Second Best Offer in Financial Products

Pablo Eduardo Duarte Tzuc
Data Engineering
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: st1809063@upy.edu.mx

D.S Gabriela Flores Bracamontes
Pétalo No.47, Col. El Reloj, Del.
Coyoacán, CDMX C.P. 04640
Email: gflores@richit.ai

Dr. Juan Vázquez Montejo
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: juan.vazquez@upy.edu.mx

Abstract

This document describes the work I did during my professional stay corresponding to the second training cycle, that is, the seventh four-month period at the Universidad Politécnica de Yucatán, with a total of 120 hours.

During this period I worked on the project ”Financial Analytical Models”, among the models to work on I was in charge of the ”Second best offer of financial products” in the company RICH IT. The project was in the research phase so I was in charge of developing a solution to the problem, which was a combination of a market basket analysis and four different classification models, which can be seen in this document.During the solution of the problem I performed the data science life cycle, i.e. business understanding, data collection, data preparation, exploratory data analysis, modeling, model evaluation. Except model deployment for reasons explained in this document.

Index Terms

Market basket analysis, classification models, data collection, data preparation , exploratory data, analysis, second best offer.

Second Best Offer in Financial Products

I. Introduction

This document shows an alternative to recommend the second best product to different customers of a bank, but this does not mean that it can be applied to other fields such as e-commerce, auto service stores or in general to any company engaged in the sale of different products. Nowadays, making recommendations that go according to customers is important for both companies and the customer, since this means higher income for a company and better purchase satisfaction for a customer [1].

The document first shows a theoretical framework of the concepts used during the project, then it will show the different phases of the solution to the problem starting with the understanding of the problem, selection of the dataset to be used, a light EDA where the most interesting and useful graphs for the preparation of the dataset will be shown, preparation of the almost final dataset, solution to the problem and its application, results of the classification models and finally conclusions and future work.

II. Objectives

Since the project was in the initial research phase. The main objective was to develop an alternative to recommend the second best offer to customers, this was achieved, however due to lack of time I had to make some adjustments to the dataset to see if they improved the classification models described in the paper.

III. State of the Art

In the paper entitled "An Empirical Evaluation of Intelligent Machine Learning Algorithms under Big Data Processing Systems" the H20 and Sparkling Water platforms were used to handle big data.The paper does not mention a robust preprocessing, it only mentions how some errors and missing values were treated, the paper mentioned that 60% of the data was used for training and 40% for testing for the realization of 48 different models (one for each product), i.e. 24 in H20 and 24 in Sparkling Water [2].

The models have the same configurations that were experimentally set as follows: 10-folds cross-validation using the validation dataset, 5 hidden layers with 10 neurons each, Shuffle training data and the rest were put in default configuration.

The metrics used to evaluate the models are: Accuracy, Area Under the Curve (AUC), F1-score, Precision, Recall, Specificity, and training Time. the time and the AUC measures were omitted.

Experimental results showed that both platforms have achieved converging results in terms of accuracy, f1-score, precision, recall, and specificity, with the Sparkling Water platform subtly outperforming the H2O platform in terms of model accuracy. However, the H2O platform achieved a significant result in terms of model training time [2].

IV. Methods and Tools

A. Market basket analysis

Market basket analysis is a data mining technique used to better understand customer buying patterns with the help of association rules. Although its name is derived from supermarket or refers to supermarkets, this does not mean that it can be applied only to these sites, in general it can be applied wherever there is any collection of items to identify affinities that can be exploited in some way [3].

B. Buying patterns

Buying patterns are those that allow us to know the How? and Why? of the purchase decisions of a product [4]. This is achieved with the association rules that generates the Market Basket analysis.

C. Association rules

The Association rules allow us to find relationships within a set of transactions, i.e. products that tend to occur together. It is given by x $\rightarrow$ y, where "x" is the antecedent and "y " is the consequent [5]. In the market basket analysis the three main rules of association are support, confidence and lift.

D. Support

The support indicates the number of times the basket appears, from one to the maximum number of possible products in a transaction, in our transaction set [6]. The formula is the next:

$$support = \frac{freq(x, y)}{N}$$

where:

- freq: is the frequency of the products x and y in a basket.
- N: total of transactions.

