VEF FINAL PROJECT MARKETING CAMPAIGN

DA 2022.04

Prepared by

[TONG KHANH LINH]
[linhtong1201@gmail.com]

Objective

Predict the customer who will accept an offer for a product or service based on the customer's personal information, purchasing history and previous campaign response.

Approaches

The project follows 5 main steps:

1. Understand the context, the dataset: the meaning of each variable.
2. Define the project objective.
3. Perform Exploratory Data Analysis (EDA) to get a glimpse of the dataset's problems and solve them with feature engineering.
4. Choose suitable algorithms to perform machine learning.
5. Interpret the results and redo previous steps if needed.

Dataset Overview

1. Data source

The dataset is from SAS Institute collected by Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner method. It was posted on Kaggle on March 2022.

2. Context

Data of customers' personal information, purchases history and previous campaign response (accepted/not accepted) of a seemingly food and beverages retail company.

3. Variables

The variables can be divided into two groups: demographic (personal information) and company-related (purchasing history, campaigns, complains)

3.1. Demographic Variables

Column	Data Type	Description
Year_Birth	int	Customers year of birth
Education	char	Customers level of education
Marital_Status	char	Customers marital status
Income	float	Customers yearly household income
KidHome, TeenHome	int	Number of small children and teenagers in customers household

3.2. Company-related Variables

Column	Data Type	Description
ID	int	Customers id
Dt_Customer	char	Date of customers' enrolment with the company
Recency	int	Number of days since the last purchase
Complain	int	1 if customer complained in the last 2 years

Z_CostContact	int	Cost to contact a customer
Z_Revenue	int	Revenue after client accepting campaign
Response	int	1 if customer accepted the offer in the last campaign, 0 otherwise
MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds	int	Amount spent on Wines, Fruits, Meat, Fish, Sweet, Gold products in the last 2 years respectively
NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases	int	Number of purchases made with discount, through company's web site, using catalogue, directly in stores
NumWebVisitsMonth	int	Number of visits to company's web site in the last month
AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5	int	1 if customer accepted the offer in the 1st, 2nd, 3rd,4th, 5th campaign, 0 otherwise

Exploratory Data Analysis and Feature Engineering

In order to build machine learning model, the dataset needs to be transformed and solved potential problems. With the objective defined, within the scope of this project, 2 classifications algorithms utilized are Logistic Regression and Decision Tree.
There are 5 main parts to this step:

1. Categorical variables

The dataset has 3 categorical variables (Dt_Custome, Marital_Status, Education) and 26 numerical variables. These 3 variables will need to be transformed into numerical values.

1.1. Dt_Customer

This is the date of customers' enrolment with the company. The date ranges from 2012 to 2014. With this I assumed that the data is collected in 2014, and customer's age will be calculated by minus 2014 with Year_Birth variable.

1.2. Marital_Status

Marital_Status	
Married	864
Together	580
Single	480
Divorced	232
Widow	77
Alone	3
Absurd	2
YOLO	2

There are 8 levels, 3 of them accounted for only a small fraction which are Alone, Absurd and YOLO. These 3 levels will be changed into Single. Then, this variable will be one-hot-encoded.

1.3. Education

Education		Educational_years
Graduation	1,127	12
PhD	486	21
Master	370	18
2n Cycle	203	18

Basic	54	5

Create a new variable called Educational_years which is the number of years corresponded to each educational level.

2. Outliers

Outlier is a big problem that could skew and produce undesirable results.

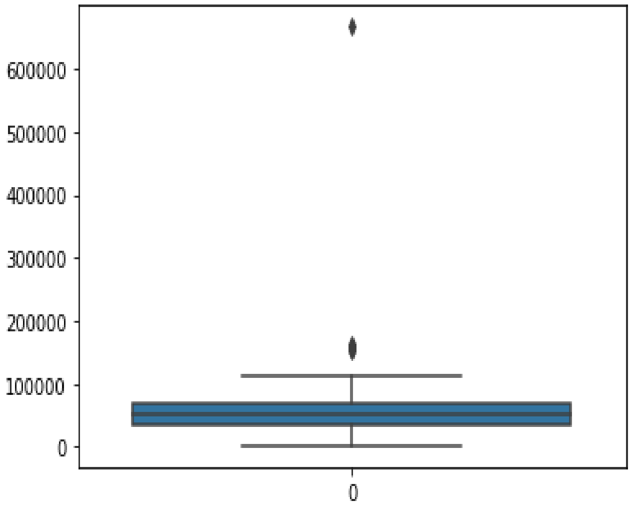


Figure 1. Income Boxplot

Income has one very extreme value; this could be entered by mistake or not a trustworthy answer and it will be dropped. Other outliers outside the IQR are unclear, so they will be kept for analysis.

