In summer 2021, I started an internship at the Financial Technology Department of China Construction Bank. With an objective to increase the marketing efficiency of a balanced fund and maximize the benefit of our promotion, my task was to build models that could accurately predict the response probability of new customers to the promotion and select customers with a high response probability over 90%. Through a series of feature selection, data preprocessing and model building in a sample of over 60,000 potential customers, I eventually pinpointed 231 targeted customers with response rate greater than 90%. By targeting only 0.4% of the customer base, I managed to reduce the marketing cost significantly. Most importantly, the first-hand insight gained in the full participation of such a real-world project allowed me to apply my analytical and creative problem-solving skills.

I participated in the whole process of the project, from data selection, data preprocessing, feature engineering, to modeling and evaluation. Starting with feature selection, we were required to identify potential attributes in either positive or negative correlation with customer's likelihood in purchasing the balanced Fund. After a discussion with my mentor, we perceived that these five categories of features would indicate customers' purchase reference – customer basic information, purchasing habit, asset & products summary, and credit history. The five categories contained 300 features which would later be reduced in feature engineering stage.

Next, we cleaned up the data through data preprocessing including null value imputation, encoding categorical variables (one-hot encoding), and splitting dataset, to prepare the dataset for further analysis. To reduce the features prepared in the beginning, we conducted feature engineering such as information Value and collinearity, and eliminated insignificant features to further reduce the features to only 87. It was during this phase of feature engineering where I encountered the most difficulties. When converting dataset into weight of evidence (WOE) to impute categorical variable and better distinguish responsive from non-responsive customers, the return always contained null values for certain bins of a few variables. Numerous methods attempting to solve the problem failed, including binning methods based on bin edges and sample quantile, until I tried feature engineering. Locating the root cause of the problem on incomplete data preprocessing in highly unbalanced categorical variables, I was able to solve these issues using downsampling and upweighting.

With all these previous steps, I then moved forward to model building in logistic regression, XGBoost, LightGBM, and Catboost models, which were evaluated by measure of lift. The whole model was used to indicate the chance of finding positive customers and compare it to that of random selection. Take CatBoost model as an example, it achieved the highest lift of 3.17, indicating that our model was able to receive 3.17 times more positive responses than offering random promotion. Eventually, we identified target customers with response likelihood over 90%. Our model significantly increases the marketing efficiency by feeding promotion to only 0.4% of the previous feed customers.

This precision marketing project can not only improve operational efficiency by applying the automated machine learning workflow; more significantly, these machine learning models are much more powerful than traditional marketing campaigns based on business intuition because they can incorporate feedback into the loop that continually updates the model, making it more accurate over time.