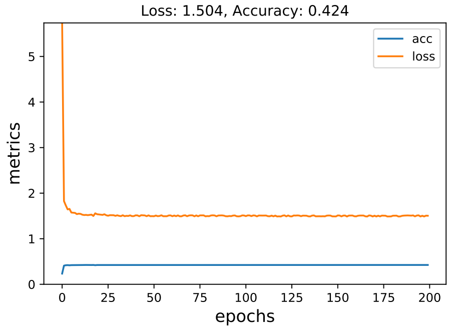

FIG. 6: Loss and accuracy of the deep neural network with 12 units in the hidden layer as described in the text. This DNN is used as a multiclass classifier of the customers by age according to the average money spent by department.

In order to improve the performance of the DNN we might try different architectures with more hidden layers. The risk of adding more layers is the overfitting which might learn the data used in the training process but being unable to classify correctly new data.

III. TIME SERIES ANALYSIS

A. Analysis of sales per week

If we look at the average sales value per week we obtain a time series that can be described by an ARIMA model as a first attempt to predict sales for future weeks. First, we begin by checking stationarity. To do so we compute the rolling mean and the standard deviation of the time series. Then, we can remove the moving average of the time series as shown in Fig. 7.

Then, by computing the autocorrelation function and the partial autocorrelation function we can infer the pa-

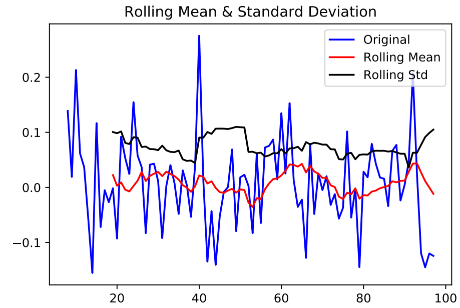

FIG. 7: Moving average and standard deviation of the original time series.

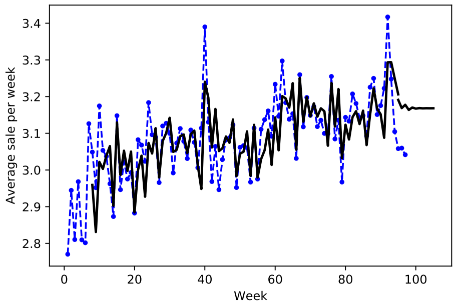

FIG. 8: AR model (black solid line) adjusted to the average sales value per week. The parameters of the ARIMA process are $p = 1$, $d = 1$ and $q = 0$.

rameters (p, d, q) for the ARIMA process. In this case, a simple AR model seems to work well where the parameters of the model are $p = 1$, $d = 1$ and $q = 0$. By going back to the original values we can try to predict some future sales values for week ≥ 98 as shown in Fig. 8.

B. Time series of money paid by customers per week

Until now we have looked at quantities averaged in time. Now, we are going to see the money spent by week for customers of different ages. One exercise would be to find outliers within a group so that we identify customers with anormal buying patterns using Isolation Forests, for instance.

Also, we can build a recurrent neural network with long-short-term memory (LSTM) cells to analyse the evolution of the money spent by customers of a certain age daily or weekly[3]. In particular, the DeepAR model as proposed in[4] seems a good candidate to train a network with all the time series of a given age group if we assume they are related. The DeepAR model in[4] seems to be a good model to describe the time series for the following reasons. Since the magnitudes of the time series differ widely and the behaviour is intermittent and bursty (see Fig. 9), it seems that a likelihood function such as the negative binomial distribution might be appropriate. Regarding the input values of the time series, one might divide the autoregressive inputs by a scale factor. If instead of looking at the money spent per week we look at the daily money spent by customer we can use as a covariate the week.

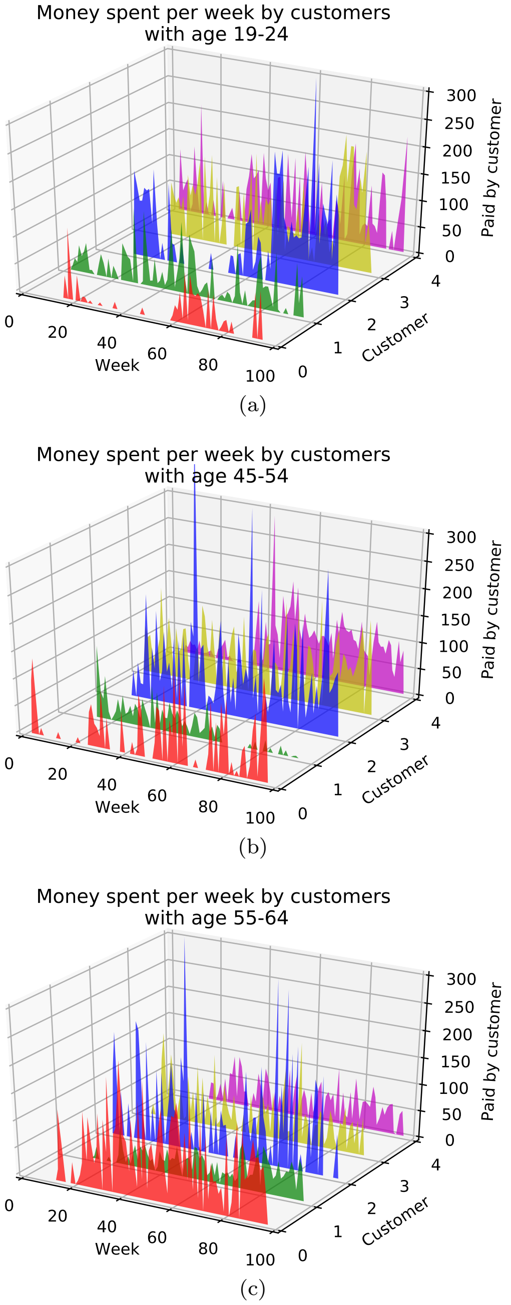

FIG. 9: (a) Total money spent by 5 customers between 19 and 24 years old per week. (b) By 5 customers between 45 and 54 years old per week, and (c) by 5 customers between 55 and 64 years old per week.

[1] W. Li, IEEE Transactions on information theory **38**, 1842 (1992).

[2] X. Gabaix, The Quarterly journal of economics **114**, 739 (1999).

[3] R. Carbonneau, K. Laframboise, and R. Vahidov, European Journal of Operational Research **184**, 1140 (2008).

[4] V. Flunkert, D. Salinas, and J. Gasthaus, arXiv preprint arXiv:1704.04110 (2017).