**Direct Marketing Data Challenge**

**Objective:** Objective of this project is to maximize revenue from direct marketing campaigns for consumer loan, mutual fund and credit cards. For this analysis I would be using python.

**Data Exploration:** We were provided with the four tables which consist of demographic details of the client, Products owned and Volume of products, Inflow and Outflow transactions for past three months of the client and target data for predicting sales and revenue of consumer loan, mutual funds and credit cards.

**Statistical Analysis of Data:** The data provided consist of 1615 unique clients out of which only target for only 946 clients were other then these clients other clients are new client and we have to predict products to maximise the revenue.

Before applying any Machine Learning Algorithm, we have to make sure that all the columns are numeric types as ML algorithms only work with numeric features. While checking the type of feature, Sex is having Object type which needs to be converted to numeric type.

In statistical analysis I can see that all the rows are unique and have no duplicate values. After checking duplicate rows, I will check for null values in the data set, and I found that there are multiple null values in the data set.

1. Three rows don’t have any values in Sex column.
2. Almost 70-80% of data were missing in different columns of products and volume-based table.
3. 1.9% of data which is 18 rows were missing in transaction-based table for clients.

These missing values need to be handled first before applying any Machine Learning models since mostly the Machine Learning Algorithms are perform poorly with the null values.

In statistical analysis we can also see that some of the features in the data set are having very big range and some are having very small range of values which needs to be handled because if do not get all the features in the similar scale some of our model will not perform well.

While doing the analysis I noticed that minimum value for Age column is zero which needs to be handled.

**Data Pre-Processing Steps:**

1. First, I will impute Sex column with ‘U’ for null values which means unknown, considering that some of the clients don’t want to reveal their gender.
2. After imputing Sex Column for null values, I will change the type of Sex column for object to numeric and imputing ‘M’ with 1, ‘F’ with 0 and ‘U’ with 2.
3. Columns related to products and volume of products for client which are having null values (Around 70-80%) such as Account balance in Savings Account/Current Account or Overdraft etc., I will impute these values with zero considering clients is not having these services from the bank or not availing the products from bank. As if they don’t have CA/SA than they don’t maintain the balance also in the bank.
4. Columns related to transaction history of the client which are having null values (Around 1.9% or 18 rows in different columns) such as number of all debit transactions, number of all debit transactions cashless, monthly number of debit cashless transactions via card etc., I will impute these values also with zero considering client might not be active in past three months and not doing any transactions.
5. After doing all the imputations for null values and changing the type of the column I would check for values in Age column since it has minimum value as zero. I would check if there is any row where Tenure with the bank is greater than Age of the client which is obviously not possible. After analysing this there are 34 rows where Age is less than the tenure, so we have to impute these values also as they all are current account holders. I would impute these values with 10 years plus tenure with the bank.
6. After imputing values will mark the ages with NA where age is less than 10 years considering that even to have a student account client must be at least 10 years. I would impute these values with KNNImputer which works on the principal of K nearest neighbour.

**Training, Testing and New Data Set:**

Before applying any machine learning model I divided the data set into three parts first in training and new data. Training data is 60% of the data set for which target is provided and we have to train our model on that. New data is something that I would be using to recommend items. From training data, I would divide it further in training and testing to evaluate our model. I would divide in 80-20 ratio as I want to give more data to train our model.

**Model Evaluation and Selection:**

**For predicting sale of product- classification algorithms:**

It is class imbalance problem with ratio of 80:20 in sale of mutual fund, 75:25 in sale of credit card and 70:30 in consumer loan. I will use models which work well with these kind of imbalance data set.

I would use XGBoost and AdaBoost since it is a boosting algorithm and adds weight to every incorrect classification and improve the accuracy. Boosting algorithm works good with imbalanced data set. I would also use Random Forest and logistic regression by adjusting the class weight because I can see some linear relationship between features and also categorical variables.

I would evaluate these models and then decide which algorithm needs parameter tunning for improvement and which model will be used for final prediction. I would be doing scaling of data before applying any non-tree algorithm because tree-based algorithm don’t need any scaling.

For all the three-propensity model logistic regression and Adaboost classifier works well. I would perform parameter tunning to improve the performance of the model. Finally logistic regression will be best model, and I will use this model for final prediction. Metric that I will be using is Precision, Recall, Accuracy and ROC-AUC score.

**For predicting revenue of product- Regression Problem**

In this also we will be using multiple models and evaluating those models and then deciding which model to use for final prediction. We evaluate multiple models because in multi space different feature behave differently with target. So, we use multiple models to capture as many possible conditions as possible and test performance of the model.

We will be using Random Forest and Ada Boost regressor and along with this we will use Ridge Regression which penalizes the data for incorrect prediction. These three algorithms were used to capture as much as possible. Random Forest algorithm is used because there are some outliers in the data set with balance columns and random forest can work with this data set. Random Forest can also work with highly corelated data as it selects the features at random and create trees for prediction.

For model evaluation I would be use Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Square Error (MSE). I will try to minimize these error metrics as much as possible.

Random Forest and Ridge performed well with the given data set and after parameter tunning random forest is best performing model. I will use this for my final prediction of revenue.

In both classification and regression problem I would be using cross-validation technique for evaluating and finally training our ML models.

**Expected Revenue Calculation:**

To calculate expected revenue, I will first predict the sale of particular item using our trained model and then will calculate revenue generated from that client.

To calculate expected revenue, I would multiply the probability with revenue and then see for the maximum expected revenue from all the three items for final prediction. Condition provided to us is that one customer can be contacted only once and with one offer only so we will contact a client with maximum revenue, to maximize the revenue.

The maximum revenue which is generated from this strategy is almost **1022 Euros** the condition is that we would be contacting only 100 people from 646 with direct marketing approach for maximising the revenue of the bank.