**Step 1: Business Problem Understanding**

A small, new bank called Universal Bank is looking for ways to convert an abundance of liability customers into personal loan customers, and they have collected a decent amount of records with various attributes. The goal of this case study is to utilize the provided records to see what attributes or combination of attributes would make someone more likely to accept a personal loan.

**Step 2: Data Understanding and Collection**

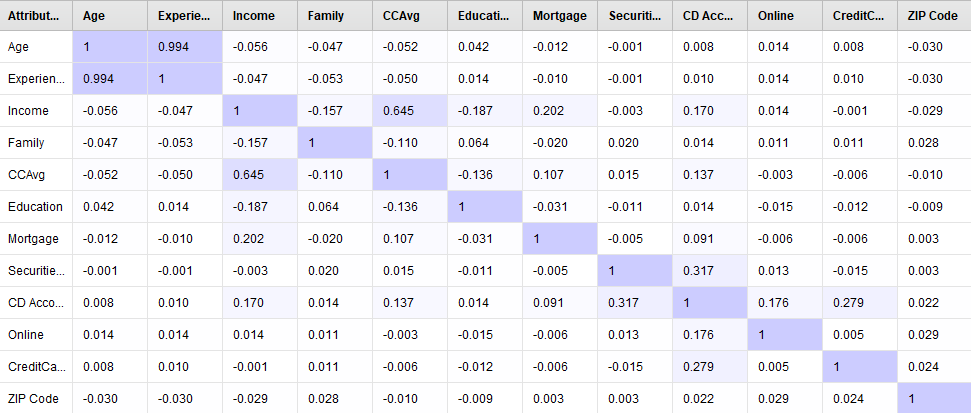
For this study, the bank has provided 5,000 records with 14 variables that include: ID, Age, Experience, Income, ZIPCode, Family, CCAvg, Education, Mortgage, Personal Loan, Securities Account, CD Account, Online, and CreditCard. All of the variables in the dataset are numerical; however, many of the variables like: Personal Loan, Securities Account, CD Account, Online, and CreditCard are intended to determine whether a variable is true or false rather than measure how much of a variable is present. ID is a variable intended to connect a customer to a record and is not intended to measure anything or influence results. The Personal Loan variable is the special attribute, and it is intended to measure whether a customer accepted a personal loan. More specifically, a couple attributes in the dataset are more complex and need to be explained like Family. This results for this attribute are intended to measure how large the customers family is, while the education attribute is recorded from 1-3 with 1 representing the customer has an undergraduate degree, 2 representing the customer has a graduate degree, and 3 representing the customer has an advanced degree.

**Step 3: Data Preparation and Feature Selection**

To prepare the data, I first ensured that there were no missing values by running the dataset and inspecting the statistics tab and missing column and noticed that there was no missing data. Then, I used the set role operator to make the ID variable an ‘ID’ in the records so RapidMiner would not try to do calculations with that variable that would falsely influence the results. Additionally, I searched for outliers by using the ‘detect outlier’ operator and then filtering the examples to keep all the useful data in the model. As a result, 10 rows of data were removed, and 4,990 records remained after the missing values and outliers were removed. Additionally, the data was rewritten to a new excel file that will be used in the model. Finally, I used a correlation matrix to see if any variables were correlated or measured similar things. I noticed that Age and Experience were highly correlated, but I do not believe they are identical, so I will leave them in the model for now. Additionally, income and CCAvg are relatively correlated, but the correlation is not strong enough to consider removing them either.

