# Introduction

As Junior Data Consultants, we were approached by the Bank with the marketing challenge, 2 weeks ago. The Bank's Marketing Department wanted to understand which set of clients should be targeted based on their previous marketing initiatives, while using machine learning and data science. This selection should result in the maximum potential revenue from the Bank. To answer this question, our Team's approach was to run a predictive analytics project and this report summarizes the steps we took in order to answer the challenge that was introduced to us.

Bank currently has 3 product offerings in their portfolio:

1. Mutual Fund (MF)
2. Credit Card (CC)
3. Consumer Loan (CL)

Thus, our Team's goal was to determine 100 clients from the dataset provided and select the best product to offer them (1 out of 3 previously mentioned), all while having in mind maximum potential revenue. To accomplish this, we broke down our goal into two smaller questions that also represented the two objectives of our product:

**Q1: How likely is the client to accept an offer from each product?**

Objective is to predict how likely the customer is going to accept an offer from each product and so to select the 100 most likely clients to be receptive to the bank offer.

**Q2: How much revenue would be gained if the client accepted the offer?**

Objective is to forecast the expected revenue based on such a chosen strategy and how much would the company gain for each customer if they accept this offer.

*Note: Each of these 100 people can only receive 1 offer for 1 product. The 100 people should not have been part of the previous marketing campaign (See description of data below).*

To address the questions, our Team defined an approach that consisted of six key steps:

1. Data Exploration
2. Dataset Preparation
3. Models Exploration
4. Shortlisting the Best Models
5. Fine Tuning the Models
6. Key Findings

In the next 6 chapters, we will describe in more details our process how we plan to use the data to maximize revenue from marketing campaigns.

# Chapter 1: Data Exploration

The purpose of this project is to optimize marketing spending for a bank by identifying the 100 clients from the database that should be targeted with the offer. Optimizing the spending also means maximizing revenue from each client. Therefore, It is crucial to understand the datasets in a way that delivers maximum insights.

## Understanding Data Provided

Before we start our project, our goal is to understand the data, study each variable and its characteristics. For all variables across the five spreadsheets we have defined the name, description and the type. Also, we have provided a brief example per worksheet to better understand the scope of the data frame.

Description of data in the given data set (excel spreadsheet) consists of information points for 1,615 of the bank's clients.

The spreadsheet has the following 5 worksheets:

1. The first excel sheet contains descriptions of all 39 variables from the dataset;
2. Social-demographic data (age, gender, tenure) - available for 100% of clients;
3. Products currently owned + product balances (current account, saving account, mutual funds, overdraft, credit card, consumer loan) - available for 100% of clients;
4. Behavioral data from client's products - available for 100% of clients;
5. Sales and revenue data from a previous marketing campaign - available for 60% of clients.

The data provided is labeled, thus we are dealing with the supervised learning problem - function that maps an input to an output based on example input-output pairs. Data in worksheet *5. Sales and revenue data from a previous marketing campaign* is our **target data** in this project, as it contains variables necessary for the further predictions.

