PERSONAL LOAN MODELING IN BANKING

Classification Project

MARCH 26, 2023

CYRIL KOMBEM
McDANIEL COLLEGE

ABSTRACT

Loans are one of the most profitable sources in banking system and banks try their best to select reliable customers and offer them personal loans. However, customers can reject bank loan offers which is a wasted effort but if the right people get these offers it serves as good profit for the bank. Therefore, our aim of this study is to predict acceptance rates of the bank loan offers using the Random Forest algorithm. In this context, Random Forest was used to predict results with a grid search algorithm for better prediction and cross validation for much more reliable results. Research findings show that the best results were obtained with a kernel as 98.8% accuracy. Some precision and recall values are above average, like 0.98 and 0.88 which is great! This study recommends the use of Random Forest models in banking system for predicting acceptance of bank loan offers.

1. Personal Loan Modeling in Banking

1.1. Introduction

Nowadays with the growth in technology and analytics tools, companies need to find ways to improve their results without accumulating insignificant costs. Most companies analyze personal and transaction data to best understand customers' needs and make informed speculations of future needs. Most technological savvy companies will deliver insights from analytics to customers based on the context of interest. These techniques could be used strategically to increase sales by effectively targeting specific customers for specific products.

Therefore, in this project, we attempt to predict the success rates of introducing personal loan offers to liability customers. Just to clarify on our terminology, liabilities are items that the bank owes to someone else, this generally includes customer deposits, and debts from other institutions. But a liability customer is a person or an entity that owes the bank. This mostly involves customers who owe the bank.

This case is about a bank (Thera Bank) whose stakeholders wish to explore methods of converting its liability customers to personal loan customers (while retaining them as depositors). They had initially run a campaign on this goal and noticed an interesting healthy conversion success rate of over 9%. This result encouraged the marketing department for retail customers to build a better model to increase the conversion success ratio with a minimal budget. This is where we come in.

1.2. Literature Review

Banks are one of the oldest institutions I, our era. They have been around since the first currencies were minted and they helped wealthy people safeguard their money and precious items. Originally banks only made loans and issued notes for money deposited, therefore their primary role was ensuring that loans were given to the right people who would pay back and keep the business going. The challenge came in at the point where banks found it difficult to identify suitable candidates for personal loans who will not only accept loan offers but who in turn will save their money and safeguard their belongings. Nowadays, with the growth of the banks customer base, it has been a challenge identifying which customer is cable of loan repayment despite the odds. Most banks have been faced with this problem for a long time and rendering solution will go a long way to educate and enable these institutions to make informed business decisions and save lots of money from general advertisement and marketing.

One of the best ways forward will be using data science tools to build a model which can help the bank identify and target suitable customer from their massive portfolios of customers. That is where machine learning techniques come in. They are extremely beneficial in predicting outcomes when dealing with big data. If the model can predict those who are suitable for loans, this will go a long way to reduce manual work and boost income for the bank. In our work, we will build, train, and test multiple types of models to find a good fit our purpose. This will involve but not limited to logistic, decision trees (CART), gaussian naïve bayes, random forest classifiers, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) models just to make sure we use the best option based on their performance. We will build and test these models to get the best fit, but we also looked at similar work done and will mention their respective approaches below.

Amira Kamil Ibrahim Hassan, Ajith Abraham (2008) uses a prediction model which is constructed using three different training algorithms to train a supervised two-layer feedforward network. The results show that the training algorithm improves the design of loan default prediction model.

Kathe Rutika Pramod, Panhale Sakshi Dattatray, Avhad Pooja Prakash, Dapse Punam Laxman, and G horpade Dinesh B used decision tree algorithm in machine learning methods which efficiently performs both classification and regression tasks.

Kritika Pathak, Shazia Shaikh used the logistic regression model to make predictions on loan approvals.

Wanjun Wu applied the Random Forest and XGBoost algorithms to train the model and compare their performance in prediction accuracy. Random Forest is a decision tree based supervised learning algorithm, which implements the classification by constructing multiple decision trees. The metric we choose for splitting attributes in decision trees is the Gini index.

Hitesh K. Sharma, Tanupriya Choudhury, Prashant Ahlawat, Sachi N. Mohanty, and Sarika Jain used decision trees, Random Forest and the logistic models to determine the loan prediction accuracy required for the project.

1.3. Data Description

The file Bank.xls contains data on 5000 customers. The data include customer demographic information (ID, Age, Experience, Income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). We will talk more about these variables when we take a deep dive into data exploration.

Exploratory Data Analysis: We aim to explore the data's main aspects to easily target customers who will accept personal loan offers.

- Preparing the data to train a model: We will start by collecting data for 5000 customers of the bank and cleaning data for analysis using Python.
- Model training and making predictions using a classification model: Would conduct analysis using Python to improve model accuracy and precision.

- Model evaluation: The final stage will be to evaluate (test) the model and ensure it works in properly identifying potential candidates for personal loans who will have a positive likelihood of accepting the bank's offer.

There are no empty values in the dataset. The dataset has a blend of numerical and categorical attributes, but all categorical data are represented with numbers.

1.4. Data Attributes

- ID: Customer ID
- Age: Customer's age in completed years
- Experience: No. of years of professional experience
- Income: Annual income of the customer ($ 000)
- ZIP Code: Home Address Zip Code
- Family: Family size of the customer
- CCAvg: Avg. Spending on Credit Card per Month ($ 000)
- Education: Education Level. 1: Undergrad; 2: Graduate; 3: Advanced / Professional
- Mortgage: Value of house mortgage if any ($ 000)
- Personal Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities Account: Does the customer have a securities account with the bank?
- CD Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Does the customer use internet banking facilities?
- Credit Card: Does the customer use a credit card issued by this Bank?

1.5. Methodology

Exploratory Data Analysis: We aim to explore the data's main aspects to easily target customers who will accept personal loan offers.

- Preparing the data to train a model: We will start by collecting data for 5000 customers of the bank and cleaning data for analysis using Python.
- Training and making predictions using a classification model: Would conduct analysis using Python to improve model accuracy and precision.
- Model evaluation: The final stage will be to evaluate (test) the model and ensure it works in properly identifying potential candidates for personal loans who will have a positive likelihood of accepting the bank's offer.

Methodology chapter is a chapter where I will discuss on what kind of methods that I will use or apply for my research study. We will start off with importing data and proceed with exploratory data analysis. We will start by collecting data for 5000 customers of the bank and cleaning data for analysis using Python. We will import all the necessary packages to ensure our analysis run smoothly.

After setting up the platform we will explore the data's main characteristics, structure, and content. This will involve summary statistics, finding outliers such as negative values, missing values, and verifying for invalid values to make sure everything is unique, test for noise in the dataset.

After dissecting and inspecting the data we will decide on data cleaning measures if applicable to ensure that data is ready for analysis. We could verify the categorical attributes of the variables at hand to better understand the variables we are dealing with. At this point we will test for correlation between variables as we select our key variables (dependent and independent variables). It will be good to visualize the relationships between the variables to set a better visual and statistical leverage for selecting our dependent and independent variables. These observations will give us better ideas on where to lean towards as we work on building a good model.

When all has been put in place, there are various methods which we could go with building a good model for our prediction project. We could go with the classification models or the decision tree model. Splitting the data into training and test data will be key to test the final model's performance capability of targeting customers who will be eligible for personal loan offers. Going by the rule of thumb the training data is 70% leaving 30% to test the model performance.

Training the model and making predictions using the preferred model of our choice. We would conduct analysis using Python, if need be, to improve model accuracy and precision.

The final phase will be evaluating the model to ensure it functions properly in helping the bank better identify potential candidates for personal loans offers who will have a positive likelihood of accepting the bank's offer. We would be able to compare models to see which will have the most accurate prediction capability for our projects goal.

1.6. Objectives

In this project, we attempt to predict whether a personal loan offer to a liability customer (customer who has no debt engagement with the bank) for a Bank is likely to result in success. We believe that this research will go a long way to build stronger relationships between the bank and its customers, personal loan customers will positively impact the bank's profitability currently and in the long run.

