A Predictive Model for Bank Loans

Martize T. Smith

Introduction

The business financial sector is incredibly diverse. There are multiple businesses entering this market and new products and services being created for the purpose of advancing clients and customers life while also being profitable. The market for basic products and services such as personal loans are highly competitive. With the competition in the personal loan markets being fierce, businesses are using every strategy and technique they can to conduct business in the most efficient and best manner possible. One of many ways to help grow a business enterprise is leveraging data to extract value then translate that into actions for driving business activities that give an expected result on average in a certain area. In this case, a bank would utilize data science methods to build a predictive model to help identify the potential customers who have a higher probability of purchasing a personal loan. Statistics from LendingTree show that 23.2 million Americans owe a combined $241 billion in personal loans, which is more than double the $117 billion owed in 2017. A different statistic from Forbes tells us that a Forbes Advisor survey found that 68% of Americans have applied for a personal loan in the past 12 months from May 2023. One more statistic, from US News, showcases the average personal loan debt per borrower is $11,692, which is up from about $8,780 before the pandemic that started in year of 2020. Statistics and case studies show there is a strong market for personal loans, so with this, businesses are using data to help guide their business decisions. I analyze a dataset of business customers from a bank to uncover what factors contribute the most to a customer deciding to buy a personal loan then create a random forest predictive model to predict what future customers have a higher probability of buying the personal loan.

Audience

This project report is for both technical and non-technical professionals because it helps advance the field of data science regardless of who decides to read it. The world of researchers, scientists, innovators, data experts and other professionals we learn and grow from each other’s work and contributions to subject matter expertise. From this expertise, we can better think critically and move our projects or ambitions forward in a direction we think will help us get the most out of what we're choosing to accomplish. This body of work present here uses a general data science framework to observe, analyze then create a predictive model, from a bank's customer data, to enhance the performance yield to better identify potential customers who have a higher probability of purchasing a loan. One scenario is that other data professionals may seek greater intelligence and application of data science methods in this area of predictive modeling. Another scenario may involve a Chief Executive Officer (CEO) or Chief Marketing Officer (CMO) investigating independently their own broad overview of how data science and analytics can help their business enterprises grow or relay this report or concept to their data science teams. Altogether, technical and non-technical personnel benefit from this report.

Data Source

The data utilized in this report was downloaded from Kaggle.com. Kaggle is a global leader in the data science community with substantial amounts of data science resources. Data leaders often visit and learn from this body of knowledge and resources to keep their mind and analytical skills sharp and active. The data consists of a bank's customer's information and more. This dataset has 5000 observations. This bank's data will be used to formulate a predictive model for the identification of the potential customers who have a high likelihood of buying loan products. Ultimately, this data source is in alignment with implementing data science methodologies within the marketing and customer segmentation context to learn and apply the best ways to attain superior business results overall.

Data

All data is from one business and it consists of 5000 observations with 14 columns which are the following:

- ID
- Age
- Experience
- Income

- ZIP Code
- Family
- CCAvg
- Education
- Mortgage
- Personal Loan
- Securities Account
- CD Account
- Online
- CreditCard

This data was taken through the data cleaning phase in which I checked for missing values, improper data types, duplicate values and misspelled words. I concentrated on “Personal Loan: Did the person accept the loan in the last campaign” as my dependent variable, which will be represented as “Y”. I choose five independent variables for this project and they are represented as “X”. The five independent variables have high correlations to customers deciding to purchase the loan products.

Method

This Kaggle dataset is customer segmentation information from a bank. The final goal is to determine what factors propel the performance yield to skillfully identify potential customers that will have the most likelihood of buying a personal loan through developing a random forest predictive model. I applied the cross industry standard process for data mining (CRISP-DM) to express an overview of the methods and procedures current in this report, which is as follow: (1) investigate and select independent and dependent variables (2) Visualize & analyze the data (3) Identify high correlations among variables (4) Build random forest model (5) Evaluate the Model. As I focused on the bank’s data, the dependent variable I choose is “Did this customer accept the personal loan offered in the last campaign?”. Furthermore, I choose five independent variables, which are Income, average spending on credit cards per month, does the customer have a certificate of deposit (CD) account with the bank?, value of house mortgage if any and Education Level 1: Undergrad; 2: Graduate; 3: Advanced/Professional because they had the highest correlation to what customers decided to buy the loan relative to other contributing factors for selecting these independent variables. Once choosing the variables and undergoing data cleaning and exploration the next phase was model formulation. In data science and related subjects there are numerous ways to build predictive models. Here are some typical predictive techniques:

Classification Model:

Classifications are supervised machine learning techniques and the model attempts to predict the accurate identification of provided inputted data. During classification, the model is completely trained using the training data. Next, the model is assessed on test data before being utilized to perform prediction on fresh invisible data.