## Project Dataset Definition - Merging the Datasets

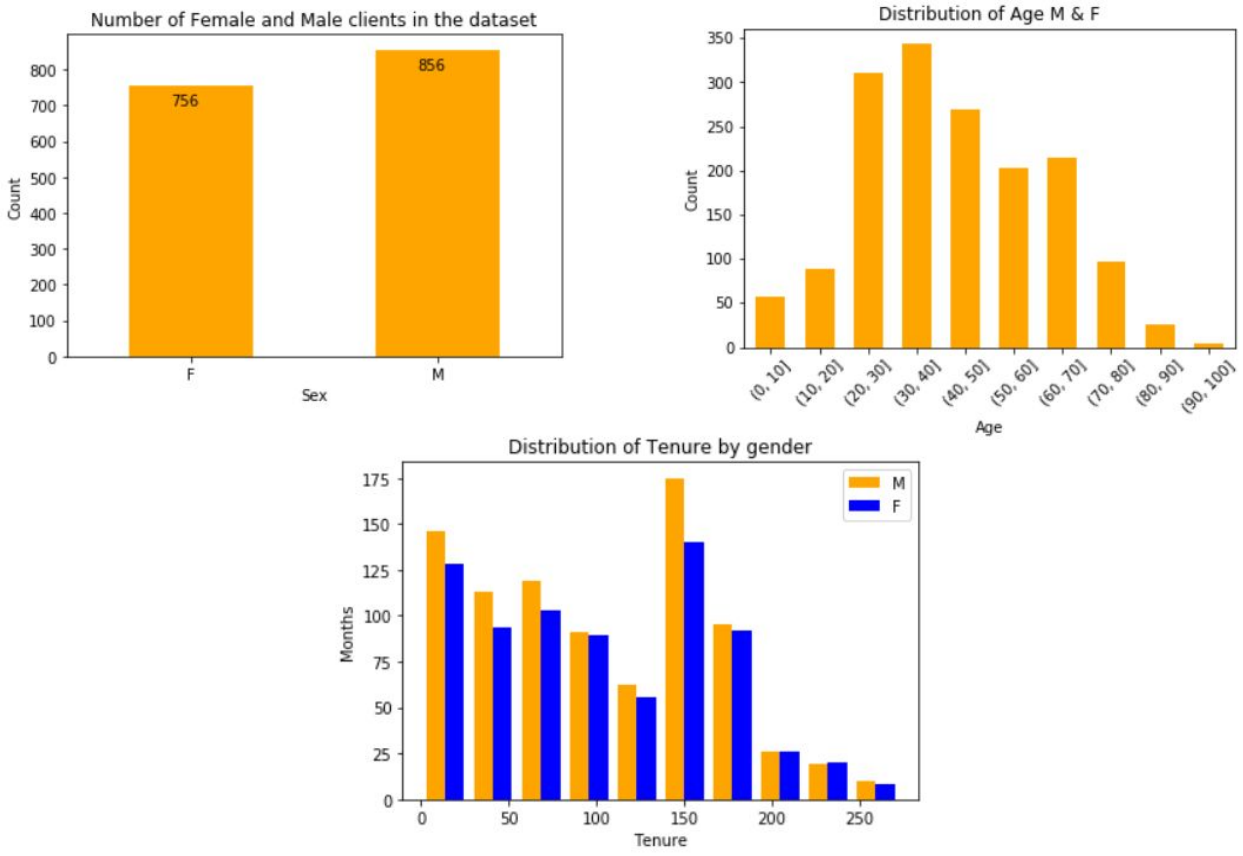
We first proceeded by joining the first and second datasets using an inner join on the Client ID column (result : df_1a). The second step was to left join the result of the previous step (df_1a) and the third dataset on the same Client ID column (result : df_1b). Next, we an inner join

between the df_1b and the dataset four, creating the base containing all information for the 60% of clients. The following code shows the steps in detail:

## Summary Statistics and Visualizations

First of all, our goal was to gain a deeper understanding of the data provided for the 100% of Bank's clients.Accordingly, we analyzed the data in worksheets 2, 3 and 4. Below are the preliminary graphs that came as a result of the data exploration.

*Image 1. Summary of Exploratory Graphs*



## Aggregated Target Data (Sales and Revenues) Overview

Next, we wanted to understand the outcomes and success rate (*target data exploration*) of the Bank's previous marketing campaign performed on 60% of their clients. 969 clients in the data have this information, and this data should be used for further modelling. We started by implementing the simple mathematical functions, to frame the bigger picture of the situation and determine the KPIs that can help us in further analysis. We performed two analyses:

- Analysis 1: Targeting **sales** results

Our goal is to understand the **success rate** of the campaign. In our case we can conclude that 59% of Bank's campaign can be treated as successful and impacted in positive call to action (CTA). Next we split the data into 4 sections: (1) Zero responses, (2) 1 response, (3) 2 responses, (4) 3 responses and determined the breakdown for each of the sections.

| | #Client | %Partition |
|---|---|---|
| 0 response | 388 | 40.8 |
| 1 response | 421 | 44.3 |
| 2 responses | 128 | 13.5 |
| 3 responses | 14 | 1.5 |

