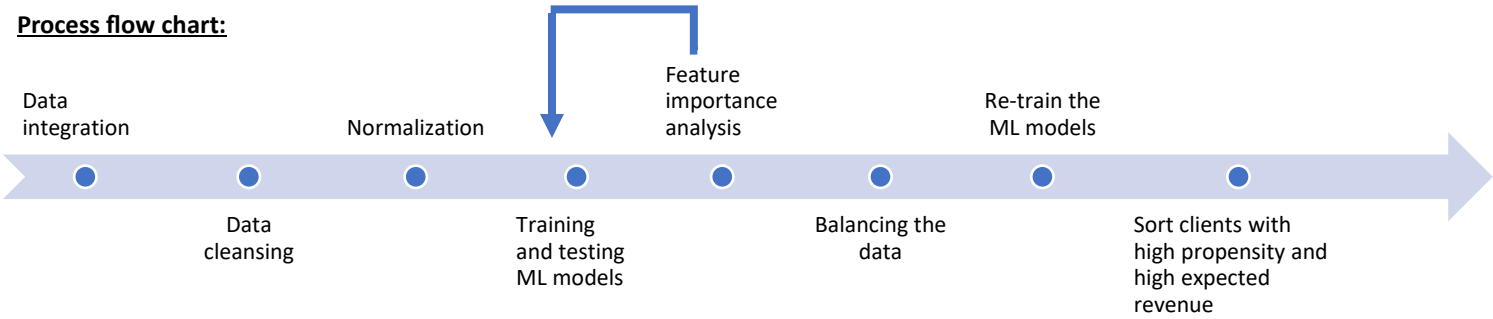
# Executive Summary: Direct marketing optimization

## Goal:

- To build customer propensity models for Mutual Funds (MF), Credit Card (CC) and Consumer Loan (CL) based on different categories of data provided.
- To generate a list of clients and the corresponding product that they can be targeted with.
- To maximize revenue based on the propensity scores.

Towards addressing these goals, three machine learning models have been designed: Logistic regression, Decision tree classifier and XGBoost for each of the products mentioned above.

## Process flow chart:




1. Data integration: Merging data sources, using Client_ID as unique identifier.
2. Data cleansing:
    a. One hot encoding of the categorical variables such as Gender, adding additional gender field 'Other' where data is missing.
    b. Replacing missing values in numerical variables with a value of 0, instead of deleting the records.
    c. Data manipulation based on assumptions.
3. Normalization: Min-max scaling of the data, except Client_ID and target variables. This is done to scale features in the range [0,1] and thereby optimize the learning capacity of machine learning models.
4. Training and testing ML models:
    a. Splitting the processed data into training and testing (85% - 15% proportion)
    b. Training data: all features from the processed data. Testing data: Binary target variables (Sales_MF, Sales_CC, Sales_CL) selected based on the product to be modelled.
    c. Evaluation of machine learning models based on metrics such as Precision, Recall, F1 scores for class values of 0 and 1. Note that accuracy is not used to compare model performance, as the data is observed to be highly imbalanced (Data available for Sales value of 0 >> Data available for Sales value of 1).
5. Feature importance analysis: Here, a decision tree classifier has been used to analyse feature importance (can be used for dimension reduction in larger datasets) for each of the propensity models. This step is followed by re-training the machine learning models to look for any improvement/deterioration in model performance based on the reduced feature set. In this case, there was no major impact on model performance, hence the entire set of features have been used for further processing.
6. Handling imbalanced data: All the machine learning models returned higher scores (Precision, Recall and F1 score) for Sales target value of 0, but the corresponding scores for Sales target value of 1 were very low. Hence, the distribution of target data (Sales field) has been studied. Figure 1 shows the imbalance in proportion of data for Sales target values of 0 and 1. To address this, random oversampling was done to increase the proportion of data belonging to sales value of 1.


(a) (b)

Figure 1: Target variable distribution (a) Before oversampling (b) After oversampling

## Assumptions in data manipulation:

- Delete records where client age < Tenure in years.

