Executive Summary of the Analysis

Methodology

The method to solve this marketing optimization problem can be described as follows:
- All the features in "Soc_Dem", "Products_ActBalance" and "Inflow_Outflow" tables are merged together to make the prediction.
- For each of the products "MF", "CC" and "CL", a classification model is built up to predict the buying propensity.
- For each of the products "MF", "CC" and "CL", regression models are built up to predict the expected revenue and its 90% confidence interval.
- Based on the predicted buying propensity and the predicted revenue for each potential client, expected total revenue from buying all the three products is calculated.
- The 100 clients with the highest expected total revenue from buying all the three products are selected as the suggested target clients for carrying out the direct marketing campaign.

Brief Description of Data Analysis and Data Preparation

- 951 valid records are used to train the classification models for buying propensity prediction.
- Among the 951 valid records, for each product, only the records with non-zero revenue are used to train the regression models for predicting expected revenue and its confidence interval.
- 636 Client IDs with valid feature values but not being included in the "Sales_Revenues" table are used to apply the prediction models.
- All the missing values of the existing numerical features are imputed as 0.
- One-hot encoding is used to convert the categorical feature to numerical features.
- There are high imbalance in the values of the target variables "Sale_MF", "Sale_CC", and "Sale_CL"; the target rates are quite low.

Classification Modeling

To generate balanced dataset, different oversampling techniques are investigated. Also, different ensemble modeling algorithms are used to build up the classification models, and model quality has been compared between these algorithms based on cross-validation. Hyperparameter searching has been performed to optimize the final model training.

Test results show that the classification models have good quality. The test ROC curves for predicting the buying propensity of the three products are shown below:
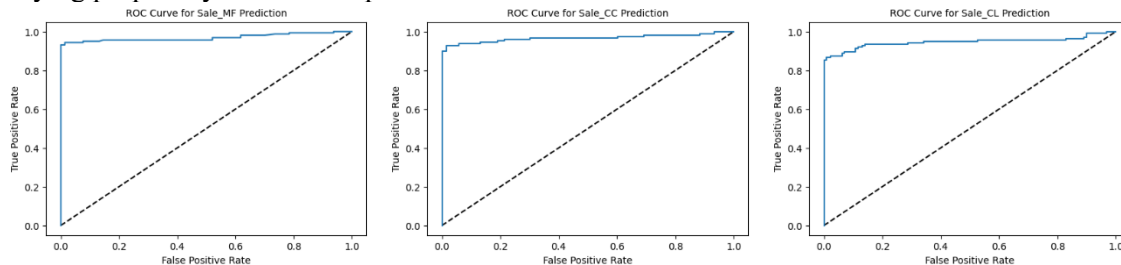
Regression Modeling

Different ensemble modeling algorithms are used to build up the regression models, and model quality has been compared between these algorithms to select the one for building final regression models. Since the size of the training data is small for all three revenues, it is hard to build up the regression models in a good quality. GradientBoostingRegressor is used to build the regression models to predict the revenue, as well as to predict the 90% confidence interval of the revenue.

The test prediction and predicted intervals for the three revenues are illustrated below. It can been seen that the prediction of the revenues are not accurate, while the actual value of the revenues are usually within the prediction interval.
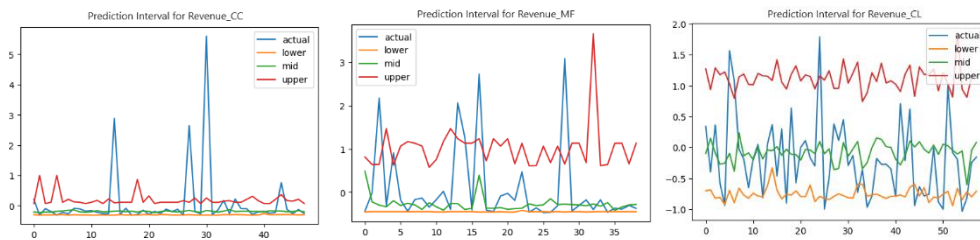
Targets and Expected Results Marketing Campaign

Based on the predicted buying propensity and the predicted revenue for each potential client, expected total revenue from buying all the three products is calculated, and the 90% confidence interval of the expected total revenue is also estimated. The 100 clients with the highest expected total revenue from buying all the three products are selected as the campaign targets.

The model-based expectation of the campaign can be summarized as follows:
- Expected Number of Clients to Buy MF: 24
- Expected Number of Clients to Buy CC: 31
- Expected Number of Clients to Buy CL: 46
- Expected Total Revenue: 898.87
- 90% Confidence Interval of Expected Total Revenue: [178.26, 2881.51]

The 100 target clients and their buying expectation are listed in the worksheet "Campaign_Targets" of the attached Excel file *Campaign_Targets.xlsx*.