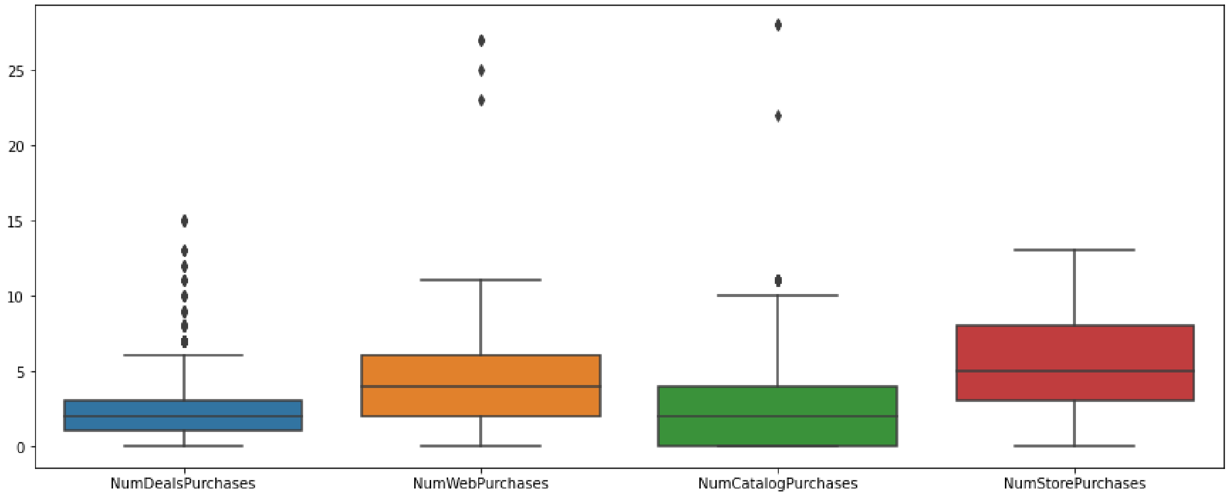


Figure 2. Number of purchases by categories boxplot

Number of purchases by categories and Monetary values of previous purchases by categories boxplot have outliers. However, still it is unclear and risky to drop those values. Instead of, Standard Scaler will be used to decrease the affect of outliers since this scaling method is robust for this situation.

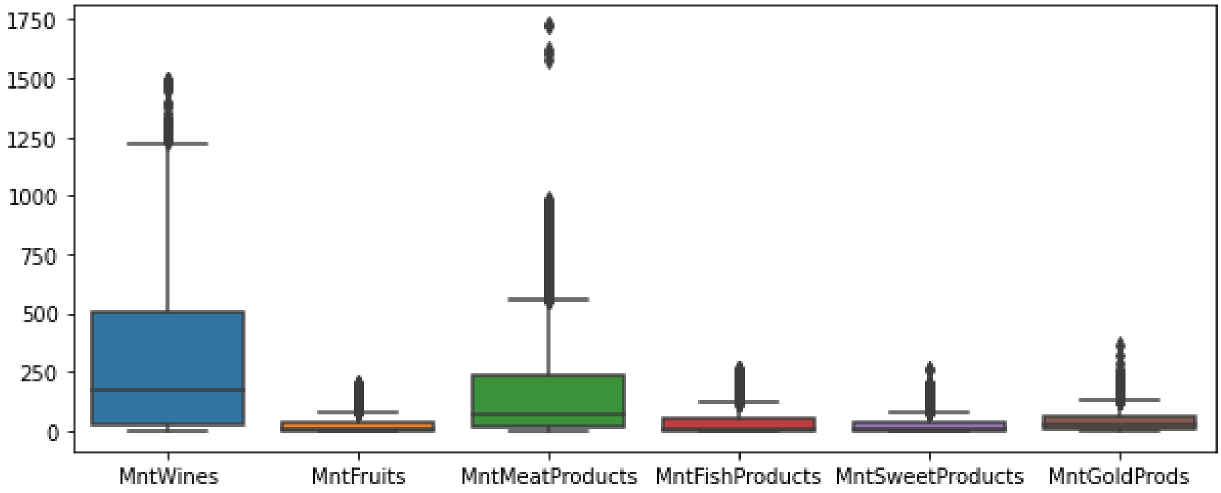


Figure 3. Monetary values of previous purchases by categories boxplot

Age variable was created in the previous step by subtracting 2014 by Year_Birth. This new variable, however, contains illogical values. 3 customers have ages of over 100 years old so they will be dropped, leaving the maximum age at 74.

