Issues in Information Systems
Volume 24, Issue 4, pp. 25-39, 2023

DOI: https://doi.org/10.48009/4_iis_2023__103

Predicting credit risks

Erin Carmody, *University of Charleston West Virginia, erincarmody@ucwv.edu*
Taiwo Ajani, *University of Charleston, West Virginia, taiwoajani@ucwv.edu*

Abstract

Credit risk has become an issue for banks and financial institutions as default payments from customers result in lost funds for these companies. Machine learning techniques are being used to assist financial institutions in determining clients with the highest probability of repaying their loans. This study determined factors associated with high-risk clients based on demographic data and credit history. Three supervised learning algorithms were explored in classifying clients into high and low risk categories including logistic regression, random forest, and k-nearest neighbor. Models were evaluated for performance using the confusion matrix technique. Results showed that Random Forest was the best fit model for this study.

Keywords credit risk, default, logistic regression, random forest, k-nearest neighbor

Introduction

As inflation rises and consumers recover from the recent Covid-19 pandemic, loan defaults remain a major concern for financial institutions. According to Schulz (2023), Americans have accumulated about $986 billion in credit card debt; rising interest, inflation, and other economic factors will continue to exacerbate this problem. Many are not able to pay back debts, resulting in huge losses for lenders. To reduce the rate of loan defaults and minimize losses, financial institutions perform screenings to assess borrowers' risk status regarding willingness or ability to repay loans, by conducting scientific investigation of the factors associated with such risks. A thorough examination of a client's history helps to gather needed information for these assessments.

Traditional methods for this process centered around the five C's, character, capital, capacity, collateral, and conditions, when determining the borrower's credibility (Li, 2019). However, these proved to have limitations as the evaluation required in-depth knowledge of the borrower and was dependent on the previous experiences of the evaluator. In essence, it is time consuming. Even after conducting these assessments, there still is no guarantee that borrowers will not default on loans. Financial institutions have also focused more on credit scores to highlight client's creditworthiness. These credit scores were created using a formula made by the Fair Isaac Corporation (FICO) that is believed to be based on the ratio of debt to available credit and adjusted for aspects like payment history, negative events, income changes, and number of credit applications (Arya et al., 2013). Formulating these scores not only requires mathematical algorithms, but skilled individuals to properly interpret and predict a client's creditworthiness.

Machine learning has since been adopted by researchers to change the way institutions and banks assess individual borrowers. Using a variety of deep learning algorithms and classification techniques, it can take the payment history data of the borrower and immediately predict whether they are credit worthy, making the process easier for banks and other financial institutions.

Existing research today regarding credit risk analysis with machine learning lacks ease of interpretability from its models. Scholars have used different algorithms to predict credit risk, but many put little focus towards which elements related to an individual contribute to their risk level. Feature importance studies regarding this topic, especially in the banking industry, are not heavily focused on. By narrowing the focus to concrete factors, it can assist credit risk analysis across all sectors even further. They can use this information to determine potentially risky customers before approving them of a loan.

This paper addresses multiple business challenges. The focus highlighted in this study is determining which factors can be used to identify high-risk clients. High-risk clients refer to individuals that are more likely to default on their loans than others. Three supervised machine learning models including logistic regression, k-nearest neighbor, and random forest were trained to classify clients into high or low risk categories. The models were subsequently validated for their efficiency at predicting whether a new credit loan applicant will or will not default on their loans. The study also assessed the association between risk and socio-economic factors (e.g. gender, occupation, or family status); as well as to determine the average income level of high-risk clients.

Literature Review

Request for loan is at an unprecedented increase. To meet high demands, financial institutions must look for alternative techniques to optimize their loan assessments for the most accurate findings. Thus, the shift from professionals conducting evaluations based on their expertise to automation in order to increase effectiveness and efficiency.