*Image 2. Insight - Sales Summary*

- Analysis 2: Targeting **revenue** results

Analysis 2 was performed with the final objective of understanding what was the aggregated revenue of the campaign and if the client responded positively to the marketing campaign - how much revenues did it bring.

| | Revenue_MF | Revenue_CC | Revenue_CL |
|---|---|---|---|
| sum | 1865.340000 | 2603.080000 | 3472.739286 |
| mean | 1.961451 | 2.737203 | 3.651671 |
| std | 10.033797 | 17.855762 | 7.784000 |

*Image 3. Insights - Revenue Summary*

*Note: The 100 clients to be selected as the final answer should be chosen from the remaining 40% (the 646 clients that are not included in this sheet).*

- Analysis 3: Cross **Sale-Revenue** Results

Next, we want to see how much was the revenue per client & product type. Finally, we will segment the amount of sales per product (CL,CC and MF) in a representation below.

| | #Client | %Clients | %Revenues | TT Sum Revenues | TT Mean Revenues | CL Mean Revenue | MF Mean Revenue | CC Mean Revenue |
|---|---|---|---|---|---|---|---|---|
| 0 response | 388 | 40.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 response | 421 | 44.3 | 59.4 | 4715.5 | 11.2 | 12.0 | 9.7 | 11.0 |
| 2 responses | 128 | 13.5 | 30.5 | 2423.6 | 18.9 | NaN | NaN | NaN |
| 3 responses | 14 | 1.5 | 10.1 | 802.1 | 57.3 | NaN | NaN | NaN |

*Image 4. Sales Revenue Summary*

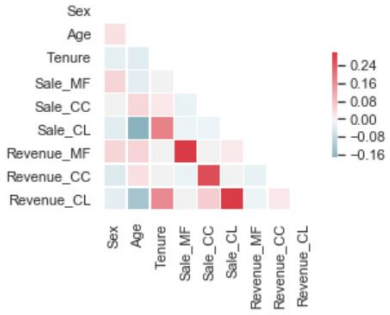
Among the clients buying:

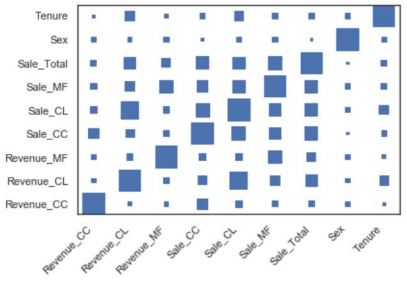
- 40.8% bought 0 products;
- 44.3% bought only 1 product. Total revenue generated: 4,715.5 EUR;
- 13.5% bought 2 products: Total revenue generated: 2,423.6 EUR;
- 1.5% bought all 3 products: Total revenue generated: 802.1 EUR.

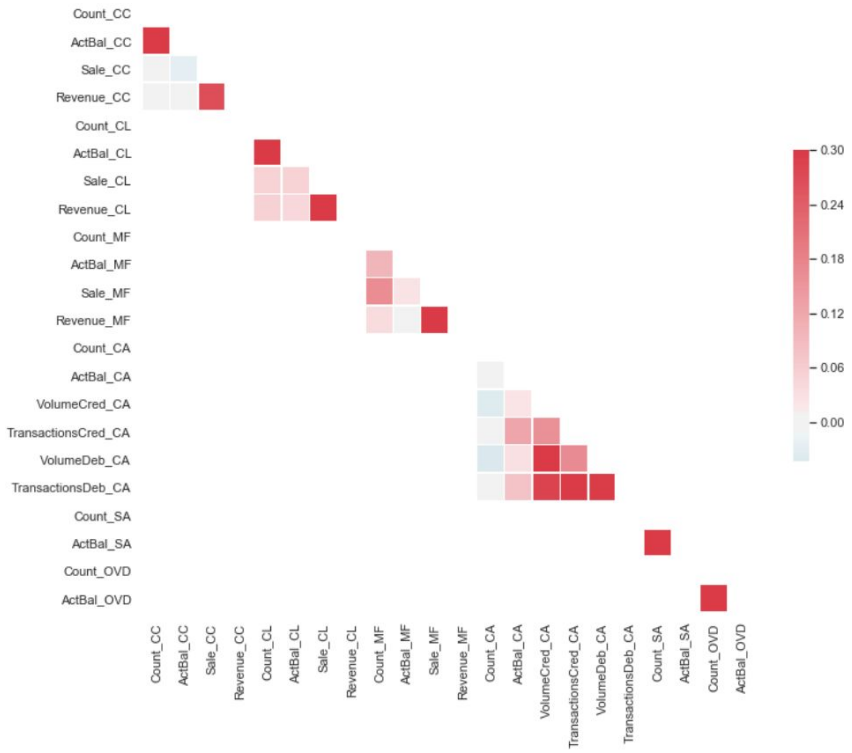
In conclusion, it seems that it is easier for the bank to sell only one product. Among those products, customer loans are the easiest to sell.