Clustering Model:

A clustering model divides data into separate categories based on similar characteristics. Next, the model utilizes the data from each group to distinguish enormous outcomes for each cluster. Furthermore, this model functions by implementing two common types of clustering. One type is hard clustering, which categorizes data by deciding whether each point entirely fits to a certain cluster. The second type is soft clustering, which pinpoints a probability to each data point verses dividing them into specific clusters.

Decision Tree:

A decision tree is an algorithm that diagrams various sources of data into a tree-like framework to illustrate the possible scenarios of output of diverse decisions. The decision tree displays diverse decisions into branches and then expresses possible scenarios underneath each decision. Businesses and organizations generally apply decision trees to govern the most critical variables in a specific dataset.

I decided to select the random forest technique for this project and report. In data science, the random forest method can be used and act as both a classification and regression technique during machine learning.

Data Wrangling & Cleaning

With the Kaggle dataset I did basic data cleaning, which is listed below.

Checked for Duplicates

It can be common for data from all sizes to be likely to have duplicated values. Duplicates can sometimes come from human mistakes when human is entering the data or completing a form while making a mistake in the progress. Mistakes can also come from computers as data is being pulled from one source then put somewhere else.

Checked for Wrong Data Types

It is important to make sure data types are correct so when moving along the data science process that exploration and analysis is properly conducted. Generally, many numbers are the common data type that will have to be converted as data cleaning takes place. Frequently numbers are entered or identified as text but for data processing to take place these numbers have to appear as numerals.

Checked for Missing Values

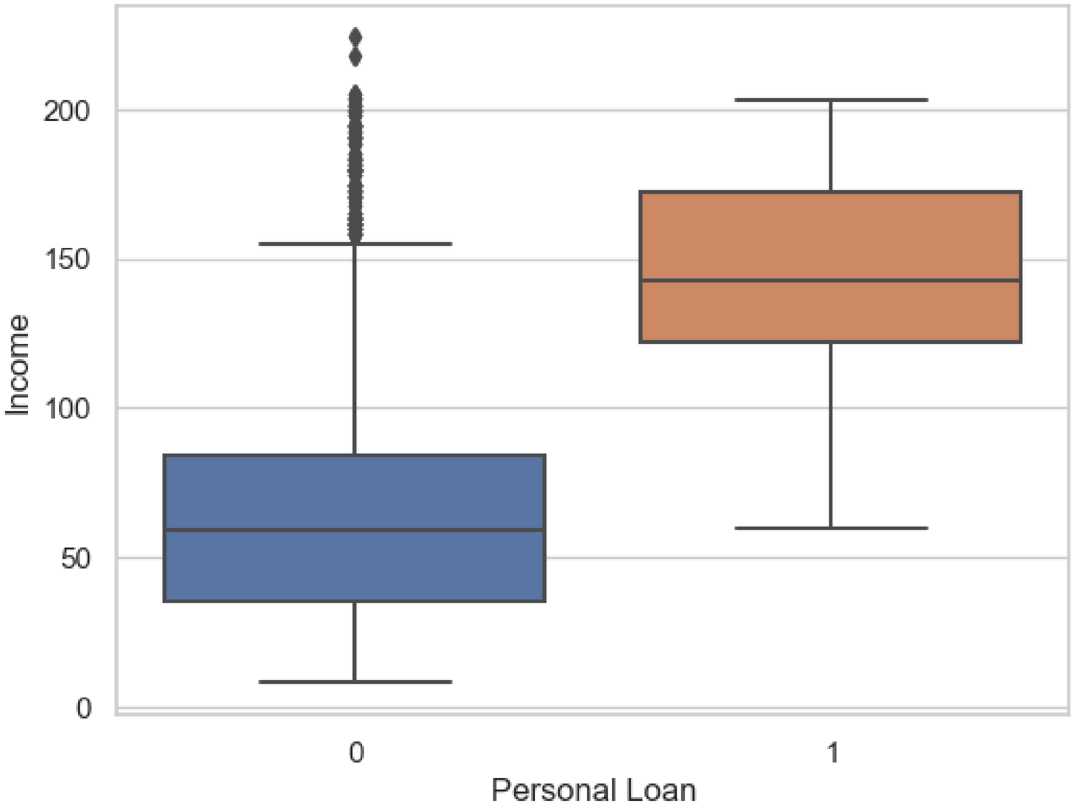
When observing datasets missing values are another common obstacle seen and must be dealt with. Two ways to handle missing values is to deleted them or replace the missing values. The right decision is based on what the overall project or purpose of the data science application is.