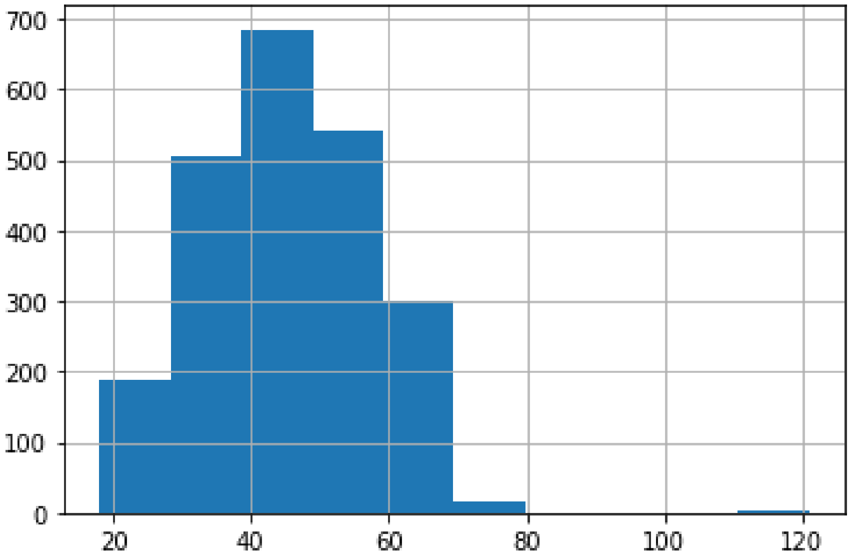


Figure 4. Age histograms

3. Imbalanced data

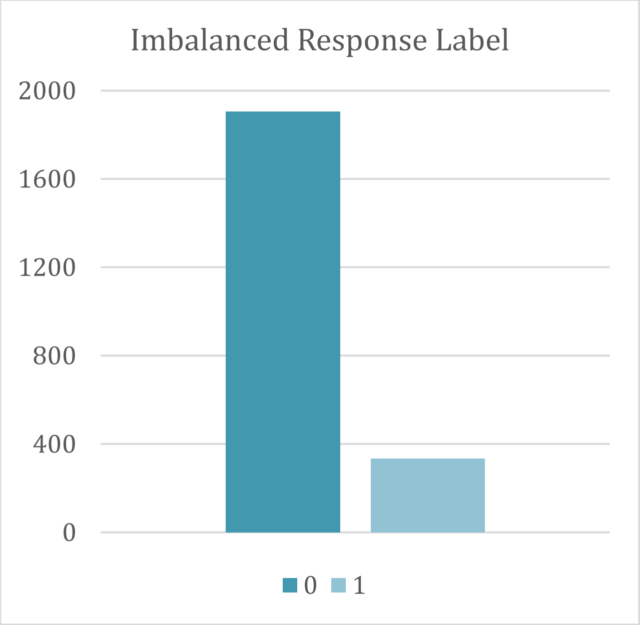
The ratio of label 0:1 is equal to 1906:334 which made the Response variable highly imbalanced. As far as I know, there are 2 possible methods to this problem:

- Down-sampling the majority label
- Up-sampling the minority label

With 2240 observations, down-sampling could result in losing valuable information, so 2 ways of up-sampling are chosen for balancing Response labels: Class-weight and SMOTE.

- Class-weight: directly modify the loss function by giving more (or less) penalty to the classes with more (or less) weight. In other words, one is basically sacrificing some ability to predict the lower weight class (the majority class for unbalanced datasets) by purposely biasing the model to favor more accurate predictions of the higher weighted class (the minority class).

- SMOTE (Synthetic Minority Oversampling Technique): creates new observations of the minority class by randomly sampling from a set of "similar" minority class observations.



4. Missing values

The dataset has one variable containing missing values which is Income. To fill them, chi-test of Independence was using to test the dependency of Income variable with few other demographic variables: Education, Marital_Status, child, age_range.
child variable was created by the sum of KidHome and TeenHome. age_range was segmented by range of 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and older on age variable.

