Personal Loan Customer Conversion Analysis

Samridhi Srivastava

Use case and Objective

- The bank wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.
- Using the given Data set, Machine Learning Models have been built which can predict if a Customer would take up Personal Loan if the targeted Marketing Campaign is done.
- This presentation summarized the approach and observations.

Basic Data Analysis

From initial viewing of the dataset, the following was observed:

- There were 5000 rows and 13 columns in the dataset
- Income and Family Size had some missing values
- Education, Internet Banking and Personal Loan columns were of object type, rest were numerical
- There were no duplicated rows

```
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  5000 non-null   int64
 1   Age                 5000 non-null   int64
 2   Experience          5000 non-null   int64
 3   Income              4980 non-null   float64
 4   Postal Code         5000 non-null   int64
 5   Family Size         4991 non-null   float64
 6   CCAvgSpending       5000 non-null   float64
 7   Education           5000 non-null   object
 8   Mortgage            5000 non-null   int64
 9   Investment Account  5000 non-null   int64
 10  Deposit Account     5000 non-null   int64
 11  InternetBanking     5000 non-null   object
 12  Personal Loan       5000 non-null   object
dtypes: float64(3), int64(7), object(3)
```

Exploratory Data Analysis - Univariate

By using various visualization plots (available in the Jupyter Notebook), the following things were found:

- Most of the customers had a family size of 1 followed by 2, 4 and 3
- Most of the customers had Undergrad degree followed by Advanced then Graduate
- Most of the customers didn't have an Investment or Deposit account with the bank
- There were around 60% customers using internet banking versus 40% who weren't
- Only 9% customers took Personal Loan, hence the dataset was imbalanced
- Age and Experience had close to normal distributions
- Income, CCAvgSpending, and Mortgage had right skewed distributions, so log transformations were performed on them

Exploratory Data Analysis - Bivariate

Similarly, the following things were found from Bivariate plots:

- Variables Age and Experience did not have much differences in their distributions with respect to the target variable(Personal Loan)
- For Income, CCAvgSpending the distribution had higher mean when Loan was taken and lower when not
- For Mortgage the median was 0 for both Loan taken and not taken
- More customers with Family Sizes 3 and 4 took the loan compared to 1 and 2
- Lesser people with Undergrad degree took loan compared to other degrees
- People with deposit account had a much higher chances of taking loan
- Investment Account and Internet banking had less influence on taking loan

Data Pre-processing

- The variable ID was dropped
- Yes and No values of Internet Banking and Personal loan was replaced by 1s and 0s
- One hot encoding was done for Education
- Missing values of Income and Family Size was replaced by their medians with respect to the target variables
- Postal Code was stripped down to have the first 2 digits and then it was one hot encoded
- A new variable indicating spending to income ratio of a customer was created
- After seeing the correlation heatmap, Experience was dropped as it was highly correlated with Age
- Income, CCAvgSpending, and Mortgage was also dropped as they have their log transformed versions

Model Evaluation Criteria

- The model could have made two types of wrong predictions, namely:
  - False Negative: Predicting a customer will not take the loan but in reality, the customer will
  - False Positive: Predicting a customer will take the loan but in reality, the customer will not
- Here both the cases were important as:
  - If we predict that a customer will not take loan but he/she would actually have then the bank will lose on a potential customer
  - If we predict that a customer will take the loan and he/she doesn't, the bank would lose its time and resources on that customer
- So, we would want the **F1 Score** to be maximized, the greater the F1 score, the higher the chances of minimizing both False Negatives and False Positives

Model Building

- The data was split into train and test sets with 70-30 split, using stratify on target variable
- MinMaxScaling was performed on the features as it is helpful for some models
- The below 4 models were built:
  - Logistic Regression
  - Support Vector Machine(SVM)
  - Random Forest
  - XGBoost
- For dealing with class imbalance, model parameters like class_weight and scale_pos_weight were used

Model Results

Performance of different models has been summarized below:

- For Logistic Regression, the average F1 score ('macro avg' from classification report) for test set was 85%, and the model was not overfitting as there was very little difference between the scores of train and test data
- For SVM, the average F1 score for test set was 89%, but the model was slightly overfitting with a difference of 7% between the scores of train and test data
- For Random Forest, the average F1 score for test set was 94%, and the model performance using K-fold cross validation was ~96%, hence the model was performing very well
- For XGBoost, the average F1 score for test set was 96%, which was the same for K-fold cross validation as well, hence this model was also performing very well
- The common important features identified from the models include – Income, CCAvgSpending, Education, Family Size and Deposit Account

Conclusion and Business Recommendations

- Both Random Forest and XG Boost model can be used by the bank to predict customers who are likely to take Personal loan with an F1 score of 96%, this can help the bank in doing targeted marketing campaign.
- The models can be retrained using the complete data for achieving better performance before predicting on new customers.
- Customers with high Income and Credit Card spending have higher chances of conversion along with the ones having higher Education degrees and Family Size, so the bank can give special attention to these customers by offering them better deals like lower interest rates, increased limits, faster processing time, etc.
- Customers having a Deposit account with the bank also have higher chances of conversion, this could be because it helps increase their familiarity and trust on the bank. So for customers not having Deposit account, the bank can first offer them good deals on opening deposit accounts before approaching them for Personal loan.

Thank you.