E. Confidence

The confidence is the probability P(X—Y), i.e., the probability that a transaction contains X, given that it already has Y [6]. The formula is the next:

$$Confidence = \frac{freq(x, y)}{freq(x)}$$

F. Lift

The lift is the probability that all the elements of a rule (A→B) occur together, the further the value of lift is from 1, it means more evidence that the rule is not due to a random artifact, i.e., the more evidence that the rule represents a real pattern [6]. The formula is the next:

$$Lift = \frac{support}{support(x) * support(y)}$$

G. Classification models

Classification models describe and distinguish classes and concepts of data, it is responsible for identifying to which of a set of categories (sub-populations) a new observation belongs [7], for example, suppose we want to buy a used vehicle, but we do not know when a used vehicle is good and when it is bad, so with the help of a database containing characteristics, opinions about used vehicles and whether it is a good vehicle or not, we can predict with a classification model whether it is a vehicle that will be a good buy or not.

H. Confusion matrix

The confusion matrix and associated metrics are a fundamental part of a data scientist's toolbox, helping to determine whether a classification model is appropriate to the problem [8].

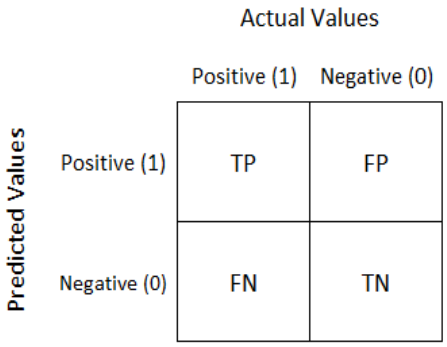

Fig. 1. Confusion matrix for binary classifier.

Figure 1 shows a confusion matrix for a binary classification, for example if we are predicting whether a mushroom is poisonous or not, the diagonal containing TP (true positive) and TN (true negative) are the number of correct predictions, i.e. TP (true positive) would be the number of mushrooms that are poisonous and TN the number of mushrooms that are not poisonous. As for the diagonal containing FP (false positive) and FN (false negative), the FP would be the mushrooms that the model classified as non-poisonous but are poisonous and FN would be the mushrooms that the model classified as poisonous but are not. Other metrics such as accuracy, precision and recall f1-score are derived from the confusion matrix [8][9].

I. Accuracy

The accuracy is the percentage of correct predictions for the test data [10][11] and its formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

J. Precision

The precision is the fraction of relevant examples (TP) among all those predicted to belong to a given class. answer the question, what percentage of positive predictions were correct? [11][12] and its formula is the next:

$$Precision = \frac{TP}{TP + FP}$$

K. Recall

Recall is the percentage of cases that were predicted to belong to a class relative to all the cases that actually belong to the class and answers the question, what percentage of positive cases were captured? [11][12], its formula is the next

$$Recall = \frac{TP}{TP + FN}$$

L. F1-score

F1-score is used when we want to find a balance between precision and recall metrics, F1-score could be a better measure if there is an uneven class distribution (large amount of real negatives) [11][12]. its formula is the following.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

V. Development

The dataset selected to address the problem can be found at https://www.kaggle.com/c/santander-product-recommendation, it was provided by Santander bank in order to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers. If you want to know about every variables contained in the dataset you can go to the link above.

it contains 13,647,309 instances and in total 48 variables, therefore given the amount of data in the dataset, we worked with a sample of 6,823,654 records (almost half of the original dataset)

fecha_dato	0
ncodpers	0
ind_empleado	27734
pais_residencia	27734
sexo	27765
age	0
fecha_alta	27734
ind_nuevo	27734
antiguedad	0
indrel	27734
ult_fec_cli_1t	6812876
indrel_1mes	104797
tiprel_1mes	104797
indresi	27734
indext	27734
conyuemp	6822769
canal_entrada	117688
indfall	27734
tipodom	27734
cod_prov	62732
nomprov	62732
ind_actividad_cliente	27734
renta	1225863
segmento	118604

Fig. 2. Missing values 1

A. Missing values

In this part I only show the initial values of the sample we selected, their treatment will be shown later, since some preprocessing of the data was done with the help of exploratory data analysis (EDA).

