Santander Product Recommendation

Pablo León Alcaide
Andrés Montoro Montarroso
Jaime Cepeda Villamayor
Pablo Ráez García Retamero
Mª Dolores Sesmero Pozo

Introduction

The problem that is tried to solve is the one proposed in a competition of Kaggle by Santander. It consists, given the behavior of customers over fifteen months, to predict what new product will acquire in the sixteenth month.

Different approaches have been proposed to solve this problem. In most of them, we agree on the use of XGBoost as an algorithm to create the predictive model. It has also been pointed out that most of the useful features that allow the best predictions to be reached have to do more with products and less with customer demographics. Perhaps the explanation is that the products that a customer has at a given moment summarize their demographic characteristics

To solve the problem, it is hypothesized that it will be possible to predict which product a customer will hire based on the products he owns, which he had and was discharged, and some demographic characteristic that conditions the client's behavior. These demographic characteristics should be selected after performing the exploratory analysis of the data and not using knowledge of the problem domain due to the synthetic nature of the data.

The first part of this document will explain the characteristics of the starting data set. Also the transformations that have been made on these data to obtain a set of useful data to train the prediction model. The second part will discuss the algorithms used in both the exploratory analysis and the creation of the prediction model. In the third, we will explain the process carried out in the exploratory analysis and the conclusions obtained. In the last part we will comment on the prediction model created.

Data

The data that gave us the problem of Kaggle is a database that shows the behavior of a series of customers of the Santander bank for fifteen months to be able to predict what kind of products will be acquired in the sixteenth month. The size of the database provided is 13.647.310 rows with the following fields and their meanings:

- Column Name: Description
- fecha_dato: The table is partitioned for this column
- ncodpers: Customer code
- ind_empleado: Employee index: A active, B ex employed, F filial, N not employee, P pasive.
- pais_residencia: Customer's Country residence
- sexo: Customer's sex
- age: Age
- fecha_alta: The date in which the customer became as the first holder of a contract in the bank.
- ind_nuevo: New customer Index. 1 if the customer registered in the last 6 months.
- antiguedad: Customer seniority (in months)
- indrel: 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month).
- ult_fec_cli_1t: Last date as primary customer (if he isn't at the end of the month)
- indrel_1mes: Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
- tiprel_1mes: Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
- indresi: Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
- indext: Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
- conyuemp: Spouse index. 1 if the customer is spouse of an employee
- canal_entrada: channel used by the customer to join
- indfall: Deceased index. N/S
- tipodom: Addres type. 1, primary address
- cod_prov: Province code (customer's address)
- nomprov: Province name
- ind_actividad_cliente: Activity index (1, active customer; 0, inactive customer)
- renta: Gross income of the household
- segmento: segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
- ind_ahor_fin_ult1: Saving Account
- ind_aval_fin_ult1: Guarantees
- ind_cco_fin_ult1: Current Accounts
- ind_cder_fin_ult1: Derivada Account
- ind_cno_fin_ult1: Payroll Account

- ind_ctju_fin_ult1: Junior Account
- ind_ctma_fin_ult1: Más particular Account
- ind_ctop_fin_ult1: Particular Account
- ind_ctpp_fin_ult1: Particular Plus Account
- ind_deco_fin_ult1: Short-term deposits
- ind_deme_fin_ult1: Medium-term deposits
- ind_dela_fin_ult1: Long-term deposits
- ind_ecue_fin_ult1: e-account
- ind_fond_fin_ult1: Funds
- ind_hip_fin_ult1: Mortgage
- ind_plan_fin_ult1: Pensions
- ind_pres_fin_ult1: Loans
- ind_reca_fin_ult1: Taxes
- ind_tjcr_fin_ult1: Credit Card
- ind_valo_fin_ult1: Securities
- ind_viv_fin_ult1: Home Account
- ind_nomina_ult1: Payroll
- ind_nom_pens_ult1: Pensions
- ind_recibo_ult1: Direct Debit

The training version of the data have many null fields, so we had to clean the data. To handle this large amount of data we have transformed the training data provided by the competition selecting only those rows in which products were contracted. In addition, we have added some features as if a customer was unsubscribed from a product in the past. Therefore, the final training set has 447904 rows, a much more manageable size than the initial one.