5. Grouping and dropping unneeded variables

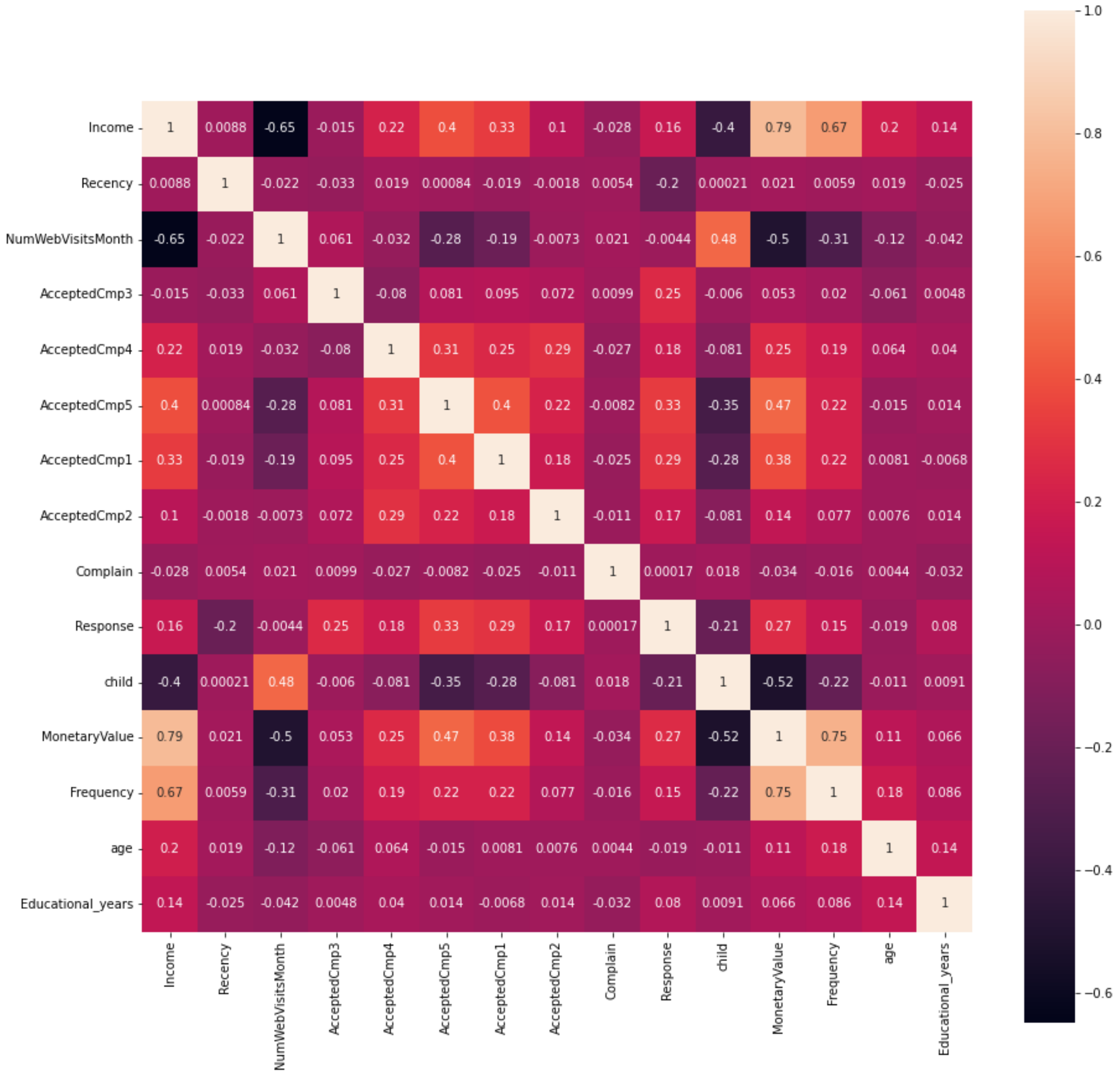
RFM analysis is a common method for customer segmentation. I group the variables based on these 3 metrics to add new features for modeling,

- Recency
- MonetaryValue: MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds
- Frequency = NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases

Results

2 classification algorithms Logistic Regression and Decision Tree was used on 2 dataframe: df_RFM_cmp and df_full.

- df_RFM_cmp included Dermographic variables, Recency, Frequency, Previous Campaign, Complain after dropping drop highly-correlated features (MonetaryValue, Frequency) as shown in correlation matrix.



