Technical exercise PrestaShop

Author: Gabriel DELGADO (gabriel.delgado@alumni.polytechnique.org)

The ensuing study is performed using Python as programming language and the libraries Sklearn, Pandas and Matplotlib. For more details consult the attached Jupyter notebook.

Dataset description

The dataset represents the account history until 2017-12-13 of 2089 merchants using (or having used) the PrestaShop READY solution. This solution comes today in two flavors or plan offers: Trial and Start. The Trial plan offer allows the merchant to use the PrestaShop READY solution for free during 30 days. After this period, the merchant must decide to either disable his account or to subscribe to the Start plan offer.

Within the dataset, each merchant is characterized by 43 features describing the contextual and usage characteristics of his e-shop. In order to facilitate the following analysis, we split these features into 3 groups. The features with the * symbol are only available for those merchants that have subscribed to the Start offer:

Categorical features: Merchant_ID, Merchant_Country, Merchant_Language, Merchant_Billing_Country*, Merchant_Billing_City*, Merchant_Billing_Zip*, Merchant_Plan_Offer, Merchant_Plan_Periodicity*, Shop_Status, Shop_Name, Shop_ID, Shop_Image_Version.

Timestamp features: Merchant_Creation_Timestamp, Merchant_Subscription_Timestamp*, Merchant_Days_To_Subscribe*, Merchant_Hours_To_Subscribe*, Merchant_End_Trial, Merchant_Creation_Year, Merchant_Creation_Month, Merchant_Creation_Day_Of_Week, Merchant_Creation_Time, Merchant_Creation_Time_Period, Shop_Last_Bo_Connection_Date, Shop_Last_Theme_Change_Date, Shop_Last_Shipping_Method_Config_Date, Shop_Last_Payment_Method_Config_Date, Shop_Last_Order_Date, Shop_Last_Logo_Change_Date, Shop_Last_Favicon_Change_Date.

Other numerical features: Shop_Existence_Days, Shop_Products_Count, Shop_Shipping_Methods_Count, Shop_Payment_Methods_Count, Shop_Bo_Connections, Shop_Bo_Connections_Last_7_Days, Shop_Bo_Connections_Last_30_Days, Shop_Orders_Count, Shop_Orders_Count_Last_7_Days, Shop_Orders_Count_Last_30_Days, Shop_Gross_Sales, Shop_Gross_Sales_Last_7_Days, Shop_Gross_Sales_Last_30_Days.

We remark that all the listed merchants in the dataset are related to only one e-shop[1].

Objective

PrestaShop would like to estimate de probability of a merchant using their solution of switching from a Trial plan offer to a Start plan offer. This information could be useful to define different merchant profiles (for instance hot, warm or cold) according to its chances to subscribe, and then apply adapted marketing actions to each profile.

The objective of this exercise will be thus to predict the churn probability of a merchant using the PrestaShop READY solution once reached the end of its trial period (i.e. 30 days after the creation of its account). We will say that a merchant churns if during his trial period (time frame for this study) he does not switch from a Trial plan offer to a Start plan offer. Indeed, below 30 days, the dataset shows that the churn ratio is zero[2]. Equivalently, this study can be formulated as a conversion rate

[1] Since the merchant ID and the Shop ID are independent, one merchant can manage different shops.
[2]Merchants wait the end of their trial period before disabling their account.

prediction problem such that the conversion occurs when the merchant buys the Start plan offer solution.

Data analysis and results

From the 2089 merchants, only for 573 have either subscribed to the Start plan offer or exceeded the trial period (30 days) with a disabled account status[3]. From this sample of 573 merchants, only 9% are not churning.

Since the data corresponds to a snapshot of the merchant situation at 2017-12-13, the following features are added to each merchant:

- Shop_Existence_Days_Corrected: Corresponds to the number of days of existence of the shop if the merchant subscribed to the Start offer or 30 days otherwise.
- Shop_Bo_Connections_Ratio = Shop_Bo_Connections[4] / Shop_Existence_Days_Corrected
- Shop_Orders_Count_Ratio = Shop_Orders_Count / Shop_Existence_Days_Corrected
- Shop_Gross_Sales_Ratio = Shop_Gross_Sales / Shop_Existence_Days_Corrected

According to their pertinence, the chosen variables to build the churn classifier are:

Predictors = ['Shop_Products_Count','Shop_Bo_Connections_Ratio','Shop_Orders_Count_Ratio', 'Shop_Gross_Sales_Ratio','Shop_Shipping_Methods_Count','Merchant_Country','Merchant_Language']

Target = ['Shop_Status']

We drop all the lines in the dataset with NaN (or rather missing) values for the aforementioned columns. The final dataset possesses therefore 572 samples of which 51 cases belonging to the not churning class.

Given the above data distribution within each class, we can say that this is an imbalanced classification problem (9% of the data belongs to the interest class). Solutions to deal with this kind of problems are for instance to up-sample minority class, down-sample the majority class, use penalize algorithms or apply tree based algorithms. In our case we will apply the latter two solutions. Furthermore, since our attention focuses on the prediction power of the classifier for the minority class, we will seek a classifier with a high recall value as performance measure without penalizing too severely the model precision.

Next, we split the data into training and test datasets (80%/20%) and we apply different classification models in order to compare their performances. We remark however that since the size of the dataset is relatively small, we can expect non-negligible variations of performance among the classification models while choosing different random seeds for the dataset (training/test) partition.

Four classification algorithms are applied to the dataset. Their performances are measured as the area under the ROC curve (AUC ROC) and resumed in the following table. The AUC ROC represents the likelihood of the model to distinguish observations from 2 classes.