2. Exploratory Data Analysis

2.1. Import and Load Libraries

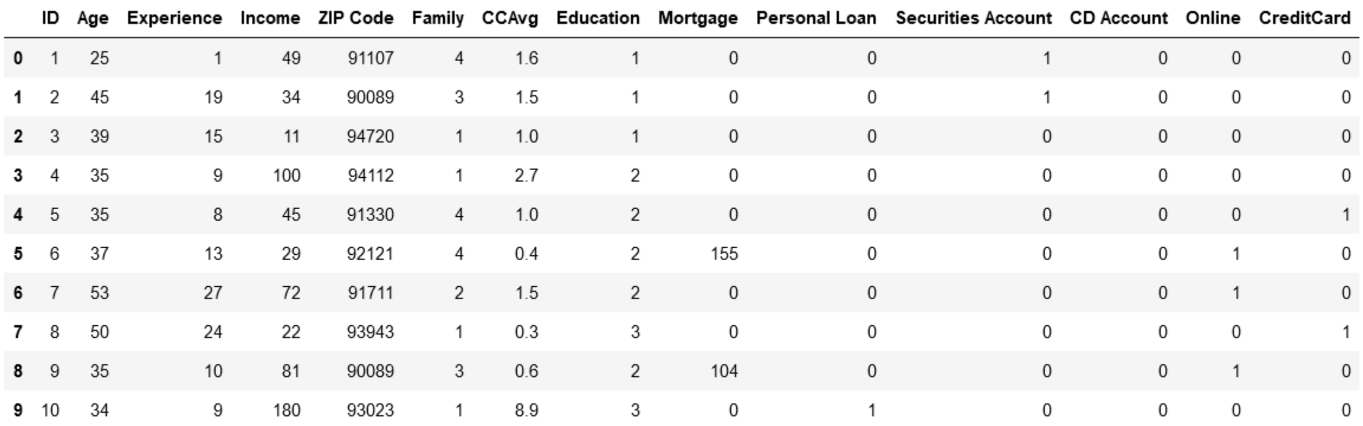
Data exploration is a pertinent phase in ensuring that our project is successful, but we will need to install some packages which will enable us to analyze the data. This will consist of built-in packages like json and third-party packages like pandas, seaborn, matplotlib.pyplot, sklearn, and plotly just to name a few.[1]

2.1. Import Data

[1] This are all carried out on python.

Read and visualize data as a data frame.

With a code in Python, we extracted data from the Excel CSV file into Jupyter notebook. We displayed the first 10 rows as illustrated in the snippet below.[2]



	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1
5	6	37	13	29	92121	4	0.4	2	155	0	0	0	1	0
6	7	53	27	72	91711	2	1.5	2	0	0	0	0	1	0
7	8	50	24	22	93943	1	0.3	3	0	0	0	0	0	1
8	9	35	10	81	90089	3	0.6	2	104	0	0	0	1	0
9	10	34	9	180	93023	1	8.9	3	0	1	0	0	0	0

2.2. Data Exploration

We started by exploring the shape of the data and got (5000, 14). The shape of the Data Frame is: (5000, 14), which means there are 5000 rows and 14 columns.[3]

```
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 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              5000 non-null   int64  
 4   ZIP Code            5000 non-null   int64  
 5   Family              5000 non-null   int64  
 6   CCAvg               5000 non-null   float64
 7   Education           5000 non-null   int64  
 8   Mortgage            5000 non-null   int64  
 9   Personal Loan       5000 non-null   int64  
 10  Securities Account  5000 non-null   int64  
 11  CD Account          5000 non-null   int64  
 12  Online              5000 non-null   int64  
 13  CreditCard          5000 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 547.0 KB
```

We have **13** integer columns as illustrated above, 1 **float** column out of which personal loan, securities account, CD account, online, credit card has **binary values**.

[2] See table 1 in the appendix.

[3] See figure 1 in the appendix.

Keep in mind that the column “**personal loan”** is our target variable (dependent variable y).

Information on the features or attributes
The attributes can be divided accordingly:

- The variable **ID** does not add any interesting information. *There is no association between a person's customer ID and loan, also it does not provide any general conclusion for future potential loan customers*. We can neglect this information for our model prediction.

The binary category has **five** variables as below:

- **Personal Loan** - Did this customer accept the personal loan offered in the last campaign? **This is our target variable.**
- **Securities Account** - Does the customer have a securities account with the bank?
- **CD Account** - Does the customer have a certificate of deposit (CD) account with the bank?
- **Online** - Does the customer use internet banking facilities?
- **Credit Card** - Does the customer use a credit card issued by Universal Bank?

Interval variables are as below:

- **Age** - Age of the customer
- **Experience** - Years of experience
- **Income** - Annual income in dollars
- **CCAvg** - Average credit card spending
- **Mortage** - Value of House Mortgage

Ordinal Categorical Variables are:

- **Family** - Family size of the customer
- **Education** - education level of the customer

The Nominal Variables are:

- **ID**
- **Zip Code**

Basic Data Summary Statistics

We went further to calculate the mean, standard deviation, average and other key variables of our dataset. This threw some insights on what type of data we are dealing with and what we need to adjust on. We notice that our **Experience** variable has a negative value. This is not logical. We will need to take note and make necessary adjustments before we begin analysis.

Identifying Unique Data

This step just helps the analyst identify what data points are repeated and which are unique. Only ID has 5000 unique columns. [4]

```
ID                    5000
Age                     45
Experience              47
Income                 162
ZIP Code               467
Family                   4
CCAvg                  108
Education                3
Mortgage               347
Personal Loan            2
Securities Account       2
CD Account               2
Online                   2
CreditCard               2
dtype: int64
```

Identifying Variable Relationships

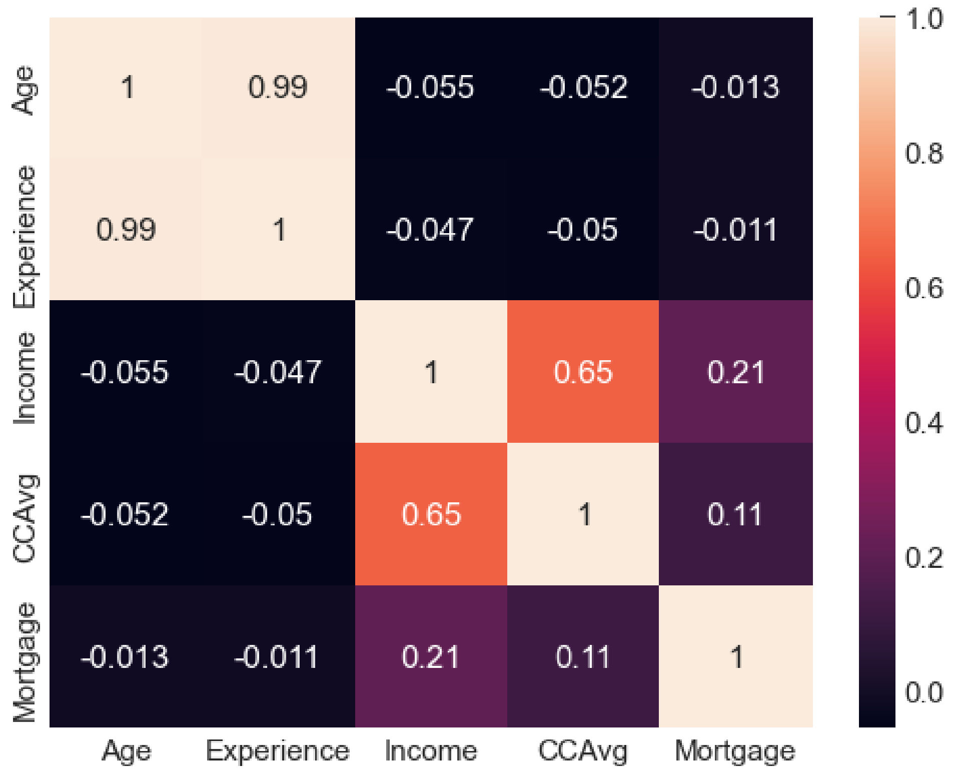
We decided to use the pair plot to understand the best set of features which would explain the **relationship between two variables** or those which have the most separated clusters. This method can be used to **form simple classification models** by drawing simple lines or make linear separation in our dataset.[5]

- Age features are normally distributed with majority of customers falling between 30 years and 60 years of age. We can confirm this by looking at the describe statement above, which shows mean is almost equal to median.
- Experience is normally distributed with more customer having experience starting from 8 years. Here the mean is equal to median. There are negative values in the Experience. This could be a data input error as in general it is not possible to measure negative years of experience. We can delete these values because we have 3 or 4 records from the sample.
- Income is positively skewed. Majority of the customers have income between 45K and 55K. We can confirm this by saying the mean is greater than the median.
- CCAvg is also a positively skewed variable and average spending is between 0K to 10K and majority spends less than 2.5K.
- Mortgage 70% of the individuals have a mortgage of less than 40K. However, the max value is 635K.
- From above pair plot we could decern that customers with higher income are more likely to accept Personal Loan offers.