Algorithms

In this section we discuss the algorithms used for both the exploratory analysis phase and the construction of a prediction model.

Algorithms used in the exploratory analysis phase:

- K-means: This clustering algorithm has been used with three different objectives: Reduce the size of the data to be able to apply a hierarchical clustering on them, to measure the silhouette of the grouping performed and in cases where a characteristic is used Continue, make it discreet. In the first two cases, the square root of the number of elements is used as the number of clusters to be obtained. In the latter case, it depends on the characteristic itself.
- Hierarchical Clustering: It has been used to visualize dendrograms and to make decisions based on them. The method used to make the groups has been the complete one

Algorithms used in the prediction phase:

- Random Forest: This algorithm has been used to determine which products a new customer acquires in the following month. For their training, the data of the acquisitions of new products, the products that they had in the past and if the customer was used, have been used, labeled with the products that they will buy in the following month.

Exploratory Analysis

The objective of this phase is to identify demographic characteristics of the customers that are related to the products that they own. In order to carry out this analysis, the characteristics of the customers and the products they had in the first month were considered. In addition, due to the size of the data, randomly selected samples were collected from all clients.

The procedure carried out is as follows:

1. A demographic characteristic is selected.
2. The data set is divided by the values of that characteristic. If it is continuous, a K-means is applied to discretize the values into groups. In the case of rent, for example, before using K-means, it is scaled by provinces.
3. A K-means applies to products that have customers with a certain value for that characteristic. The silhouette is measured in order to obtain information on how "good" clustering has been.
4. A hierarchical clustering is performed on the groups obtained in the K-means in order to observe the similarity between the different groups obtained in the K-means.
5. Based on the silhouette and shape of the dendrogram it is determined whether the selected demographic feature is related to the products that the client has.

This procedure has been carried out with various characteristics such as the country of residence, the channel of entry, the income, the province or the age. Only in the latter case has it been possible to determine that it has influence on the products contracted by a customer.

For the characteristic of age, it is divided into two groups: minors and adults. Based on the value of the silhouette, which in this case is high, and the shape of the dendrogram, it is not determined to be related to the products. However, k-means groups virtually all elements in the same cluster for the underage persons, therefore it is concluded that underage clients may exhibit similar behavior.

Problem Solution

The training data is formed by the occurrences of the clients in the initial data in which the acquisition of some new product takes place. From each of these appearances only the products that the client had last month are preserved, and the age of the client indicating if the client is under or over age. Features have also been added indicating whether a customer has been discharged from a product in the past. Each of these elements (a customer's products, past customer's products and if minor) are labeled with the products they acquire the following month.

With this training data the Random Forest model is trained. First, in order to measure the accuracy of the model, the training data were split into two sets. With the former the model has been trained and the products acquired by the customers of the latter have been predicted. Precision has been measured in two different ways: If several products are predicted but at least one matches the prediction is considered correct, or that all products must match to be considered correct. With the first way of measuring, 82% of the new acquisitions have been achieved. With the second 75%.

Subsequently, the model has been trained with all available training set and the acquired products have been predicted from the training set supplied by Kaggle.

Source code

The code for the entire project can be found in the following repository: https://github.com/PabloLeon23/ML-Final-Work.

Conclusions and future work

Given the results obtained, it can be concluded that the products that a customer will acquire in the future is defined mainly by the products that already have or have had in the past. The demographic characteristics of a client have less influence and in many cases their consideration worsens the results of the prediction.

In order to improve this model, it is possible to expand the exploratory analysis carried out, studying the rest of demographic characteristics and their influence on the contracted products. You can also study adding new features related to products that can increase accuracy, such as the time a customer takes without hiring a product. It is also possible to study ways of dealing with the acquisition of many products at the same time, since they are rare occurrences, and the fact of trying to predict them all may worsen the general precision of the model.

References

https://www.kaggle.com/c/santander-product-recommendation

http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://www.kaggle.com/apryor6/santander-product-recommendation/detailed-cleaning-visualization-python