Checked for Irrelevant Data

Irrelevant data is sometimes in the dataset. Irrelevant data can give inaccurate or incomplete analysis. When examining data, the data professionals must know what is relevant and what is not before advancing to the data cleaning phase. An example is if you are investigating the age range of your customers, you would not need to add their email addresses.

Exploratory Data Analysis (EDA)

During the EDA phase I utilized summary statistics of the entire dataset and then narrowed it down to just showing summary statistics of all five of the independent variables. Additionally, I explored and visualized correlations, through a heatmap, from my dependent variable and independent variables to further understand data relationships. Box-plots are shown as illustrations of the relationships of the dependent variable and just the two independent variables with the highest correlations to customers choosing to buy loan from bank, income and average spending on credit cards per month. A standard bar plot is expressed that captures the education levels of customers who choose to get a loan verses the ones that choose to not get one. Lastly, a combination bar and line graph are utilized to detail the insights and relationship between all five independent variables with the highest correlations to the dependent variable. There are many visuals so I only display the most relevant as figures which are labeled below.

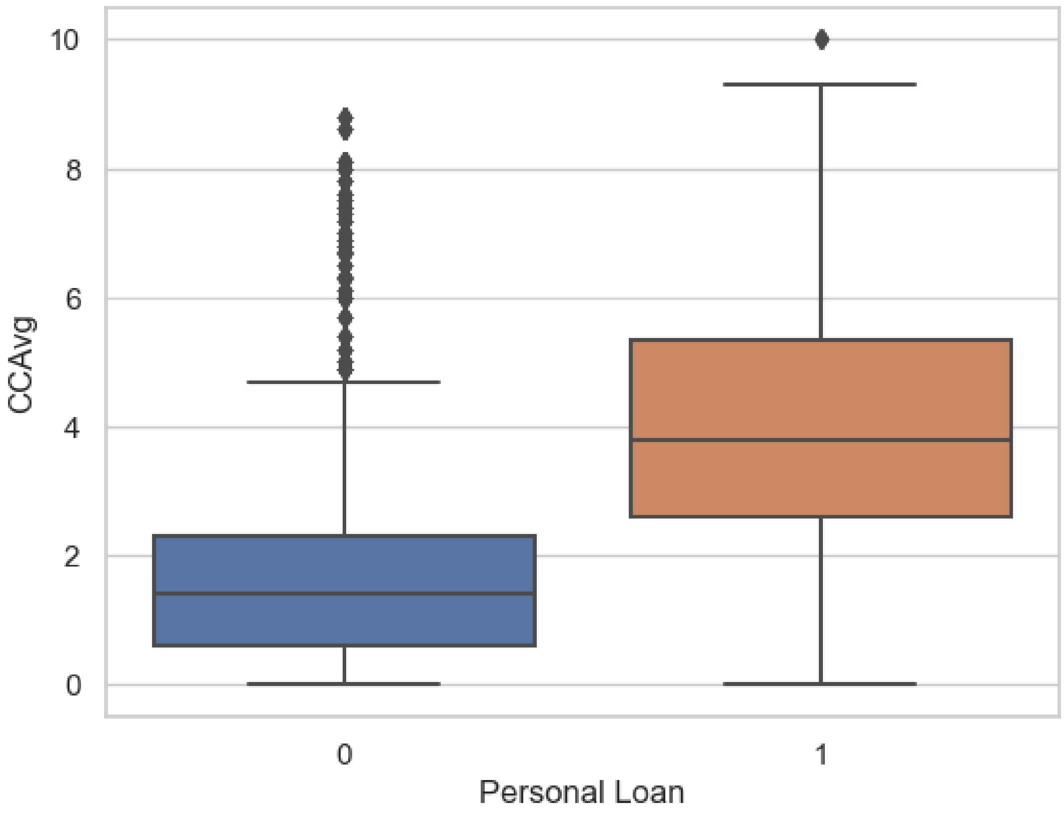


[1]

Figure one is box-plot visual of the bank's customer's income compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan). Figure one demonstrates that the customers with