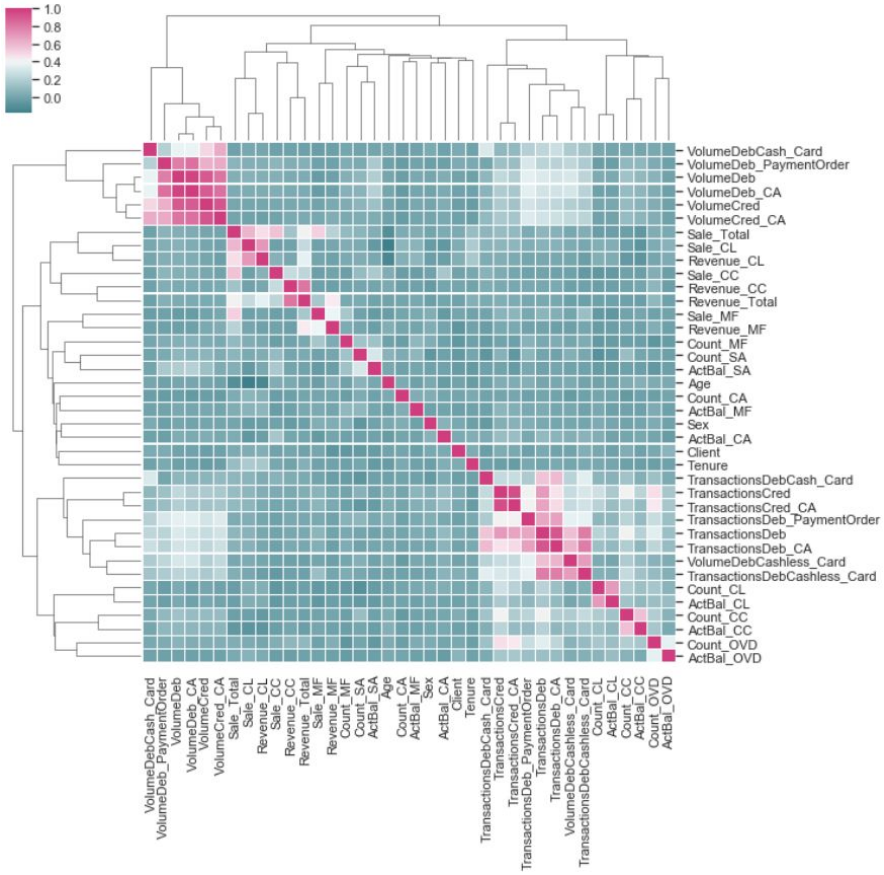
## Studying the correlations between the attributes

Studying correlations was very important at first, before starting modeling and tuning models, it allowed us to dig into variables and uncover correlations among them. Indeed we discovered that some variables are directly correlated with others : Sales/Revenues, ActBal/Count, TransactionsCred/TransactionsDeb, etc, as the following graph illustrates it. Correlations indeed can impact the way models react and are directly modifying results when trying to fit the model to our dataset. What's more it is important to then have a look at the correlations liked to each target in order to later optimize the models.