Diagram

Description automatically generated



**Step 4: Modeling Development**

The creation of this model required three processes in order to get results that were accurate and valuable to the goal of the study; but first, I am going to explain the problems with the first model that requires a separate process that created another dataset with an equal number of True and False records for the Personal Loan attribute. The first logistic regression model I created used the entirety of the clean data and the results were very unbalanced because the model had an abundance of records that were Personal Loan results were false, so the model had heavy bias to assume that almost everything was false. To fix this, I had to create a model that had an equal number of customers that accepted and denied the offer for Personal Loans. So, I started by taking two of the clean datasets and placing them separately in the process and each of the datasets was filtered using the filter example operators. One filter was made to include the True records and the other filtered in the false records. Then, the false set was sampled for 479, which is the number of positive records in the dataset. To finish this process, the two separately filtered datasets were combined into another data set with a total of 958 records that included an equal number of true and false records, and this data set was written into a new excel sheet that was then used to make the combined even data set. After the dataset was clean and even, I created another ROC process to see which type of model operator would produce the best results. Within the ROC operator I included a logistic regression, deep learning, and decision tree operator and ran the process. The decision tree produced the best results, and this was the operator I chose for the last process. Then, I created the logistic reasoning process that started by utilizing the select attributes operator to include the CCAvg, CD Account, CreditCard, Education, Family, ID, Income, Online, and Personal Loan variables since their p-values were under 0.05 and within acceptable. I then used the set role operator to set the role of the Personal Loan variable to a ‘label’ and the ID variable to an ‘ID’ in this separate process. I also included the numerical to binomial operator to convert the Personal Loan variable, so the logistical regression model would be able to function. Finally, the cross-validation operator is included and will perform 100 folds. The decision tree operator is on the training side of the cross-validation operator and the apply model and performance operator are on the testing side. The decision tree operator was also edited to change the criterion to information\_gain since it yielded the best results for the model.

Diagram

Description automatically generated

Chart

Description automatically generated

A picture containing diagram

Description automatically generated

Diagram

Description automatically generated

**Step 5: Model Evaluation and Interpretation**

The logistic regression model that was created in step 4 preformed with an accuracy of 96.26%, which is very reliable and means that the model is making correct decisions 96.26% of the time. The model has 22 False Negatives, 465 True Negatives, 14 False Positive, 457 True Positives which means that the model has a misclassification rate of 3.74%. This means that the model is very accurate and able to predict whether a customer is going to accept a personal loan with tremendous accuracy, and the odds of making a correct determination are dependent on the use of the model.

Table

Description automatically generated

A picture containing timeline

Description automatically generated

**Step 6: Model and Results Communication**

Ultimately, the model can make predictions with great accuracy with little probability of encountering errors with such a small misclassification rate. This proves that the presence of certain attributes that were selected have a large amount of influence in predicting whether a customer is willing to consider taking a personal loan. I believe the numerous processes that create the model form a dataset that includes clean and evenly distributed records that equally represent customers that accepted and refused offers on personal loans. As a result, I think the accuracy of the model is very nice and based on reliable data that can be trusted to make business predictions. The model’s accuracy of 96.26% is certainly an advantage, but I believe a disadvantage may be the fact the that results seem too good to be true and this could lead to some potential distrust in the model as well as potential complacency and dependency on the model. Thus, I believe the bank employees can utilize this model in order to predict whether a type of customer would be open to the idea of accepting a loan and market personal loans to customers with similar traits and backgrounds.

**Step 7: Model Deployment (Assess and Reflection)**

The study certainly has significant value to the Universal Bank and accomplishes their goal of seeing which combination of present attributes would make a certain customer more likely to accept a personal loan from their bank. I would recommend the bank use the model to create an advertising campaign that could reel in new and existing customers with loans that benefit both parties. Additionally, I would recommend the creation of similar models that could determine the likelihood of a customer accepting offers of other services the bank could offer. This model has a wide range of potential uses for Universal Bank if it is slightly modified the meet the new jobs needs, so I would recommend they recycle the model to continue making decisions about advertising other services. Based on the current accuracy of the model, I do not believe the model can be improved upon much more than it already has been, so the return on investment this model will provide for the company it with refined, efficient, and accurate advertising that earns the bank more personal loan opportunities that make them more money as well as repurposing the model for other studies and services.