[1] Box-Plot visual of the bank's customer's income compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan)

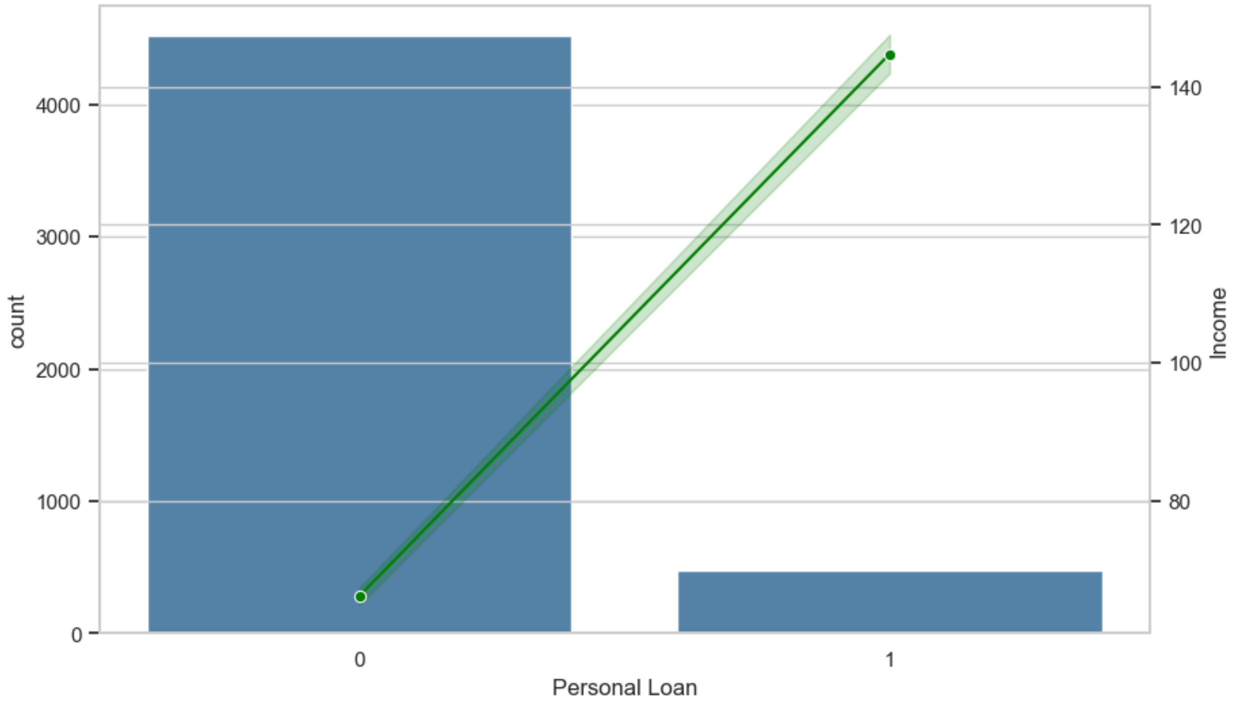
higher incomes are accepting the loans compared to customers with lower incomes who are not accepting the loan offerings.



[2]

Figure two is a box-plot visual of the bank's customer's average spending on credit cards per month compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan) Figure two shows that the customers with higher credit card spending on average are accepting the loans compared to customers with lower credit card spending on average are not accepting the loan offerings.