Count_CC
ActBal_CC
Sale_CC
Revenue_CC
Count_CL
ActBal_CL
Sale_CL
Revenue_CL
Count_MF
ActBal_MF
Sale_MF
Revenue_MF
Count_CA
ActBal_CA
VolumeCred_CA
TransactionsCred_CA
VolumeDeb_CA
TransactionsDeb_CA
Count_SA
ActBal_SA
Count_OVD
ActBal_OVD
0.30
0.24
0.18
0.12
0.06
0.00


1.0
0.8
0.6
0.4
0.2
0.0
VolumeDebCash_Card
VolumeDeb_PaymentOrder
VolumeDeb
VolumeDeb_CA
VolumeCred
VolumeCred_CA
Sale_Total
Sale_CL
Sale_CC
Revenue_CL
Revenue_CC
Revenue_Total
Sale_MF
Revenue_MF
Count_MF
Count_SA
ActBal_SA
Age
Count_CA
ActBal_MF
Sex
ActBal_CA
Client
Tenure
TransactionsDebCash_Card
TransactionsCred
TransactionsCred_CA
TransactionsDeb_PaymentOrder
TransactionsDeb
TransactionsDeb_CA
VolumeDebCashless_Card
TransactionsDebCashless_Card
Count_CL
ActBal_CL
Count_CC
ActBal_CC
Count_OVD
ActBal_OVD

# Chapter 2: Data Preparation

## Data Cleaning

In order to prevent poor model results and predictions due to “dirty” data, next step was to clean our data. Data cleaning was performed for both 60% and 40% clients data set.

Transformations performed on the 60% of clients data:

- **From categorical to numerical:** column sex transferred from F and M to 1 and 2, respectively;
- **Filling missing values**: replace NaN values with 0 – transformation performed for 6 columns in Count and 6 columns in ActBal
- **Dropping the rows:** all other rows containing NaN values were not considered in the further data manipulation.

After data cleaning 951/969 rows were kept for the analysis.

Transformations performed on the 40% of clients data:

- **From categorical to numerical:** column sex transferred from F and M to 1 and 2, respectively;
- **Filling missing values**:
    - replace NaN values with 0 – transformation performed for 6 columns in Count and 6 columns in ActBal;
    - replacing NaN values with average (mean) – for 7 columns in Volume and 9 columns in Transactions;

After data cleaning 646/646 rows were kept for the analysis.

## Data split

To finalize our data preparation process, next step was to split the data into training and test sets. We decided to train our models on the 80% of data (760 instances), and later on test in on the remaining 20% (160 instances). The split was performed on the 60% of client dataset, and was implemented in the same way for all three products. MF, CC, CL.

Moreover, to make sure that our model can generalize to the independent test set,we decided to implement cross-validation for testing all the models by re-sampling the data. We have split our dataset into 5 sections that are periodically fed into the model where error is calculated to avoid overfitting. The model’s overall performance on each partition was then average to get a better sense of its generalizability.

# Chapter 3: Models Exploration

Next we wanted to know, based on the training data, models encapture our problem best and apply them to target our customers that have not yet been informed about our marketing campaign. As the initial step, our Team has agreed that to answer this question, the combination of two supervised learning approaches should be implemented:

1. Firstly, using **classification** algos to predict **which client buys what** - Mutual Funds, Credit Cards or Customer Loan and solve the Q1 challenge;

   > *"Team 2 model decision making process:*
   > *Are we predicting a category? YES – goal is to answer*
   > *a question is the client going to buy a product (1) or no (0);*
   > *Is the data provided labeled? YES*
   > *ANSWER: Classification models to be used"*

2. Secondly, using **regression** to say how much is the expected **revenue** from each client and solve the Q2 challenge.

   > "*Team 2 model decision making process:*
   > *Are we predicting a category? NO – values are*
   > *continuous in the case of predicting the revenue;*
   > *Are we predicting a quantity? YES - understanding how much*
   > *revenue we can gain from our clients if targeted correctly;*
   > *ANSWER: Regression model/algorithms to be used"*

However, to solve the challenge and answer two previous questions, the next step was to determine which models are both the most appropriate ones in terms of estimated outcome and create the most accurate sample of customers to elect. Trying out various models on our dataset was crucial but also the longest step (time wise), while trying to find the best fit without overfitting the data. Our idea was to train many dirty models first and gain rough performance indicators, using standard parameters and later on improve their performance.