The dataset had missing values in the customer characteristics as shown in Figure 2.We can see that the number 27,734 is repeated in several variables, so this may mean that there are a number of clients that have the same missing values where the same number is repeated (same variables), this will be corroborated in the exploratory and data analysis part.

ind_ahor_fin_ult1	0
ind_aval_fin_ult1	0
ind_cco_fin_ult1	0
ind_cder_fin_ult1	0
ind_cno_fin_ult1	0
ind_ctju_fin_ult1	0
ind_ctma_fin_ult1	0
ind_ctop_fin_ult1	0
ind_ctpp_fin_ult1	0
ind_deco_fin_ult1	0
ind_deme_fin_ult1	0
ind_dela_fin_ult1	0
ind_ecue_fin_ult1	0
ind_fond_fin_ult1	0
ind_hip_fin_ult1	0
ind_plan_fin_ult1	0
ind_pres_fin_ult1	0
ind_reca_fin_ult1	0
ind_tjcr_fin_ult1	0
ind_valo_fin_ult1	0
ind_viv_fin_ult1	0
ind_nomina_ult1	16063
ind_nom_pens_ult1	16063
ind_recibo_ult1	0

Fig. 3. Missing values 2

In figure 3 we can see the corresponding missing values in the 24 different products, there are only 16063 missing values in ind_nomina_ult1 (payroll) and ind_nom_pens_ult1 (pensions).

B. Exploratory data analysis

In this part of the work I will not show the complete EDA since there are many graphs and some of them are not so relevant, so we will show a light EDA which will contain the most interesting graphs, as well as those that helped us to see errors in the dataset and correct them.

In the graph below we can see a slight white line from ind_employee (employee index) to ind_activity_customer (activity index), indicating that some customers have missing values from the ind_employee to ind_activity_customer column. This tells us that the 27,734 missing values in the different columns are from the same customers.

Fig. 4. Matrix of missing values

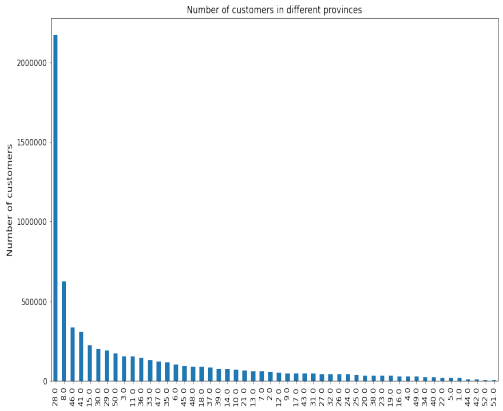

Fig. 5. Code of province

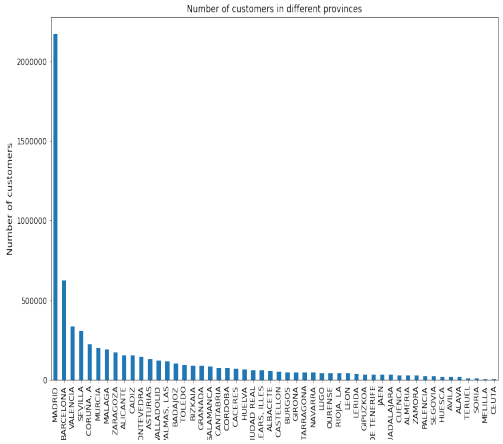

Fig. 6. Name of province

We can see in figure 5 and 6 that they are almost identical,

the only difference are the labels of the provinces, because in figure 7 this is with numbers, on the other hand in figure 8 the provinces are with names, given the above I decided to eliminate the variable cod_prov and use nomprov.

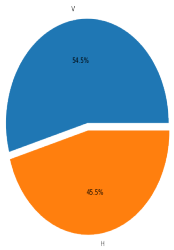

Fig. 7. Distribution of the variable sex

We can see in Figure 7 that there is a majority of women in the dataset with 54.5% and 45.5% of men, however there is not a very big difference.