Model	AUC ROC training dataset	AUC ROC test dataset
Support vector machine (SVC)	0.91	0.75
Logistic regression	0.9	0.71
Decision tree[5]	0.98	0.95
Random forest	0.99	0.8

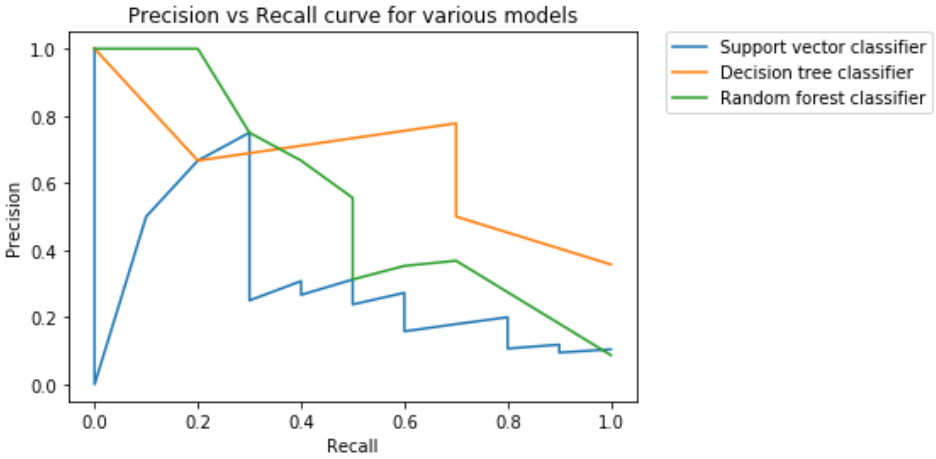
[3]We consider as outliers the few merchants who have exceeded the trial period with a Trial plan offer but remain with active accounts.
[4]Bo stands for “back-office”.
[5]We tune this model with maximum depth equal to 4

The difference of the AUC ROC between the training and the test datasets for each model is small so we conclude an overall absence of overfitting.

Furthermore, the Random forest model provides us a Feature importance score, from which we deduce that the most important features for churn classification are the 'Shop_Bo_Connections_Ratio' and the 'Shop_Products_Count'. This seems logical since both variables capture the merchant's usage and involvement in his e-shop account.

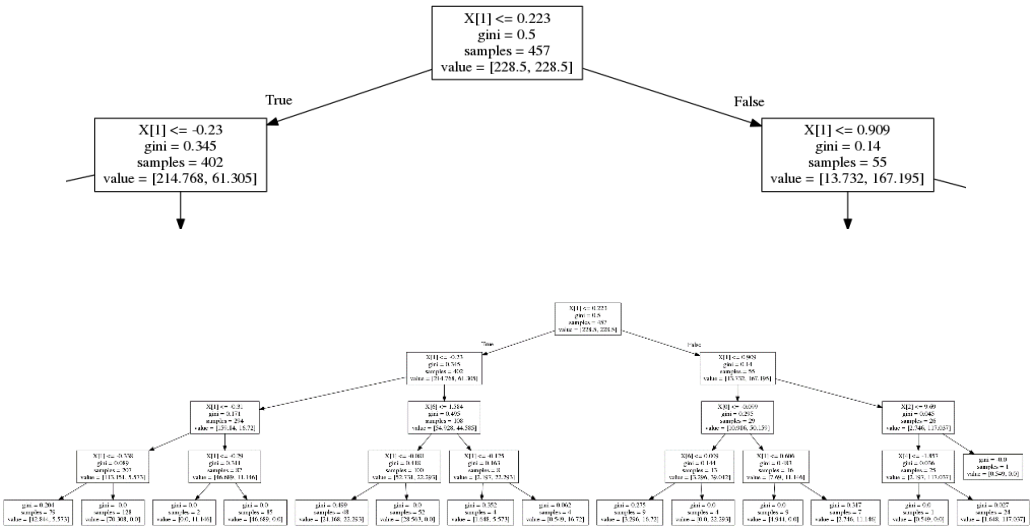
Features	Importance
Shop_Products_Count	0.2
Shop_Bo_Connections_Ratio	0.7
Shop_Orders_Count_Ratio	0.02
Shop_Gross_Sales_Ratio	0.02
Shop_Shipping_Methods_Count	0.02
Merchant_Country	0.03
Merchant_Language	0.03

Selecting the three best performance models according to the AUC ROC values for the test dataset (namely the SVC, DTC and RFC models), we plot and compare their respective Precision vs Recall curves:



The tuned decision tree classifier outperforms the other two models[6]. In particular for a given precision (let us say 0.5), the decision tree classifier maximizes the recall (0.7). The following picture represents the generated tree and a zoom on its root node (X[1] = 'Shop_Bo_Connections_Ratio').

[6] The random forest model could be tuned in order to get even better results. For a matter of limited time, we keep the default parameter values.



Next steps

Here below we list some recommendations and clues to pursue the data analysis:

- **Reframe the problem as anomaly detection**: Specific algorithms exist in order to detect outliers and rare events.
- **Use somehow the timestamp data:** In order to increase the performance of the classifier with respect to the average ratio used as predictor in this study (e.g. 'Shop_Bo_Connections_Ratio'), we could use the information related to the solution usage within the last 15 days or 30 days. The last date that some action occurred regarding the merchant creation date could also add valuable information.
- **Wait to gather more data:** According to the dataset, around 70% of the current merchants using the PrestaShop READY solution are in trial period. More information to enhance the algorithm training will be soon available.