To ensure we understood our data deeply and the respective correlation between them, we used the Heat Map to analyze just the interval variables to gain more insights.

[4] See figure 2 in the appendix.

[5] See figure 3 in the appendix.



Observation: By looking above plots we can see that **'Age' has very strong and positive association with 'Experience'**. I am also considering 'Education' to fix the negative experience error. Because as we know experience relates to the education level.

Decision: We can replace each negative 'Experience' value with the median of positive 'Experience' associated with the ‘Age’ and 'Education' value.

2.2 Data Preparation

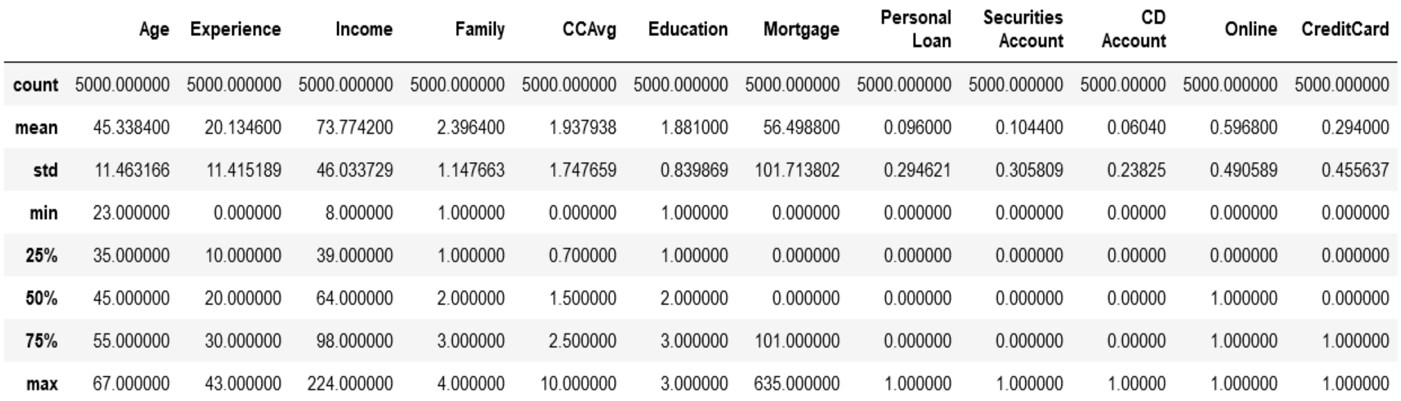
Dropping Unnecessary Data

- We see that a lot of categorical columns are being treated as integer datatypes. We'd like to convert them to categorical for our analysis.
- Finally, we'd like to drop 2 columns that we think are not relevant to the prediction: 'ID' and 'ZIP Code' from our analysis at this point.

Experience cannot be negative.

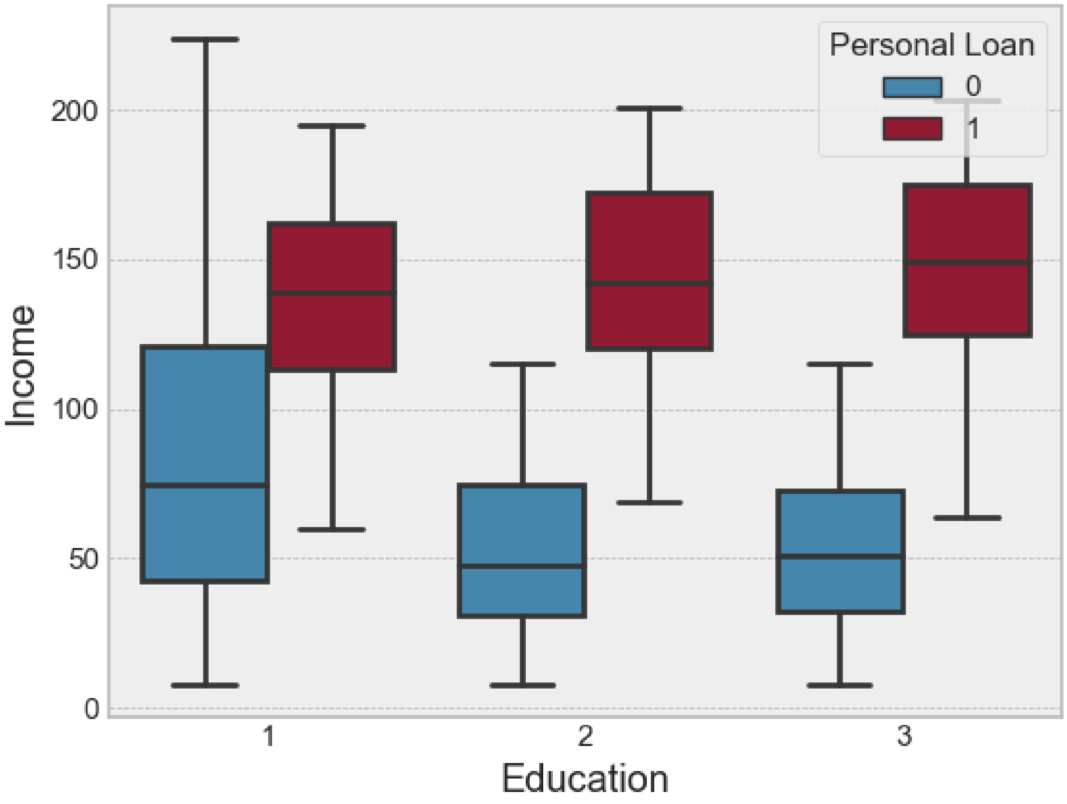
There are 52 records with negative experience. Negative experience does not make sense to our data, so we will need to clean data before proceeding. We displayed this data below. We will need to change these to positive values.

After adjusting the negative “Experience” values we can generate a realistic summary statistic of our variables. We can observe that experience does not have a negative mean as earlier indicated. Data seems ready for the next steps of exploration and analysis.



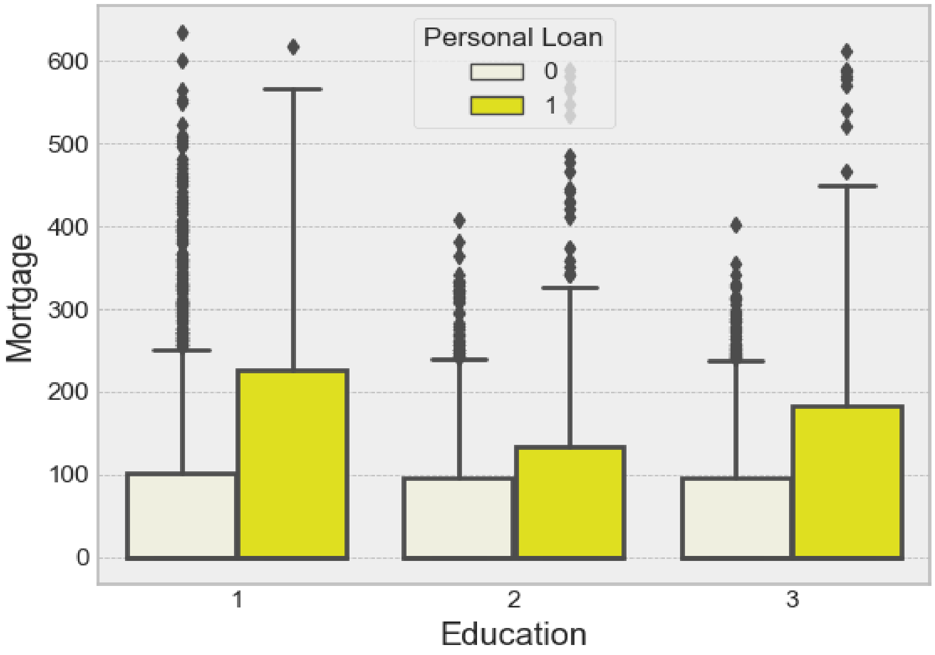
	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.00000	5000.000000	5000.000000
mean	45.338400	20.134600	73.774200	2.396400	1.937938	1.881000	56.498800	0.096000	0.104400	0.06040	0.596800	0.294000
std	11.463166	11.415189	46.033729	1.147663	1.747659	0.839869	101.713802	0.294621	0.305809	0.23825	0.490589	0.455637
min	23.000000	0.000000	8.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
25%	35.000000	10.000000	39.000000	1.000000	0.700000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
50%	45.000000	20.000000	64.000000	2.000000	1.500000	2.000000	0.000000	0.000000	0.000000	0.00000	1.000000	0.000000
75%	55.000000	30.000000	98.000000	3.000000	2.500000	3.000000	101.000000	0.000000	0.000000	0.00000	1.000000	1.000000
max	67.000000	43.000000	224.000000	4.000000	10.000000	3.000000	635.000000	1.000000	1.000000	1.00000	1.000000	1.000000