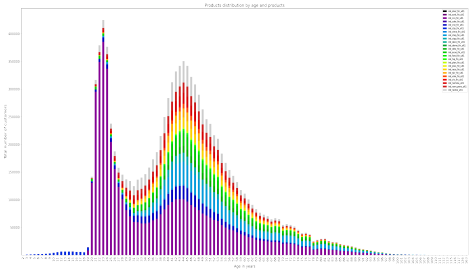

Fig. 8. Products distribution by age and products

We can see in Fig. 8 a bimodal distribution where the customers between the age range of 20 to 28 years, the product with the highest demand is ind_cco_fin_ult1 (current account), while customers between the age range of 38 to 52 years, have the products: ind_cco_fin_ult1 (current account), ind_ctop_fin_ult1 (private account) and ind_recibo_ult1 (direct debit) with the highest demand.

C. Missing value processing

This section shows the remaining missing values, since some of them were processed thanks to some graphs shown in the previous section, as well as the corresponding treatment.

Eliminated variables:

- **Conyuemp:** this variable was eliminated because it contains mostly missing values.
- **Cod_prov:** this variable was eliminated, because as can be seen in the EDA it is similar to the nom_prov variable.

Variables filled with zero:

- **ind_nomina_ult1**
- **ind_nom_pens_ult1**

These variables were filled with zero, since the dataset contains 0 when a customer does not have the product and 1 when he has the product.

Variables filled randomly:

- **sexo**
- **Indrel_1mes**
- **Tiprel_1mes**
- **Canal_entrada**
- **Nomprov**

The previous variables were filled in this way as they did not have a clear category that was dominant, as could be seen in the graph corresponding to the sex distribution, initially they were going to be filled with the KNN algorithm, but it was discarded since the remaining time of the stay was not much.

Variables filled with the dominant category:

- **Segmento**

The previous variable was filled by ”02 - PARTICULARES” for being notoriously dominant category.

Transformed variable:

- **ult_fec_cli_1t**

Data containing date were replaced with ”yes” and data with no date (missing values) were filled in with ”no”. It was filled in this way because it could be converted into a dummy variable later and used in the classification models.

D. Near final dataset preparation

Once the dataset was free of missing values, we proceeded to convert the categorical variables to dummy variables, so that they could be used in the classification models. The variables converted into dummy variables are the following:

Once the dataset was free of missing values, we proceeded to convert the categorical variables to dummy variables, so that they could be used in the classification models.

The variables converted into dummy variables are the following:

- **ind_empleado**
- **pais_residencia**
- **sexo**
- **indrel**
- **ult_fec_cli_1t**
- **indrel_1mes**
- **tiprel_mes**
- **indresi**
- **indext**
- **canal_entrada**
- **nomprov**
- **segmento**

The dataset with the dummies variable was left with a total of 385 columns, 6,795,919 records and zero missing values.

E. Solution of the problem

The solution to give the second best offer to customers was a combination of market basket analysis and classification models , i.e., first I had to perform a market basket analysis to find out which products tend to be purchased together and that it is not a coincidence.

In order to perform the market basket analysis, from the dataset that was already completely free of missing values and had all 385 columns, only the 24 different products and customer IDs were selected.

A minimum support of .03 or 3% was selected, a higher support was not chosen because it returned few transactions and a lower support was not chosen because it took too much time and the lower the support means fewer times the product appears among the records.

The minimum support of .03 or 3% means that baskets will be selected from one product up to a maximum possible number of products that appear at least 203,877 times out of our total records.

I obtain seventeen baskets with different association rules. To choose the products we relied on the lift and support metric, the lift was selected since it is not bidirectional like confidence, i.e., the lift of A $\rightarrow$ B is the same as the lift of B $\rightarrow$ A, contrary to confidence.

We choose the next products:

- **ind_cno_fin_ult1 (payroll account)**
- **ind_nom_pens_ult1 (pension)**

The lift of the selected products is greater than one which means that the presence of the antecedent increases the chances of the consequent occurring in a given transaction. The support that the selected products have is not so low compared to others, which indicates that the consequent is not a coincidence.