A study performed on a Lending Club data set used decision trees and random forest for a comparative analysis to determine whether an individual should be given a loan (Maadan et al., 2021). This loan default prediction analysis identified random forest as being the ideal model for prediction as it yields an accuracy level of 80% compared to the decision tree's accuracy of 73%. Random forest was found to be optimal in another study for credit risk assessment when compared to four different classifiers, k-nearest neighbor, decision tree, naïve bayes, and logistic regression. A comparative evaluation for deciding whether an individual should be granted credit financing produced a high performing random forest model with area under the curve at 92%, accuracy level of 96.53%, and precision of 97.16% (Wang et al., 2019). Random forest proved to be best in a study regarding P2P online lending platforms.

For predicting loan defaults, the random forest outperformed the other three models with an accuracy level, f1 score, and recall of 98% (Zhu et al., 2019). One study for assessing the probability of loan default discovered that tree-based models were the most stable models when put against other classification models. The gradient boosting model produced the highest performance in terms of area under the curve and root mean square error while using the top ten variables including EBITDA, equity, liabilities, raw financials, cash flows, profits, and flows (Shi et al., 2022). A more recent study conducted by Jumaa et al. (2023) employed deep learning algorithms to create a consumer loan default prediction model to minimize credit risk in the banking sector yielding a test accuracy percentage of 95.2%.

Model comparison for credit risk has been presented in recent research to determine the best fit model for the assessment. A comparison study performed on auto loan credit risk data combined particle swarm optimization with XGBoost and compared it to numerous other machine learning techniques to determine the ideal model for credit risk assessment (Rao et al., 2023). Krivorotov (2023) used both traditional risk models and machine learning models to identify profit score cut offs that can be used to distinguish higher risk customers from lower risk customers. Feature importance has been a key aspect in credit risk evaluation to determine certain criteria that separates individuals from being at a high or low risk for default. Another study used Iranian bank information to create a dynamic model based on bad customers on a monthly basis

and incorporated economic and political fluctuations. Using this approach, they discovered that many defaults were securitized by large collaterals and were among backed loans. Morality was key for default payments, and 5% of bad debtors on high amount loans didn't have any kind of economic issue, they simply refused to pay loan before bank forecloses (Moradi & Mokhatab Rafiei, 2019).

Barbaglia et al. (2021) conducted research to predict loan default on residential mortgage data in European countries with the help of machine learning algorithms. They discovered that interest rate and local economic characteristics for the most important variable for explaining loan default. Feature selection has been key to improving credit risk analyses by only utilizing the variables that hold importance resulting in removal of redundant variables. Fuzzy theory supported one study by increasing the prediction power through feature selection in calculating default risk for credit clients (Baser et al., 2023). With minimal research being done on specific feature selection and importance for determining criteria of good credit clients, this study aims to start pointing research in that direction.

Methodology

Data Collection

The data source used in this study was the "Credit Card Approval Prediction" gathered from Kaggle. Two CSV data files make up this data set, namely, application record and credit record. They contain various demographic factors, credit history, and bill statements of credit card clients. The application record contains a total of 438,557 observations and 18 variables, and the credit record contains a total of 1,048,575 observations and 3 variables.

Data Preparation

To create the target variable for this study, the STATUS column, and MONTHS_BALANCE columns were used to formulate the RISK column. The RISK column contains two values 0 and 1 representing the low-risk and high-risk clients respectively. Using the credit_record.csv, the values from the STATUS column were extracted for each ID. The values from the STATUS column were used to create five new columns that will be used to help group the data for the RISK column. The five columns were calculated as follows:

1. The total_past_due column sums all the overdue payments for each ID.
2. The less_90 column sums all the overdue payments that were 90 days or less.
3. The over_90 column sums all the overdue payments that were greater than or equal to 90 days.
4. The past_due_diff column calculates the difference between the loans paid off and the total past due payments.
5. The months column counts the months balance for each ID using the MONTHS_BALANCE column from the data.

If-then statements were used to create conditions using these columns to make the RISK column. A client was associated as low risk if their past_due_diff was greater than or equal to 3 or their no_loan was equal to their months. Higher positive values meant that the individual paid off their loans. Clients were also grouped as low-risk only if their past_due_diff greater than -2 but less than 3 and less_90 was greater than or equal to over_90. This was to account for individuals paying off in a small period. All other clients not meeting these conditions were grouped as high-risk.

