Team #2

Jiani Wei, Prabeen Parida, Fiona Passoni

# Machine Learning final report

# Business case

## I/ Global picture and business view

In this business case, we analyses the customer data of a bank to find trends about their behavior in term of service consumption and make predictions on various marketing offers.

Here we will target 3 products: Customer loan, Credit cards and Mutual funds. Our goal is to design probabilistic models in order to target no more than 100 clients with a marketing offer. These models will be used by the bank to determine the propensity of a client to buy after addressing him with the offer and thus target the clients with the highest propensity to buy to maximize the revenue.

Our work will help define the right target for a given product. We’re doing this in order to not market people with low probability to buy or market the same person multiple time with different products. Our objective is to limit marketing cost with a low return on investment and increase revenue. Our solution aim to reduce cost but also have a better relationship with our clients by knowing what’s the most useful for them, and what they could be interested in.

After our analysis and recommendation, performance will be measured in term of ROI (return on investment) but also in conversion rate (number of people buying on number of people targeted) in order to understand if our customer selection was right, and if it was efficient.

Looking forward, our method could also be reuse for renew contract for example, by targeting people that are close from the end date of their contract and have probabilities to renew.

We’re not trying to replace human expertise by generating this kind of model, quite the reverse, we need human to first understand the problematic and challenges, then to code, and finally to interpret result and get conclusion.

## II/ Data manipulation

## Data cleaning

We first make a work in order to clean our data. We proceed by following these steps:

At first, we gathered all the clients for whom we had all the information by merging all the excel files and fetching all the required information from the excel file to the pandas data frame(primary). Then we figured out all the clients for whom we don’t have all the information and put them all in another data frame and kept it aside. After that we pointed out all the null values in the primary data frame and replaced them with zeros.

After completing the above task, we split the data of the primary data frame into training set and test set (75 % training set and 25% test set) by using scikit learn training split function.

Then, we split our data and performed some featuring engineering on columns by converting data type of few columns to categorical.

Finally, we dropped few unwanted columns from the data frame by analyzing the co-relation among the columns of the data frame and considered only the right ones.

## Tuning the model

To increase the performance of our models we tried few tuning techniques as below:

* Hyperparameter tuning: we selected our best hyperparameters by using grid search function from scikit machine learning package in order to improve the performance of our model.
* Ensemble method: we used ensemble method to obtain better predictive performance of model. Ensemble method use multiple learning algorithms to obtain better performance than could be obtained from any of the constituent learning algorithms alone.
* At the end we performed featuring scaling on our primary data frame in order to speed up the calculation of our models.

# III/ Data exploration

The original data consist of 4 data frames, mixed (categorical and numeric) features with 40% of missing values in the sales and revenue data frame:

They fall into two major groups:

* The demographical information (age, sex, tenure)
* Information related to banking (balance, products and sales)

From these 4 datasets, exploratory analysis has been conducted by using groupby() and agg()function, displaying major trends and inner connections within the dataset by plotting, and finding the correlations among features in order to choose the appropriate ones for machine learning purposes.

The categorical variables were coded with levels [0,1] and scaled for facilitating regression prediction later in the analysis.

« Sales » and «Revenue » have been chosen as target features and a series of modelling will be conducted in the next steps.

# IV/ Solution: analysis and limitations

This report analyses the customer data to find trends and make predictions on various marketing offers like Customer loan, Credit cards and Mutual funds by designing predictive models.

This business problem could be decomposed into two parts: How to determine which clients will buy which product, and how to predict the most optimized revenue by each client.

The difficulty of this project is that each client could only be targeted once, so only one optimized product and one revenue value are needed for each client.

## Analysis

The target variables were first studied to understand the problem we were dealing with. We have determined that this is a supervised learning problem, for a specific output is demanded.

Once it was clear, what kind of algorithms we would be needing, the next step involved determining the features (columns) that would be needed for the modelling. To determine the features, a correlation matrix was computed and sorted by their correlation score with the target variable (for each case) to get the columns which are most suitable for determining the target.

In total, three different models were designed (one each for Loan, Credit card and Mutual Fund) using SVM, Random Forest and logistic regression, each model has reached at least 80% of the accuracy, and 55% of precision and recall.

## Limitations

1. The models could have been trained better with more data and better feature engineering.
2. A very low correlation was observed between the target and the features which impacted the fitness of the model.
3. Too much “0” value in the dataset, not sure if this should be tuned or not

# V/ Our conclusion

By designing data-driven models to predict on sales quantity and quality could give out statistically based marketing solutions to this bank. However, further data quality review should be conducted on the historical data recency and update accuracy. The models designed for classification and regression problems could be put into use right now in real business scenarios, but its efficiency and precision remain to be improved.

For this precise case, we decide to target clients only for credit card and customer loan service as the probabilities of client buying those products were higher. To have the best conversion rate, we choose to target only 70 clients on the 100 allowed, as the 30 clients remaining has too low probabilities to buy, and it would have not been an efficient marketing expense.