## Chapter 3-1: Solving the Q1 Challenge *Which Clients Are To Be Targeted?* With Classification

To address Q1 challenge we ran 7 models for each of the 3 products (21 in total). Overview of the models tested and their performance:

* Note: Random Forest could not be computed due to the code error.

| **Mutual Funds** | Precision | Recall | Accuracy |
|---|---|---|---|
| *KNN* | 75% | 77% | 77% |
| *Linear Discriminant Analysis* | 75% | 77% | 77% |
| *Random Forest* | / | / | / |
| *Naive Bayes* | 58% | 76% | 76% |
| *XGBoost* | 72% | 76% | 76% |
| *Logistic Regression* | 74% | 76% | 76% |
| *Support Vector Classifier* | 58% | 76% | 76% |

| **Customer Loan** | Precision | Recall | Accuracy |
|---|---|---|---|
| *KNN* | 67% | 72% | 72% |
| *Linear Discriminant Analysis* | 76% | 77% | 77% |
| *Random Forest* | / | / | / |
| *Naive Bayes* | 52% | 72% | 72% |
| *XGBoost* | 70% | 73% | 73% |
| *Logistic Regression* | 79% | 79% | 79% |
| *Support Vector Classifier* | 73% | 73% | 73% |

| **Credit Cards** | Precision | Recall | Accuracy |
|---|---|---|---|
| *KNN* | 72% | 74% | 74% |
| *Linear Discriminant Analysis* | 69% | 73% | 73% |
| *Random Forest* | / | / | / |

| | | | |
|---|---|---|---|
| *Naive Bayes* | 51% | 72% | 71% |
| *XGBoost* | 68% | 72% | 72% |
| *Logistic Regression* | 78% | 76% | 76% |
| *Support Vector Classifier* | 51% | 72% | 72% |

## Chapter 3-2: Solving the Q2 Challenge *What Is The Expected Revenue?* With Regression

To address Q2 challenge we ran 4 models for each of the products (12 in total). Overview of the models tested and their performance:

***** Note: accuracy could not be computed due to the code error.

| **Mutual Funds** | Precision | Recall |
|---|---|---|
| *Decision Tree Regressor* | 68.4% | 68.1% |
| *Neural Network* | Na | Na |
| *Linear Regression* | 73.4% | 42.4% |
| *SVR* | 100% | 100% |

| **Customer Loans** | Precision | Recall |
|---|---|---|
| *Decision Tree Regressor* | 66% | 64% |
| *Neural Network* | Na | Na |
| *Linear Regression* | 80% | 33% |
| *SVR* | 100% | 100% |

| **Credit Card** | Precision | Recall |
|---|---|---|
| *Decision Tree Regressor* | 65% | 64% |
| *Neural Network* | Na | Na |

| | | |
|---|---|---|
| *Linear Regression* | 61% | 44% |
| *SVR* | 100% | 100% |

# Chapter 4: Shortlisting the Best Models

**The most significant variables for each algorithm:** To shortlist the best models we needed to decide what was the right metric to take into consideration when choosing the model. Our options were: Accuracy, Recall, Precision, F-Score & Specificity, and the question was which one should we optimize on? Based on Towards Data Science article "Choose precision if you want to be more confident of your true positives." Thus, our Team decided to proceed with the *Precision*, as our rationale was to follow the theoretical definition of it and apply it to our case study. The key question that precision answers would be "How many clients that are labeled as "BUY" will actually BUY?"

Objective 1 (sales) models selected for deployment:

- Sale_MF: LDA
- Sale_CL: LDA
- Sale_CC: Logistic Regression

Objective 2 (revenue) models selected for deployment:

- Revenue_MF: Linear Regression
- Revenue_CL: Linear Regression
- Revenue_CC: Linear Regression

# Chapter 5: Fine Tuning the Models

## Feature Selection and Engineering

- Drop the attributes that provide no useful information for the task
- Add promising transformations of features
- Aggregate features into promising new features

By using Scikit-learn's SelectKBest() function we were able to exclude those features that reduced our model's precision since it was the metric that we were optimizing for. SelectKBest is a known function that takes two arguments:

1. **Score_func:** which can refer to a multitude of methods (`f_classif`, `chi2`, `f_regression`, `SelectFdr`... )
2. **K:** which takes either an integer to specify the top best features or the string 'all' to keep all features. It takes 10 as a default value.