Influence of Income and Education variables on Personal Loan:



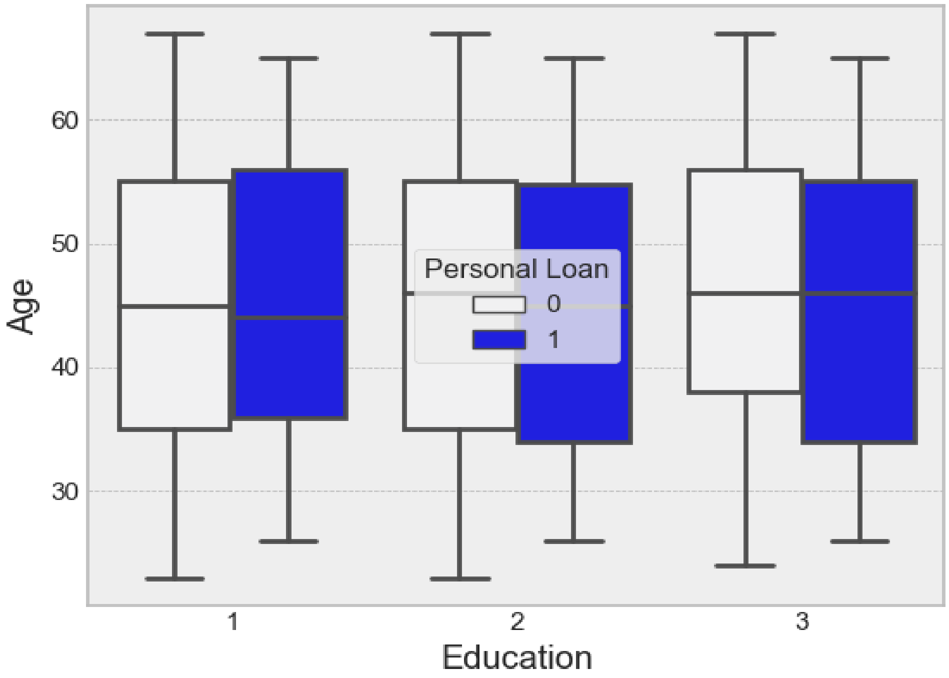
Observation: It seems the customers whose education level is 1 is having more income. However, customers who has taken the personal loan have similar income levels.

Influence of Education and Mortgage variables on Personal Loan:



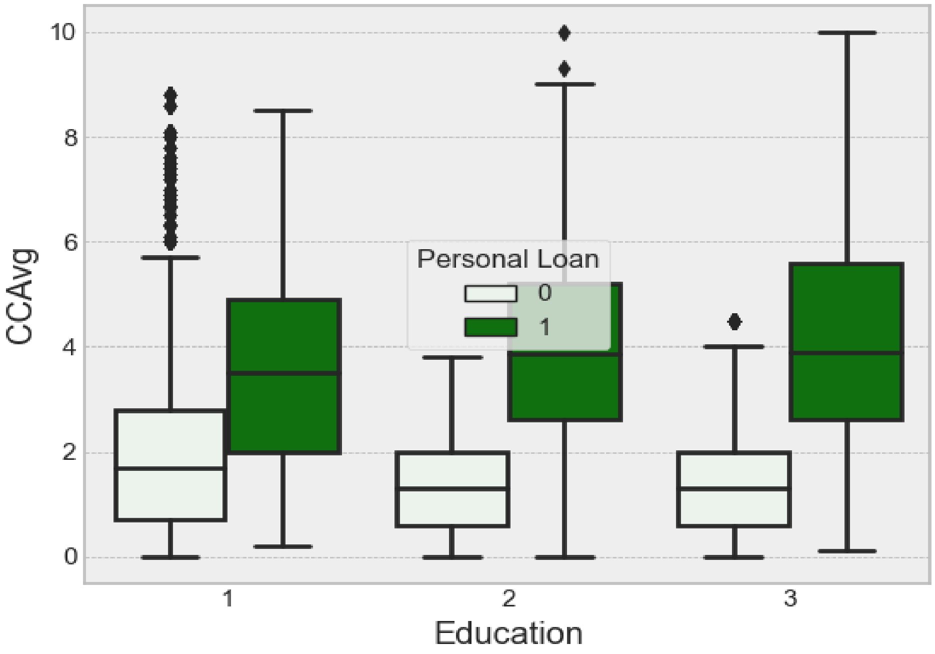
Observation: From the above chart it seems that customers who do not have personal loan and customers who currently have personal loans all have a high mortgage.

Influence of Education and Age variables on Personal Loan:

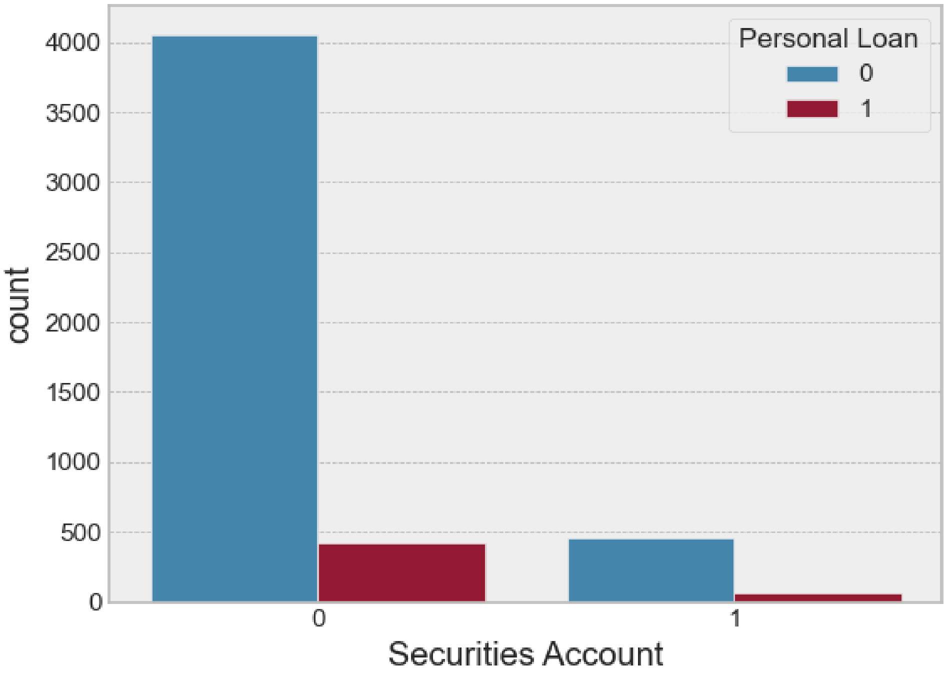


Observation: Younger customers have slightly more personal loans than older customers who are equally educated.

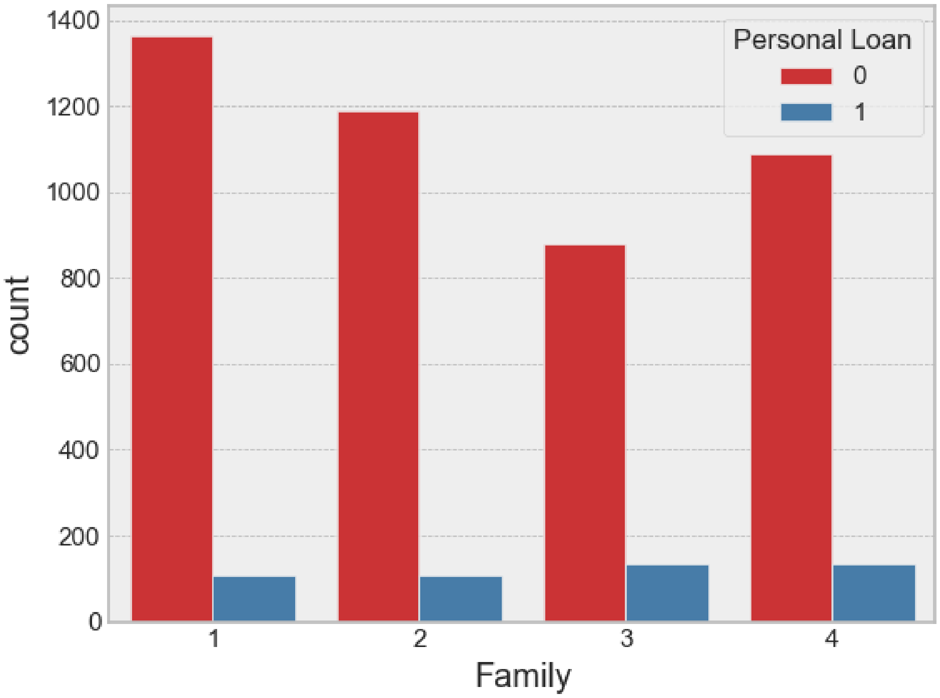
Influence of Education and Credit Card expenditure variables on Personal Loan:



