Approach to this Bank load Classification problem

Data Preprocessing

1. Understanding the data: what is the data about, how is the data.
2. Checking if there are any unexpected values inside each column to make sure that while training no error pops up.
3. Dropped 2 columns which I thought would be unnecessary for my analysis which was ID and ZIP Code
4. Removing null values as it possesses the greatest threat to any ml project. The columns gender and home ownership contained 90% missing data with such high missing percentage I thought dropping those columns would be better rather than imputing.
5. The columns Income and Online had missing data less than 2% and they were missing completely at random which is why I choose CCA to handle this problem. Complete case analysis means to drop the null values below 5% which are missing randomly.
6. After dropping the null values from Income and Online I compared their Probability density function(pdf) and both pdfs' exactly stacked in top of each other, which means that the dropping of 2% data didn't hamper the distribution which is best for the analysis.
7. There was some preprocessing which needed to be done beforehand.
    a. The Experience column had some negative values which I replace with 0 as experience will always be a positive integer.
    b. The income column's value was not in the multiple of thousands as suggested in the data dictionary which is why I converted them by multiplying them with 1000 and converted the entire column to integer.
    c. The CCAvg column was also not in the multiple of 1000 of converted it as well.
    d. The personal Loan column had one row with empty value which I removed.
    e. Mortgage column also changed into 1000's
    f. The age columns had some unexpected value. The age was above 100 which in this current scenario is not possible which is why I dropped them as well. Having an age of 976 or 250 doesn't make sense.
8. Analyzing total income and total expenditure as per family size

Feature Engineering

1. Different hidden Columns were forged out like
    a. Income_montly by dividing the provided annual Income.
    b. Savings_monthly was forged out by subtracting the newly formed Income_monthly with the CCAvg column.
    c. The newly formed savings_monthly column had some negative values which means that those families didn't save any amount which is why I converted them to 0.
    d. The given CCAvg was monthly spending and from this I crafter CCAvg _yearly by multiplying it by 12.
    e. Savings_yearly was forged by subtracting annual income with newly formed yearly spendings (CCAvg_yearly)

f. Income_person_year was forged by dividing annual income with the number of family members.
g. Spending_person_year was forged by dividing yearly spending with the number of family members.

EDA

First of all, basic univariate analysis was performed to find some basic insights. Some of key findings are:

1. In the dataset family size of 1 member was the highest following with 2 3 and 4
2. Most of the families only had bachelor's degree, 2075 families to be exact followed by advanced degree with 1453 families and lastly master's degree with 1360.
3. Most families didn't have security accounts with the bank.
4. Most families didn't have CD accounts either.
5. The distribution of families with respect to online banking services were in the ratio of 60:40. 60% took services and 40% didn't.
6. Most families didn't take credit cards from the bank.
7. The acceptance of Loan offered in the campaign was very low. Only 368 families accepted.

Then multivariate analysis was conducted to find some valuable insights:

1. Most families who earned were most probably in the range of 20-60 years of old.
2. As the annual income grew the annual expenditure also increased. There it can be concluded that the spending is directly proportional to income.
3. Not only expenditure, the annual savings of the families also grew as income increased. So, savings are also directly proportional to income.
4. It was found that average annual income, annual spendings and annual savings, all three aspects were higher for the families with qualification level of Bachelor's in comparison to masters and above.
5. As well as the average income, average annual income, annual spendings and annual savings, all three aspects were higher for the families with lower family size. Family size of 1 i.e., Individual had highest income, expenditure, and savings than the rest.
6. Age and experience column had normal distribution.
7. The distribution of columns like Income, CCAvg, Savings_yearly, Income_person_year, CCAvg_yearly, Spending_person_year, Mortgage all were right skewed.

Observations

After all the necessary preprocessing and eda I split the data as 80:20 80 for training as 20 for testing. Then I checked in which random state would the Logistic regression's model perform the best, which came out to be 738 with a whopping 93% accuracy on test data.

For modelling I utilized 3 algorithms Logistic Regression, Random Forest Classifier and SVM all 3 lies inside Supervised learning as they require labeled data from where they learn and predict on new unseen data.

In logistic regression the test accuracy was 93 and the train accuracy was 91 which is one of the reasons that explains my model isn't overfitting. The cross Val cross was also constant, but precision recall and f1-score were really bad.

Random Forest classifier had accuracy of 90% for both train and test data and the precision recall and f1-score were also above 90% which generally considered as good model. Regarding overfitting, it's important to consider the precision, recall, and F1 score in addition to accuracy. The precision of 0.98 indicates that the model has a low false positive rate, while the recall of 0.88 suggests that it captures a good proportion of the positive instances. The F1 score of 0.93, which combines precision and recall, is also reasonably high.

Support vector machine had 94% accuracy on both test and train data. The cross Val score was also constant. Here, precision was 92% but recall and f1-score were very bad.

KNN also had the same accuracy on test and train as SVM, but it was slightly better as its precision, recall and f1-score were better in comparison to SVM but not too high to be considered as a good model.

Among the 4 supervised algorithms random forest performed well in every aspect.

Furthermore, I realized that there is a class imbalance in this data set. Almost 90% of data didn't accept the loan. To make the classes balance I also approached SMOTE (Synthetic Minority Over-sampling Technique) a famous technique to address imbalance classes in datasets. This helped in improving the existing models. The precision of Logistic Regression was higher and also slightly improved Random Forest classifier.

The accuracy, cross Val score, precision, recall and f1score of random forest before and after applying SMOTE was kind of similar but ultimately, I choose Random Forest Classifier after applying SMOTE as my model as the class was hugely imbalanced.