F. Classification models

I used with the models GridSearchCV in order to obtain the most optimal hyperparameters.The classification models used with GridSearchCV were:

- **Logistic regression**
- **Decision tree**
- **Random Forest**
- **Adaboost**

VI. Results

First product					
		Precision	recall	f1-score	support
LogisticRegression	0	0.91	1	0.96	1865049
	1	0	0	0	173727
DecisionTreeClassifier	0	0.92	0.99	0.96	1864907
	1	0.62	0.1	0.18	173869
RandomForestClassifier	0	0.91	1	0.96	1864907
	1	0.93	0	0	173869
AdaBoostClassifier	0	0.91	1	0.96	1865314
	1	0	0	0	173462

Fig. 9. First product's results

As can be seen in Fig. 9 and 10 the best model is the decision tree, however it is not quite optimal since the F1-score is too low for class 1 (recommend the model), while almost all models are good at predicting class 0 (do not recommend the product), so this may indicate that our dataset is unbalanced.

Second product					
		precision	recall	f1-score	support
LogisticRegression	0	0.94	1	0.97	1908900
	1	0	0	0	129876
DecisionTreeClassifier	0	0.94	1	0.97	1908900
	1	0.64	0.07	0.12	129876
RandomForestClassifier	0	0.94	1	0.97	1908886
	1	0.96	0	0	129890
AdaBoostClassifier	0	0.94	1	0.97	1908886
	1	0	0	0	129890

Fig. 10. Second product's results

VII. Conclusion

From my point of view the way the solution to the problem was approached is one of the best things I took away, because before defining a solution to the problem, I was investigating alternatives to solve it, one option was to use matrix factoring (used by Netflix to recommend), I am sure it would have been a good alternative, but due to the number of hours I was assigned to the stay, it was going to be complicated. On the other hand some complications were the amount of data and the characteristics of my computer, but thanks to RICH IT this could be solved since they lent me a company server, which was useful to run everything related to the models, since it was what consumed more resources.

As you could see the results obtained were not all good and gives us indications that the dataset is unbalanced, this can be solved with some technique of oversampling or under-sampling, unfortunately running the models took a long time and when the results were obtained I was already in the final part of my stay, so in future work would be to balance the classes and run again all the models with the help of GridSearchCV to obtain the best hyper-parameters and finally see if the models improved and which is the best after class balancing.

Acknowledgment

I would like to thank my advisor Paulina Gomez came up with the idea of combining market basket analysis with classification models. Also to my partner Danilo Cuevas since he helped me to configure and use the server provided by the company and to Gabriela Flores since she supported us in whatever we needed and in general to all the people who helped me with the development of the project and to RICH IT for the opportunity.

References

[1] A. Klimenko, greenice, 30 05 2019. [En línea]. Available: https://greenice.net/ml-based-recommendation-system-for-marketplace-5-proven-ways-to-grow-your-profits/.

[2] SULEIMAN, Dima; AL-ZEWAIRI, Malek; NAYMAT, Ghazi. An empirical evaluation of intelligent machine learning algorithms under big data processing systems. Procedia computer science, 2017, vol. 113, p. 539-544.

[3] AMAT RODRIGO, Joaquín. Reglas de asociación y algoritmo apriori con R. cienciadedatos. net, 2018.

[4] M. S. Team, mbaskool, 22 January 2018. [En línea]. Available: https://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/10916-buying-pattern.html.

[5] SUBASI, Abdulhamit. Practical Machine Learning for Data Analysis Using Python. Academic Press, 2020.

[6] Robert Nisbet Ph.D., ... Ken Yale D.D.S., J.D., in Handbook of Statistical Analysis and Data Mining Applications (Second Edition), 2018
[7] PARRA, Francisco. Estadística y Machine Learning con R. RPubs Blog, 2017.
[8] P. R. d. l. Santos, Telefonica, 23 enero 2018. [En línea]. Available: https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/.
[9] K. Markham, Dataschool, 25 March 2014. [En línea]. Available: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.
[10] J. Jordan, Jeremy Jordan, 21 July 2017. [En línea]. Available: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/.
[11] G. G. Meza, Medium, 28 January 2018. [En línea]. Available: https://medium.com/@gogasca_/precisi%C3%B3n-y-recuperaci%C3%B3n-precision-recall-dc3c92178d5b.
[12] K. P. Shung, Towards data science, 15 March 2018. [En línea]. Available:https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.