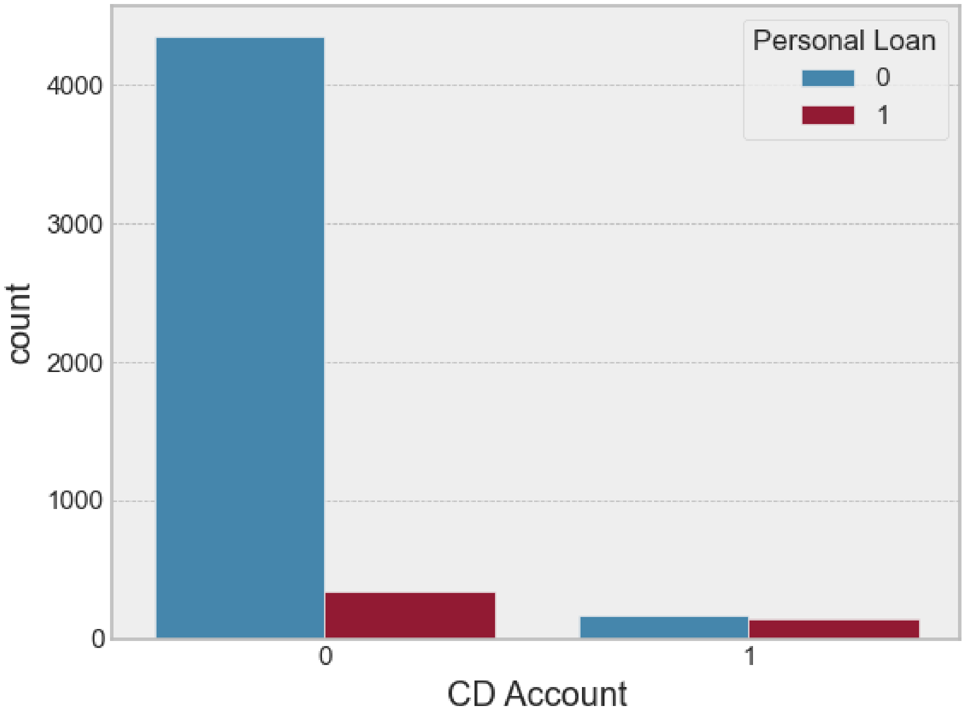
Observation: People with a higher average card monthly expenditure have personal loans. Customers without personal loans spend less on credit cards.



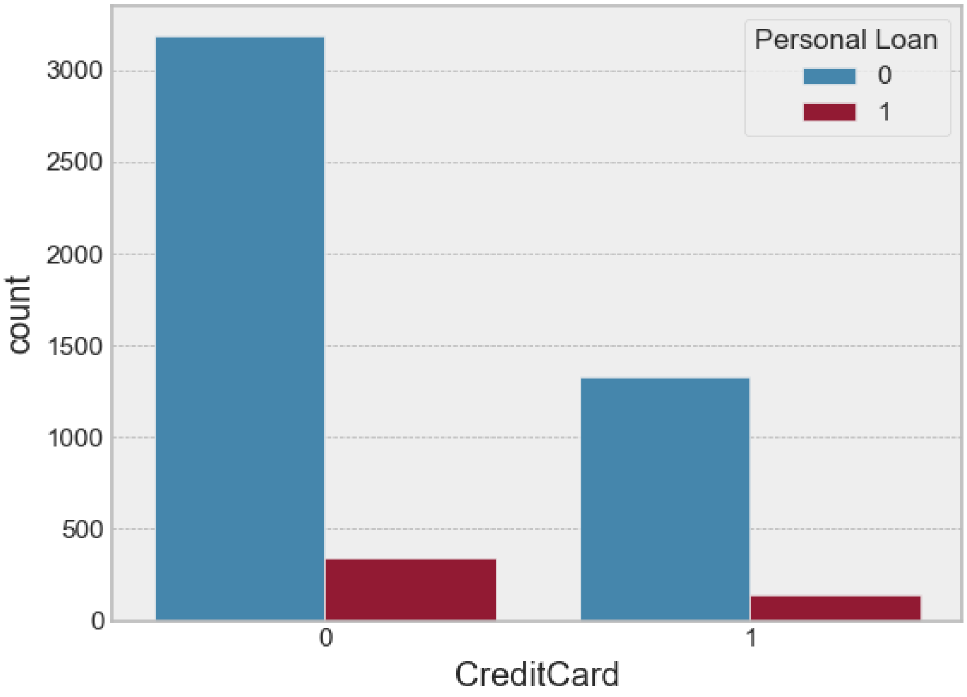
Observation: Majority of customers who does not have loan have securities account.



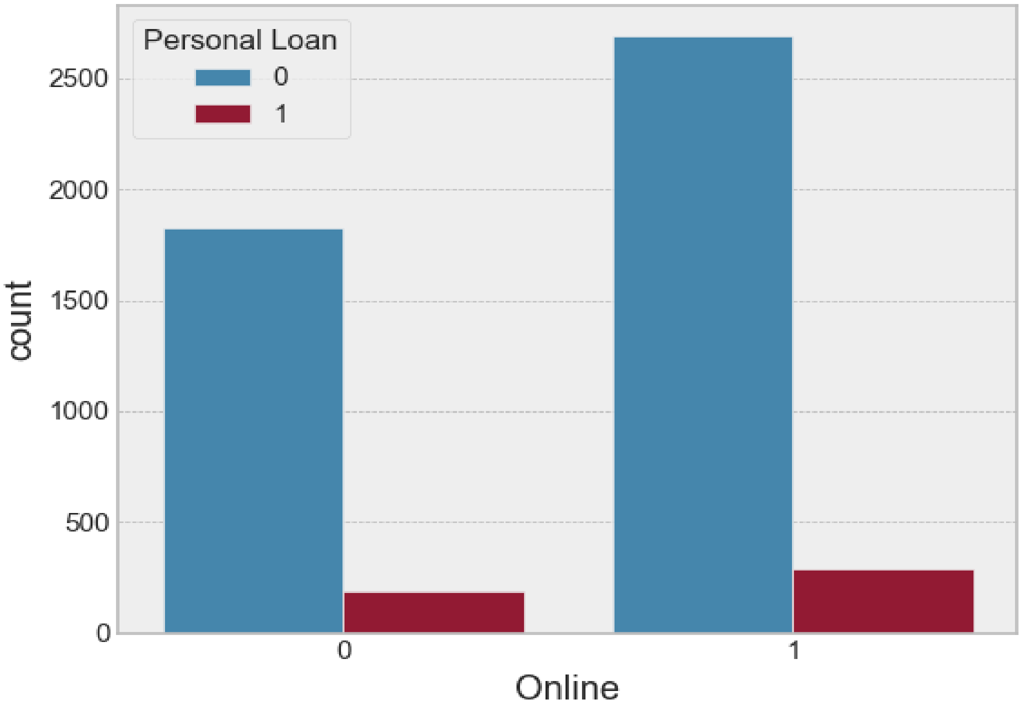
Observation: Family size does not impact the possibility of personal loan. But it seems families with size of 3 are more likely to take personal loans than a family of 1. When considering future campaign this might be good association.



Observation: Most customers who do not have CD account do not have personal loans as well. Almost all customers who has CD account have a personal loan as well.



Observation: Credit card usage compared to personal loans will have a relatively low importance in our analysis.



Observation: Customer online usage compared to personal loans will have a relatively low importance in our analysis.

3. Applying Models

First, we will need to specify the dependent variable **(y)** and the independent variable(s) **(X)**. y will represent **Personal Loans** which is our variable of interest and the other variables Age, Experience, Income, family, CCAvg, Education, Mortgage, Securities Account, CD Account, Online and Credit Card will represent the independent variables.