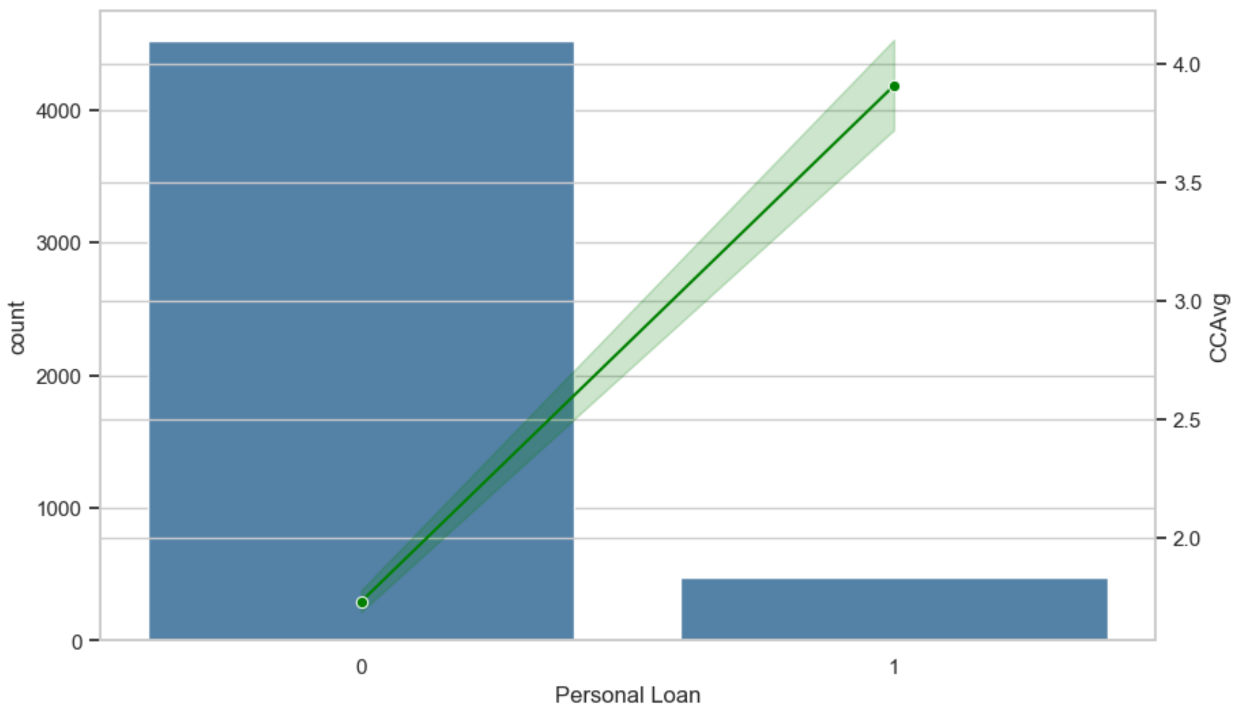
[2] Box-plot visual of the bank's customer's average spending on credit cards per month compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan)



[3]

Figure three above is a combination bar and line visual of the bank's customer's income compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan). In figure three the graph illustrates that on average customers with an annual income of about $145,000 accepts the bank's loan product offering while customers with an annual income of approximately $20,000 do not accept the loan offering.

[3] Combination bar and line visual of the bank's customer's income compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan)



4

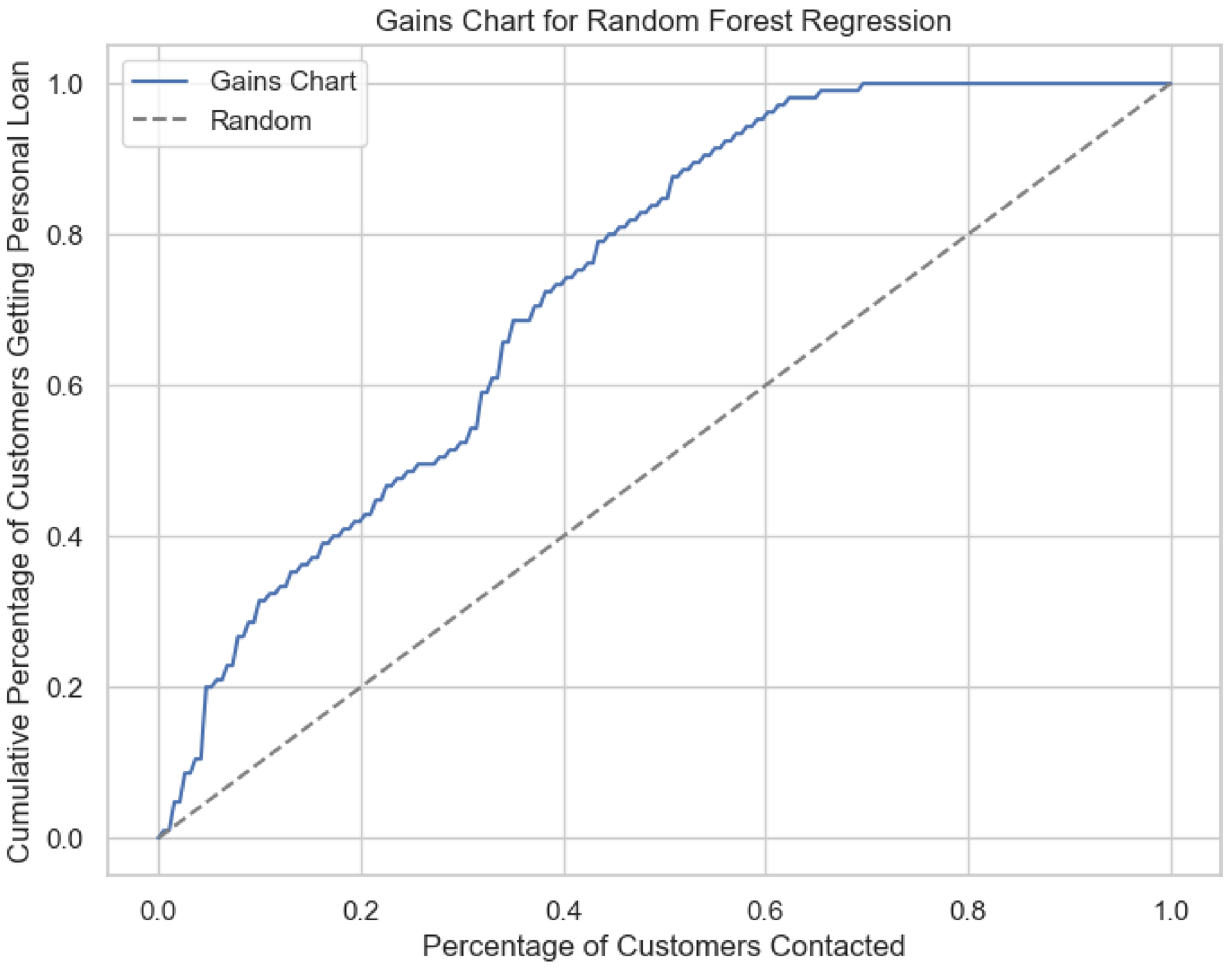
Figure four above is a combination bar and line visual of the bank's customer's average spending on credit cards per month compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan). In figure four the graph illustrates that on average customers that spend on average $3875 per month accepts the bank's loan product offering while customers that spend on average $1650 per month do not accept the loan offering.