	Income	Recency	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Response	child	MonetaryValue	Frequency	age	Educational_years
Income	1	0.0088	-0.65	-0.015	0.22	0.4	0.33	0.1	-0.028	0.16	-0.4	0.79	0.67	0.2	0.14
Recency	0.0088	1	-0.022	-0.033	0.019	0.00084	-0.019	-0.0018	0.0054	-0.2	0.00021	0.021	0.0059	0.019	-0.025
NumWebVisitsMonth	-0.65	-0.022	1	0.061	-0.032	-0.28	-0.19	-0.0073	0.021	-0.0044	0.48	-0.5	-0.31	-0.12	-0.042
AcceptedCmp3	-0.015	-0.033	0.061	1	-0.08	0.081	0.095	0.072	0.0099	0.25	-0.006	0.053	0.02	-0.061	0.0048
AcceptedCmp4	0.22	0.019	-0.032	-0.08	1	0.31	0.25	0.29	-0.027	0.18	-0.081	0.25	0.19	0.064	0.04
AcceptedCmp5	0.4	0.00084	-0.28	0.081	0.31	1	0.4	0.22	-0.0082	0.33	-0.35	0.47	0.22	-0.015	0.014
AcceptedCmp1	0.33	-0.019	-0.19	0.095	0.25	0.4	1	0.18	-0.025	0.29	-0.28	0.38	0.22	0.0081	-0.0068
AcceptedCmp2	0.1	-0.0018	-0.0073	0.072	0.29	0.22	0.18	1	-0.011	0.17	-0.081	0.14	0.077	0.0076	0.014
Complain	-0.028	0.0054	0.021	0.0099	-0.027	-0.0082	-0.025	-0.011	1	0.00017	0.018	-0.034	-0.016	0.0044	-0.032
Response	0.16	-0.2	-0.0044	0.25	0.18	0.33	0.29	0.17	0.00017	1	-0.21	0.27	0.15	-0.019	0.08
child	-0.4	0.00021	0.48	-0.006	-0.081	-0.35	-0.28	-0.081	0.018	-0.21	1	-0.52	-0.22	-0.011	0.0091
MonetaryValue	0.79	0.021	-0.5	0.053	0.25	0.47	0.38	0.14	-0.034	0.27	-0.52	1	0.75	0.11	0.066
Frequency	0.67	0.0059	-0.31	0.02	0.19	0.22	0.22	0.077	-0.016	0.15	-0.22	0.75	1	0.18	0.086
age	0.2	0.019	-0.12	-0.061	0.064	-0.015	0.0081	0.0076	0.0044	-0.019	-0.011	0.11	0.18	1	0.14
Educational_years	0.14	-0.025	-0.042	0.0048	0.04	0.014	-0.0068	0.014	-0.032	0.08	0.0091	0.066	0.086	0.14	1

Figure 5. df_RFM_cmp correlation matrix

- In the same manner, df_full included Dermographic variables, Recency, Number of Purchases by categories, Monetary value by Categories, Previous Campaign, Complain after dropping drop highly-correlated features as shown in correlation matrix. With this data frame, I dropped value with the absolute correlation score above 0.6.