3.1. Split data into Train and Test

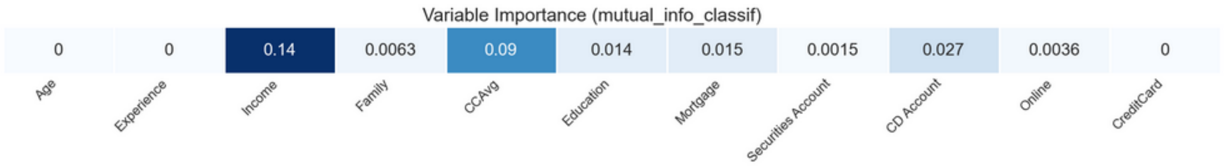
X = 'Age, Experience, Income. Family, CCAvg, Education, Mortgage, Securities Account, CD Account, Credit Card'
y = 'Personal Loan'

For this project, we started off choosing to go with the training feature are 70.0 % of dataset and training labels are 70.0 % of dataset, and the test feature is 30.0 % of dataset and test labels are 30.0 % of dataset. However, when we tried the 80/20 split, we got better results.

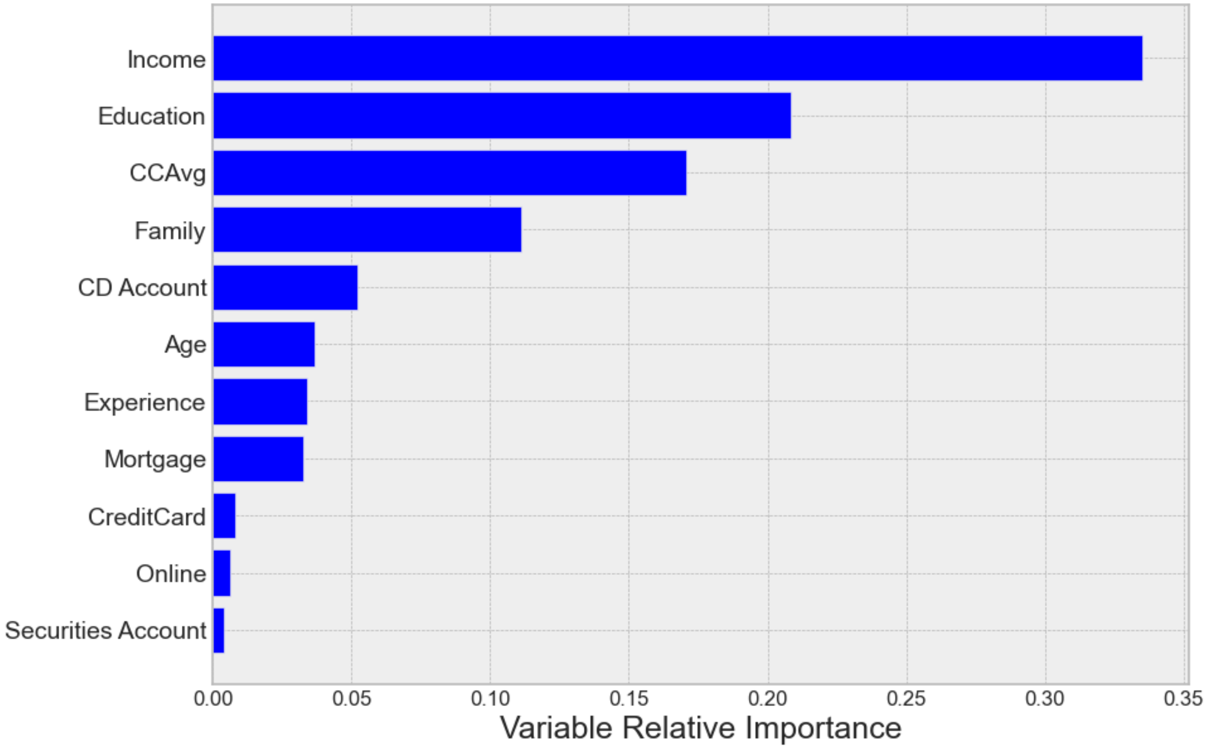
3.2. Verify Variable Importance and Relativity

We started off by splitting our data into the training and test data. we will need to specify the dependent variable (y) and the independent variable(s) (X). y will represent Personal Loans which is our variable of interest and the other variables Age, Experience, Income, family, CCAvg, Education, Mortgage, Securities Account, CD Account, Online and Credit Card will represent the independent variables. We went with splitting the train and test data using the 80/20 rule. The training data will be 80% and test data will be 20%. I believe this method works best for most models. I had initially set training data for 70% and testing data at 30% but after model adjustments and tuning it wasn’t as efficient as the 80/20 split.

The next step was to ensure that the independent variables used in our model are relevant to our analysis. To ensure that we use the most impactful variables for our analysis, we used the mutual information classifier to identify the most important variables and had the following feedback: income rated highest with 0.14, CCAvg followed with 0.09, CD Account with 0.027, Mortgage 0.015 and Education for 0.014. With these top five important matrices in mind, we noted that Age, Experience, and credit card were the least important variable with a metric of 0.



We used the random forest classifier to further clarify on the variables relative importance and ended up with income as the top variable followed by Education, CCAvg, Family and CD Account. The lowest relative importance from this analysis was recorded by owners of a Securities Account, Online and the Credit Card variable.



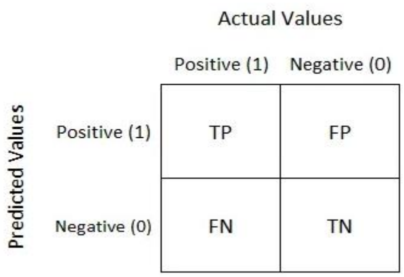
4.0. Results

This all leads to training our respective models and observing their performance. We had the intention to test out the following models.

- ✓ Gaussian Naive Bayes
- ✓ Logistic Regression
- ✓ K Nearest Neighbor
- ✓ SVM (Support Vector Machine)
- ✓ CART (Classification and Regression Tree)

✓ Random Forest

To be able to evaluate the performance of our models, we will be using the confusion matrix. The confusion matrix parameters such as accuracy, precision, recall and f1 are key in determining which model will be our best fit!



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Accuracy how many of them we have predicted correctly from all the classes (positive and negative).

Recall how many we predicted correctly from all the positive classes. Recall should be high as possible.

Recall = TP /TP + FN

Precision measures how many are actually positive from all the classes we have predicted as positive. Precision should be high as possible.

Precision = TP /TP + FP

To add more to the evaluation process, I will quote a 2018 article by Sarang Narkhede, “It is difficult to compare two models with low precision and high recall or vice versa. So, to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.” This just goes to reiterate on the importance of the F measure in selecting a model.

F1- Score = 2*Recall* Precision / Recall + Precision

Results from model training are analyzed below**;**

- Gaussian Naive Bayes
  - TP: 3406
  - FP: 214
  - FN: 184
  - TN: 196

  - Model Accuracy: 0.900
  - Model Precision: 0.478
  - Model Recall: 0.515
  - Model F1 measure: 0.496
- Logistic Regression
  - TP: 3586
  - FP: 34
  - FN: 120
  - TN: 260
  - Model Accuracy: 0.961
  - Model Precision: 0.884
  - Model Recall: 0.684
  - Model F1 measure: 0.771
- K Nearest Neighbor
  - TP: 3616
  - FP: 4
  - FN: 174
  - TN: 206
  - Model Accuracy: 0.955
  - Model Precision: 0.980
  - Model Recall: 0.542
  - Model F1 measure: 0.698
- SVM (Support Vector Machine)
  - TP: 3619
  - FP: 1
  - FN: 91
  - TN: 289
  - Model Accuracy: 0.977
  - Model Precision: 0.996
  - Model Recall: 0.760
  - Model F1 measure: 0.862
- CART (Classification and Regression Tree)
  - TP: 3591
  - FP: 29

  - FN: 33
  - TN: 347
  - Model Accuracy: 0.984
  - Model Precision: 0.922
  - Model Recall: 0.913
  - Model F1 measure: 0.917
- Random Forest
  - TP: 3515
  - FP: 5
  - FN: 43
  - TN: 337
  - Model Accuracy: 0.988
  - Model Precision: 0.985
  - Model Recall: 0.886
  - Model F1 measure: 0.933

- TP in the SVM model is better than others but TN is better in the CART (Decision tree) model.