Model & Evaluation

In the modeling phase, I choose to utilize a random forest predictive model as I did the train test split on the selected data. For evaluation, I apply a gains chart as well as the accuracy formula. A Random Forest is a supervised machine learning algorithm that grows and brings together numerous decision trees to form a "forest." The model is trained using various diverse examples of several inputs and outputs then learns how to classify all new input data it gathers in the future.

[4] Combination bar and line visual of the bank's customer's average spending on credit cards per month compared to if customers accepted the loan product from the last campaign or not (0 = Did Not accept loan; 1 = Did accept loan)

Accuracy is the fraction of predictions the model got correct. Formula: Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

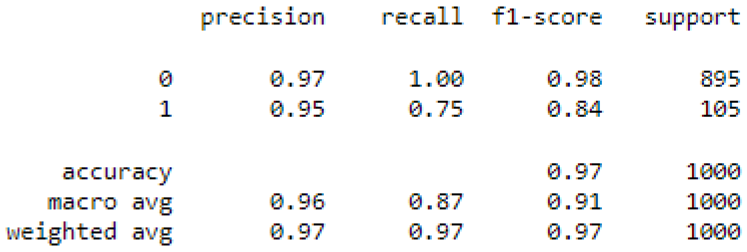


5

The gain chart in figure five, the dashed line expresses “no gain”, meaning, what companies would expect to achieve by marketing to customers at random. The nearer the cumulative gains line is to the top-left corner of the chart, the bigger the gain; the higher the amount of the customer’s getting the personal loans that are reached for the lower amount of customers contacted. As far as accuracy, the accuracy scored 97%, good range is from 70% to 90%, but this doesn’t mean that the model is perfect.

[5] Gains chart of the Random Forest Model

Predictions



```
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       895
           1       0.95      0.75      0.84       105

    accuracy                           0.97      1000
   macro avg       0.96      0.87      0.91      1000
weighted avg       0.97      0.97      0.97      1000
```

[6]

Figure six is the classification report from the Random Forest model I developed. This report displays the precision, recall, F1, and support scores for the model. This report helps express how the model will perform. Lastly, the classification report acts as a tool to analyze then decide how to enhance the model for future utilization and application.

Future Improvements

For the future, to enhance this model I would likely add more data from new incoming customers to help give deeper understandings to be applied to model for maximum optimum use. Another enhancement would be to use multiple evaluation approaches to reveal more areas that could be targeted to boost model performance. Lastly, one advancement is applying different algorithms to extend the scope of the predictive model's performance then compare all to find the best model to act as the model of choice.

[6] Classification Report of Random Forest Model