- Delete records where a client with age < 18 has either a credit card or consumer loan. This follows from the general guidelines that a minor can have a mutual fund in their name, but not a credit card or consumer loan.

**Results:**

Average F1 scores (macro average for target value of 0 and 1) have been used to compare machine learning models before and after oversampling. Figure 2 is a comparison plot to indicate that XGBoost classifier outperforms the other two machine learning models for all the propensity models, based on the test data considered here. Hence, XGBoost classifier has been chosen to generate the list of clients based on high propensity and high expected revenue.

Total expected revenue following this strategy (using XGBoost classifier) = 1652.834918

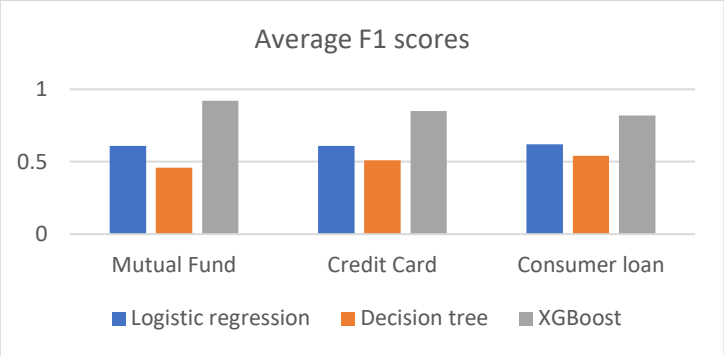



Figure 2: Comparison of machine learning model performance for different products

**Strategic points:**

- Expected revenue is calculated from the revenue given in the data and the customer propensity returned by the model. This implies higher the propensity to buy the product, closer is the expected revenue to the actual revenue furnished in the data.

Expected revenue = Customer propensity x Actual revenue

- For revenue maximization, clients have been sorted based on high expected revenue. 15% of clients with high expected revenue for each product have been merged to return the list of clients with high overall expected revenue. This list also informs the products to be targeted for each client.
- Based on the list of clients with high overall expected revenue, it is observed that there are some records where clients have been assigned an actual revenue value of 0, but the model returned high propensity scores. This can be explained with an example as follows:

| Client | Propensity | Revenue | Revenue_adjusted | Product |
|---|---|---|---|---|
| 75 | 0.986301541 | 0.004821429 | 0.004755382 | MF |
| 826 | 0.967032313 | 0 | 0 | MF |
| 963 | 0.958244205 | 0 | 0 | MF |

(a) Propensity and expected revenue returned by XGBoost classifier

| Client | Count_CA | Count_SA | Count_MF | Count_OVD | Count_CC | Count_CL | ActBal_CA | ActBal_SA | ActBal_MF | A |
|---|---|---|---|---|---|---|---|---|---|---|
| 825 | 1 | | | | | | 28.57142857 | | | |
| 826 | 1 | | 1 | 1 | 1 | | 0 | | 62103.26857 | |
| 827 | 1 | | | 1 | 1 | | 49.19357143 | | | |

(b) Count_MF and ActBal_MF for Client 826, from the Products and Balance data given

Consider Client_ID = 826, whose actual revenue given in the data is 0. But this client has a mutual Fund account with a considerable balance (as highlighted above). Hence, the model returned a higher propensity score of 0.96. This can be supported by the fact that the features 'Count_MF' and 'ActBal_MF' scored high on importance, as returned by decision tree classifier. Hence it might be important to include such clients in the target list, although the given data suggests that the actual revenue return from them is 0.

**Scope for improvement:**

- Deploying increasingly complex machine learning models like Feed forward Neural network architectures, that can efficiently learn from imbalanced datasets; other unsupervised algorithms for customer segmentation.
- Exploring other techniques to handle imbalanced data.
- Machine learning models for revenue prediction: Using the entire Sales_Revenue data (60% of the dataset) for training and predicting revenue from the remaining data.