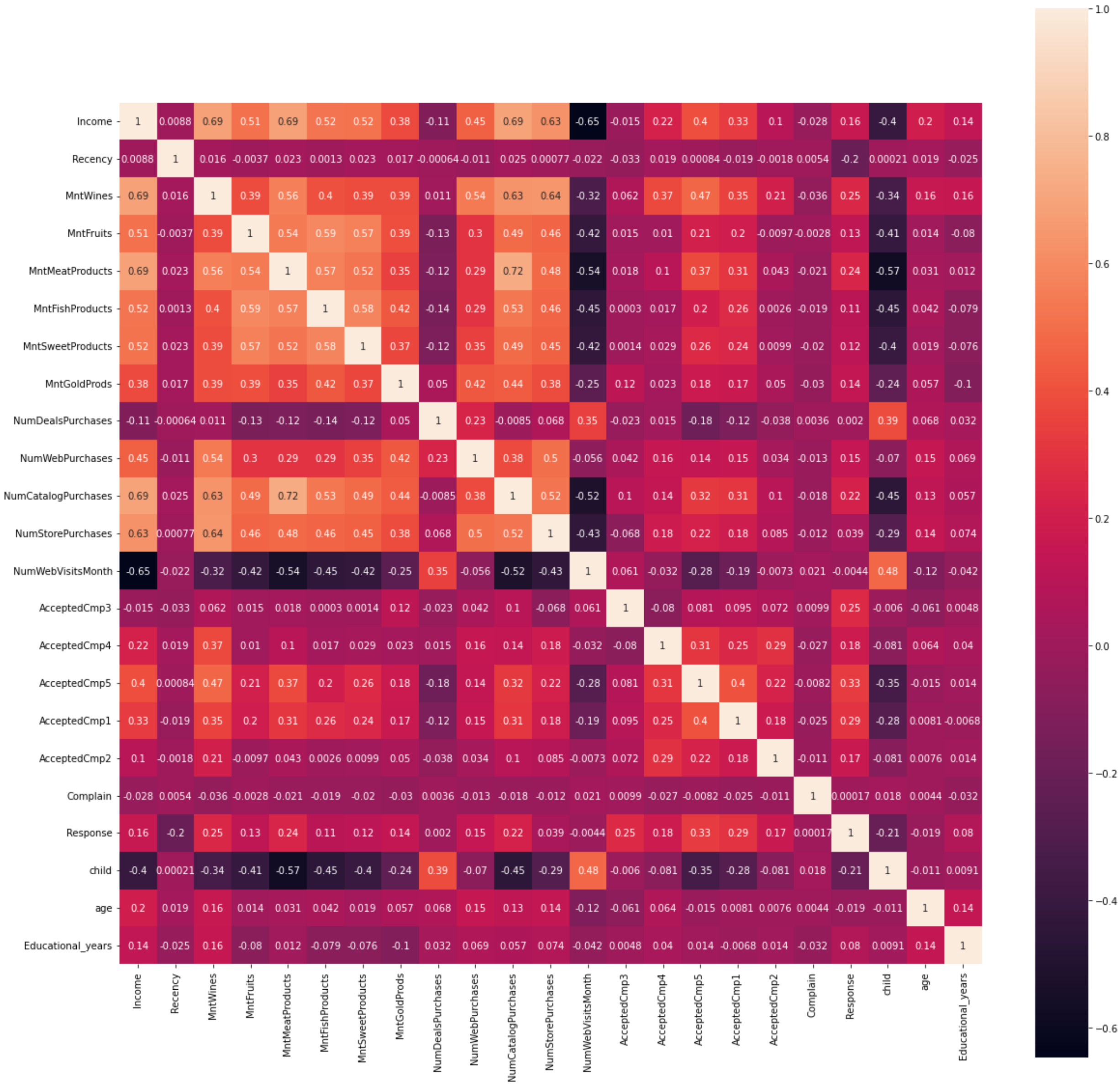


Figure 6. Figure 5. df_full correlation matrix

Below is the results table after running the 2 data frame through 2 machine learning model (Logistic Regression and Decision Tree) in 3 ways: without balancing data, balancing data using class- weight and SMOTE respectively.

Dropped highly-correlated variables		Label	Logistic Regression				Decision Tree			
			precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
df_RFM_cmp		0	0.88	0.96	0.92	0.85	0.90	0.90	0.90	0.83
		1	0.60	0.33	0.42		0.49	0.48	0.48	
	Class-weight	0	0.93	0.78	0.85	0.77	0.88	0.93	0.90	0.83
		1	0.38	0.68	0.49		0.49	0.37	0.42	
	SMOTE	0	0.93	0.74	0.82	0.73	0.88	0.84	0.86	0.77
		1	0.35	0.73	0.47		0.34	0.42	0.38	
df_full		0	0.89	0.96	0.92	0.86	0.89	0.86	0.87	0.79
		1	0.64	0.37	0.47		0.38	0.42	0.40	
	Class-weight	0	0.94	0.76	0.84	0.76	0.88	0.90	0.89	0.81
		1	0.38	0.74	0.50		0.41	0.37	0.39	
	SMOTE	0	0.93	0.73	0.82	0.73	0.90	0.82	0.86	0.78
		1	0.34	0.73	0.46		0.37	0.53	0.44	

In general, using balancing methods (class-weight and SMOTE): decrease the overall accuracy, increase the precision of label 1 prediction; but decrease label 0's counterpart. The accuracy score range is [0.73 – 0.85]. The precision score range is [0.89 – 0.93] for 0 label and [0.34 – 0.64] for label 1.

Conclusions

The overall score of label 0 score is acceptable, however label 1 precision score since the data is too imbalanced. This problem should be improved by:

- Collecting more data
- Selecting more suitable algorithms
- Choosing different methods to balance data