- TP + TN in Gaussian Naive Bayes = 3602
- TP + TN in logistic Regression = 3846
- TP + TN in KNN = 3822
- TP + TN in SVM = 3908
- TP + TN in CART (Decision Tree) = 3938
- TP + TN in Random Forest = 3952
- Random forest was able to correctly predict more true numbers than the other models.
- If those who can take loan are not predicted correctly, that means the model is not properly trained and sending out loan offers to these customers will be costly.

- Comparation between key models

- Based on our results above we noticed that the Random Forest, SVM and KNN models stood out in the model testing phase.
- FP is 1 for the SVM model compared to the Random Forest model by 4 points. That means only four additional customers for the Random Forest model will be sent a personal loan invitation which will likely be rejected. That will not be a costly error.
- The lower the number of FP is the better, because a smaller number of customers who were eligible for loan were wrongly predicted and we said no to them.
- The Gaussian Naive Bayes, Logistic Regression, and CART models will be costly when it comes to FP. Therefore, these models are not suitable for this project.
- As a result, the Random Forest model with an accuracy score of 0.988, and an F1 score of 0.933 (which combines precision with recall) has been able to predict better than the other models with

a better cost function. But it is worthy to mention that the SVM model would have been the next best choice followed by the KNN model.

4.1. Conclusion

We started off hoping to build a model that can help a bank screen and select viable customers for the personal loan program. Since their goal was to target current customers who were likely susceptible to accepting these loans, we built a few models to deem which approach was most economical for the bank. We tested multiple models and our analysis resulted in advising the bank to use the Random Forest model with an accuracy score of 0.988, and an F1 score of 0.933.We should keep in mind that this model didn’t take lead in all the evaluation tests but it was the most economical for the bank.

We also did some model tuning and updated our model to test its capability to boost precision, recall and the F1-score. This was successful and interesting but we believe it may overfit the model thus not recommended for real life models.

4.2. References

X.Francis Jency, V.P.Sumathi, Janani Shiva Sri. (November 2018) “An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients” *International Journal of Recent Technology and Engineering. Volume-7.* pp. xxx-xxx.

Aurelien, Geron. (June 2019) “Hands on Machine Learning with Scikit Learn Keras, and Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems” Second Edition, OReilly Media 2019. pp. 87-189

Wanjun, Wu. (September 2022) “Machine Learning Approaches to Predict Loan Default” Intelligent Information Management. pp. 14, 157-164.

Hitesh K. Sharma, Tanupriya Choudhury, Prashant Ahlawat, Sachi N. Mohanty, Sarika Jain. (May 2019) “Machine Learning based model for Loan Amount Prediction and Distribution”. pp 210-218.

Kritika Pathak, Shazia Shaikh. (September 2021) “Loan Approval Prediction Using Machine Learning”. *International Research Journal of Engineering and Technology (IRJET).* pp. 1897-1900.

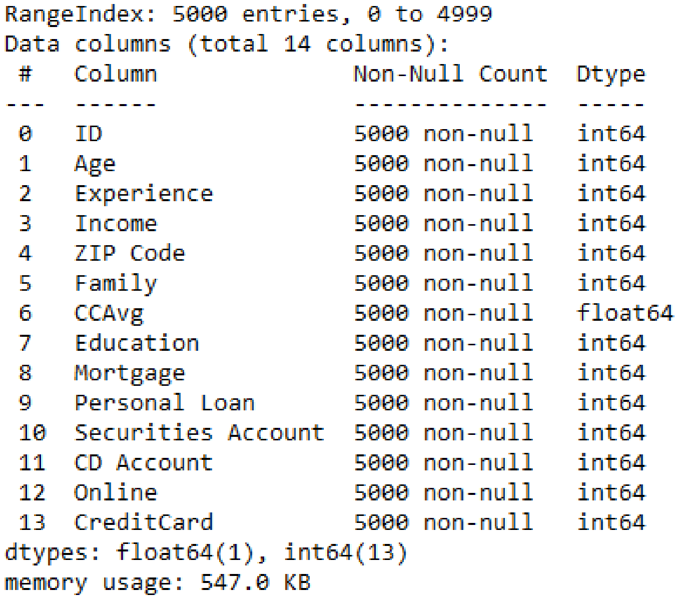
Kathe Rutika Pramod, Panhale Sakshi Dattatray, Avhad Pooja Prakash, Dapse Punam Laxman, G horpade Dinesh B. (June 2021) “An Approach For Prediction Of Loan Approval Using Machine Learning Algorithm”. *International Journal of Creative Research Thoughts (IJCRT). pp. 568-570.*

Mehmet Furkan, Akça and Onur, Sevli, . (August 2022) “Predicting acceptance of the bank loan offers by using support vector machines” *International Advanced Researches and Engineering Journal 06(02): 142-147, 2022. Volume-6. pp. 142-147.*

L. Udaya Bhanu1, Dr. S. Narayana. (June 2021) “Customer Loan Prediction Using Supervised Learning Technique”*, International Journal of Scientific and Research Publications, Volume 11, Issue 6. pp. 403-407*

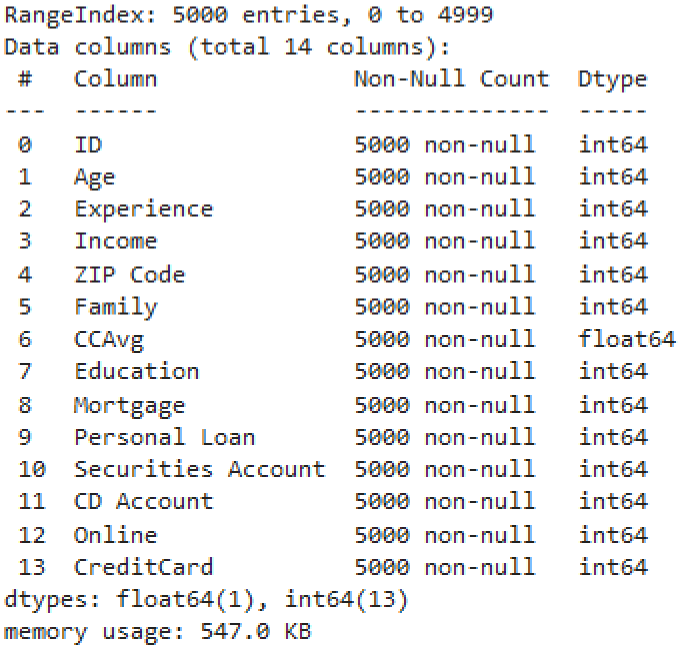
4.3. Appendix: Results

Figure 1:



```
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 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              5000 non-null   int64  
 4   ZIP Code            5000 non-null   int64  
 5   Family              5000 non-null   int64  
 6   CCAvg               5000 non-null   float64
 7   Education           5000 non-null   int64  
 8   Mortgage            5000 non-null   int64  
 9   Personal Loan       5000 non-null   int64  
 10  Securities Account  5000 non-null   int64  
 11  CD Account          5000 non-null   int64  
 12  Online              5000 non-null   int64  
 13  CreditCard          5000 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 547.0 KB
```

Figure 2:



```
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 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              5000 non-null   int64  
 4   ZIP Code            5000 non-null   int64  
 5   Family              5000 non-null   int64  
 6   CCAvg               5000 non-null   float64
 7   Education           5000 non-null   int64  
 8   Mortgage            5000 non-null   int64  
 9   Personal Loan       5000 non-null   int64  
 10  Securities Account  5000 non-null   int64  
 11  CD Account          5000 non-null   int64  
 12  Online              5000 non-null   int64  
 13  CreditCard          5000 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 547.0 KB
```

Figure 3:

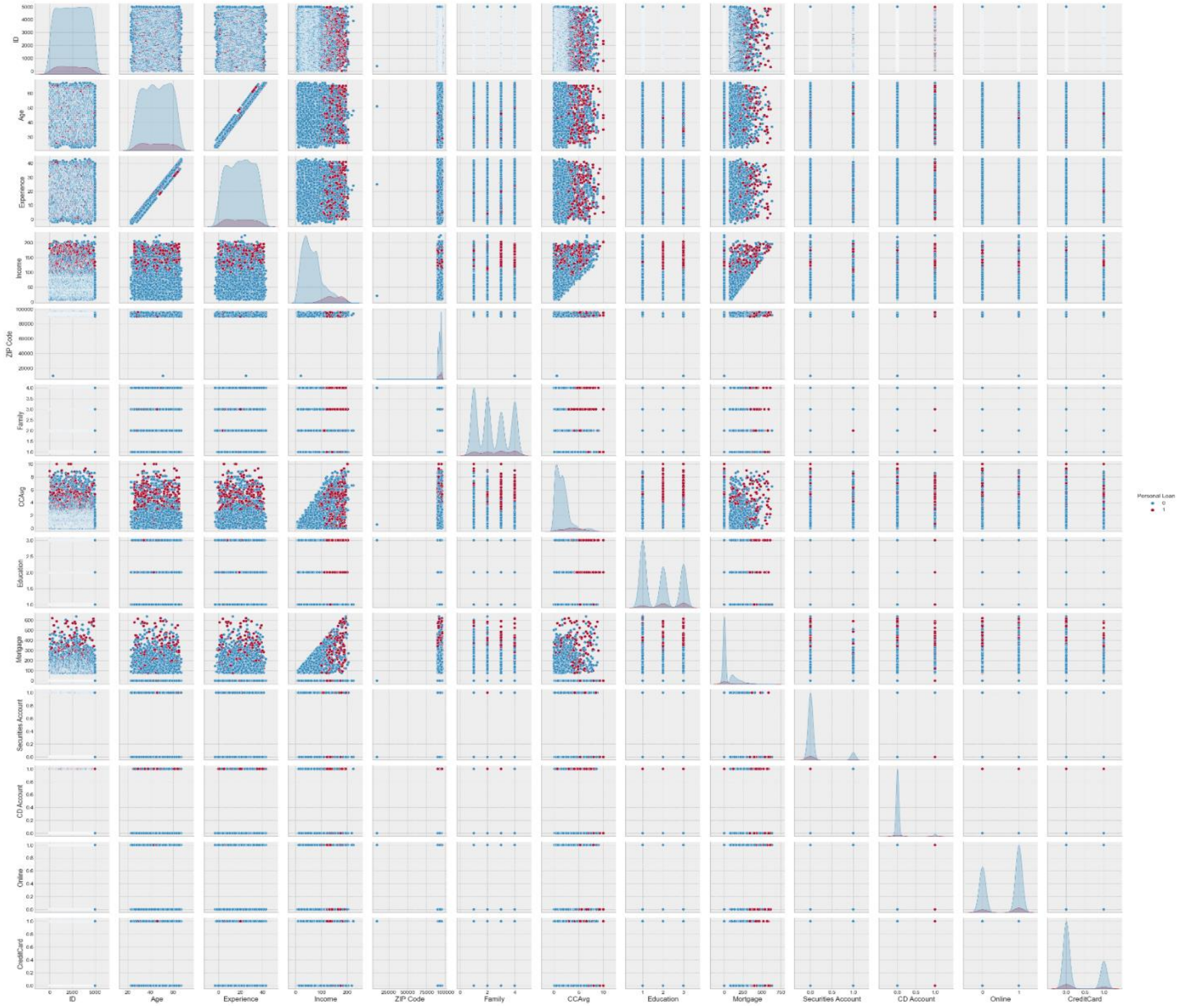


Figure 3:

Table 1:

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1
5	6	37	13	29	92121	4	0.4	2	155	0	0	0	1	0
6	7	53	27	72	91711	2	1.5	2	0	0	0	0	1	0
7	8	50	24	22	93943	1	0.3	3	0	0	0	0	0	1
8	9	35	10	81	90089	3	0.6	2	104	0	0	0	1	0
9	10	34	9	180	93023	1	8.9	3	0	1	0	0	0	0

Table 2:

	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.338400	11.463166	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIP Code	5000.0	93152.503000	2121.852197	9307.0	91911.00	93437.0	94608.00	96651.0
Family	5000.0	2.396400	1.147663	1.0	1.00	2.0	3.00	4.0
CCAvg	5000.0	1.937938	1.747659	0.0	0.70	1.5	2.50	10.0
Education	5000.0	1.881000	0.839869	1.0	1.00	2.0	3.00	3.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.00	0.0	101.00	635.0
Personal Loan	5000.0	0.096000	0.294621	0.0	0.00	0.0	0.00	1.0
Securities Account	5000.0	0.104400	0.305809	0.0	0.00	0.0	0.00	1.0
CD Account	5000.0	0.060400	0.238250	0.0	0.00	0.0	0.00	1.0
Online	5000.0	0.596800	0.490589	0.0	0.00	1.0	1.00	1.0
CreditCard	5000.0	0.294000	0.455637	0.0	0.00	0.0	1.00	1.0

Table 3:

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1
5	6	37	13	29	92121	4	0.4	2	155	0	0	0	1	0
6	7	53	27	72	91711	2	1.5	2	0	0	0	0	1	0
7	8	50	24	22	93943	1	0.3	3	0	0	0	0	0	1
8	9	35	10	81	90089	3	0.6	2	104	0	0	0	1	0
9	10	34	9	180	93023	1	8.9	3	0	1	0	0	0	0

Table 2:

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
89	25	-1	113	4	2.30	3	0	0	0	0	0	1
226	24	-1	39	2	1.70	2	0	0	0	0	0	0
315	24	-2	51	3	0.30	3	0	0	0	0	1	0
451	28	-2	48	2	1.75	3	89	0	0	0	1	0
524	24	-1	75	4	0.20	1	0	0	0	0	1	0
536	25	-1	43	3	2.40	2	176	0	0	0	1	0

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
540	25	-1	109	4	2.30	3	314	0	0	0	1	0
576	25	-1	48	3	0.30	3	0	0	0	0	0	1
583	24	-1	38	2	1.70	2	0	0	0	0	1	0
597	24	-2	125	2	7.20	1	0	0	1	0	0	1
649	25	-1	82	4	2.10	3	0	0	0	0	1	0
670	23	-1	61	4	2.60	1	239	0	0	0	1	0
686	24	-1	38	4	0.60	2	0	0	0	0	1	0
793	24	-2	150	2	2.00	1	0	0	0	0	1	0
889	24	-2	82	2	1.60	3	0	0	0	0	1	1
909	23	-1	149	1	6.33	1	305	0	0	0	0	1
1173	24	-1	35	2	1.70	2	0	0	0	0	0	0
1428	25	-1	21	4	0.40	1	90	0	0	0	1	0
1522	25	-1	101	4	2.30	3	256	0	0	0	0	1
1905	25	-1	112	2	2.00	1	241	0	0	0	1	0
2102	25	-1	81	2	1.60	3	0	0	0	0	1	1
2430	23	-1	73	4	2.60	1	0	0	0	0	1	0
2466	24	-2	80	2	1.60	3	0	0	0	0	1	0
2545	25	-1	39	3	2.40	2	0	0	0	0	1	0

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
2618	23	-3	55	3	2.40	2	145	0	0	0	1	0
2717	23	-2	45	4	0.60	2	0	0	0	0	1	1
2848	24	-1	78	2	1.80	2	0	0	0	0	0	0
2876	24	-2	80	2	1.60	3	238	0	0	0	0	0
2962	23	-2	81	2	1.80	2	0	0	0	0	0	0
2980	25	-1	53	3	2.40	2	0	0	0	0	0	0
3076	29	-1	62	2	1.75	3	0	0	0	0	0	1
3130	23	-2	82	2	1.80	2	0	0	1	0	0	1
3157	23	-1	13	4	1.00	1	84	0	0	0	1	0
3279	26	-1	44	1	2.00	2	0	0	0	0	0	0
3284	25	-1	101	4	2.10	3	0	0	0	0	0	1
3292	25	-1	13	4	0.40	1	0	0	1	0	0	0
3394	25	-1	113	4	2.10	3	0	0	0	0	1	0
3425	23	-1	12	4	1.00	1	90	0	0	0	1	0
3626	24	-3	28	4	1.00	3	0	0	0	0	0	0
3796	24	-2	50	3	2.40	2	0	0	1	0	0	0
3824	23	-1	12	4	1.00	1	0	0	1	0	0	1
3887	24	-2	118	2	7.20	1	0	0	1	0	1	0

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
3946	25	-1	40	3	2.40	2	0	0	0	0	1	0
4015	25	-1	139	2	2.00	1	0	0	0	0	0	1
4088	29	-1	71	2	1.75	3	0	0	0	0	0	0
4116	24	-2	135	2	7.20	1	0	0	0	0	1	0
4285	23	-3	149	2	7.20	1	0	0	0	0	1	0
4411	23	-2	75	2	1.80	2	0	0	0	0	1	1
4481	25	-2	35	4	1.00	3	0	0	0	0	1	0
4514	24	-3	41	4	1.00	3	0	0	0	0	1	0
4582	25	-1	69	3	0.30	3	0	0	0	0	1	0
4957	29	-1	50	2	1.75	3	0	0	0	0	0	1