In our case, and after trying dozens of combinations between Score functions and k number features, we were able to fine-tune our parameters to maximise precision. For logistic regression models for instance, the Chi squared function along with a k = 14 proved to be the best combination.

# Chapter 6: Key Findings

From our models, once finalized and deployed, we get the clear final scheme of the overall results. To obtain the insights we run our model on the 40% of data for both classification and regression challenge and tried to identify patterns among the remaining clients. What we have found out the list of the clients most likely to convert (list provided in the Python notebook), as well as the expected revenue that is 5592 EUR.

| Client | Revenue_MF_Pred | Revenue_CL_Pred | Revenue_CC_Pred |
|---|---|---|---|
| 3 | 4.339805 | 4.091717 | -0.342604 |
| 4 | -0.396705 | 6.918472 | 2.380970 |
| 5 | 1.961506 | 7.437693 | 8.837391 |
| 7 | 3.159796 | 4.118803 | 2.023971 |
| 9 | 3.306984 | 5.052171 | 6.551210 |
| ... | ... | ... | ... |
| 1606 | 1.571951 | 2.479365 | 1.405348 |
| 1609 | 1.063827 | 4.340584 | 0.666019 |
| 1610 | 4.981343 | 1.471174 | 7.996644 |
| 1611 | 1.570916 | 5.208002 | 7.168490 |
| 1614 | 1.264202 | 2.442955 | 1.223763 |

**Revenue MF:**
After running our model we get this final table that contains the client IDs that would respond to the sales offer along with a predicted revenue figure. The sum sales for Mutual funds is 1391,77. Concerning Revenues from CL it amounts to a total of 2325.41 from the following clients. Finally, for credit cards, the total estimated revenue is 1874.82.
Overall, the final predictions output is a dataframe of (100, 6) dimensions that looks like the following:

| | Client | Sale_MF_Pred | Sale_CL_Pred | Sale_CC_Pred | Sales_Total_Pred | Revenue_Total_Pred |
|---|---|---|---|---|---|---|
| 211 | 506 | 0.0 | 0.0 | 1.0 | 1.0 | 45.630002 |
| 608 | 1510 | 1.0 | 1.0 | 0.0 | 2.0 | 25.973168 |
| 63 | 153 | 0.0 | 0.0 | 1.0 | 1.0 | 25.456194 |
| 317 | 785 | 0.0 | 0.0 | 1.0 | 1.0 | 24.287195 |
| 419 | 1051 | 1.0 | 1.0 | 0.0 | 2.0 | 23.588115 |
| ... | ... | ... | ... | ... | ... | ... |
| 513 | 1271 | 0.0 | 1.0 | 1.0 | 2.0 | 11.498977 |
| 155 | 362 | 0.0 | 0.0 | 1.0 | 1.0 | 11.477636 |
| 573 | 1440 | 0.0 | 0.0 | 1.0 | 1.0 | 11.417434 |
| 523 | 1304 | 0.0 | 1.0 | 1.0 | 2.0 | 11.413596 |
| 194 | 463 | 1.0 | 1.0 | 1.0 | 3.0 | 11.410122 |

# Conclusion and Reflection

The models implemented could also overcome the business initial problem with creating clearer types of customer clustering. Indeed, not only providing them with the final list of clients to contact, such analysis could give an indication of the type profile of clients they could consider to target more generally as for further later communication campaigns. Creating patterns which the company could definitely deep on using afterwards along with our database high potential scaling model.

To conclude, we have learned that the secret of marketing optimization is to know which data truly brings companies' an added value. This project has been very insightful and challenging at the same time. Choosing the right model proved difficult and further experimentation is necessary before final adoption. However, the analysis and exploration conducted at the earlier stages did provide us with different paths that we can follow. Namely, further feature selection, regression for estimating future revenues and classification to identify the right customer IDs to target.

# References:


- Hands-On Machine Learning with Scikit-Learn and TensorFlow
- Scikit Learn
- StackAbuse
- https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f
- https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124