The two CSV files were combined on the ID column produced a total of 36,457 observations and 25 variables. There were numerous missing values located in the data set in the OCCUPATION_TYPE column, so proper cleaning methods were conducted to impute and remove these items.

OCCUPATION_TYPE contained 11,323 blank values that were able to be imputed with the value, Unknown, so that the data did not have to be lost.

The goal of this research is to classify the clients into categories of high risk and low risk. With no binary column present to represent this type of data, the target variable, Risk, needed to be created using the columns from the credit_record.csv file. The credit industry highlights several factors that contribute to evaluating whether a client is accepted or rejected for a loan amount. Delinquent accounts are one of the most important aspects that credit companies observe when assessing risk. Experian, the world's leading global information services company, states 90% of top lenders use payment history as the sole factor for calculating a FICO score (Egan, 2022).

The STATUS column in the data set highlights this factor by breaking down into ranges missed payments by days of the clients. MONTHS_BALANCE recorded the start month that the loan was distributed. Since the STATUS column was broken up into months, this made it easier to track individuals that were making and missing payments.

Additional data preparation included converting data types, removing irrelevant columns, and creating new columns. Binary categorical columns (CODE_GENDER, FLAG_OWN_CAR, and FLAG_OWN_REALTY) were converted to numerical values, so they could be used in numerical analysis and correlations. New columns were created from DAYS_BIRTH and DAYS_EMPLOYED to allow easier interpretation. Age was calculated by dividing the absolute value of DAYS_BIRTH column by 365.2422, which is the length of a year on Earth.

An unemployment column was created to determine the individuals who were not employed using a conditional statement. If DAYS_EMPLOYED was greater than 0, the individual is unemployed and given the value 1. Employed individuals were given the value 0.

Employment days were calculated by dividing the absolute value of DAYS_EMPLOYED by 365.2422. Using a similar conditional statement, if DAYS_EMPLOYED was less than or equal to 0, the formula was used. If not, the individual received a value of 0 for that column. Removal of the DAYS_BIRTH, DAYS_EMPLOYED, and FLAG_MOBIL was done as there was no longer a need for these columns. FLAG_MOBIL only contained a value of 1 for all responses, thus it would not add value to the study.

Preparation for Modeling

The clean dataset was partitioned into training and testing in 7:3 ratio. The models were trained on the former and tested on the latter. Subsequently, models were evaluated for performance and compared.

Logistic regression is a form of supervised machine learning method that helps to predict if a data point belongs to a certain class. It works with a binary target variable, which in this case will determine whether the client has high or low credit risk. Values for the target variable typically are either 0 or 1, and the model predicts an output based on different dependent variables. The model is beneficial due to its computational efficiency, simple implementation, ease of regularization, and additional multicollinearity has little impact on the result (Vyas et al., 2023). This is a classification model.

K-nearest neighbor (KNN) is another classification method that uses proximity of data points to group them into respective classes. It is often referred to as a lazy technique as it does not require any training of the data points in order to generate a model. The basic process that is performed to complete this method follows these steps: Measure the distance using either Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance, locate the nearest neighbors, and select labels (Vyas et al., 2023). This

is a non-parametric model that can be considered both regression and classification. However, it is normally used in classification problems.

Random forest consists of multiple decision trees that can produce more accurate results than a single decision tree. Each individual tree is made using various bagging and feature randomness techniques to create this forest of unrelated trees that can be used to predict the classes. Because of this, random forests can provide better results than decision trees.

The scalability, speed, noise-reducibility, and no overfitting make this model highly desirable to use for classification challenges (Vyas et al., 2021). This model can handle large data sets in an efficient manner as well. Table 1 as shown below, lays out the tuning parameters conducted on each model to improve the prediction accuracy of them.

Machine Learning Classifiers

Table 1: Classification model descriptions of tuning parameters.

Model	Type	Tuning Parameter	Description
Logistic Regression	Classification	AVS removed columns of FLAG_EMAIL, FLAG_PHONE, FLAG_WORK_PHONE, FLAG_OWN_CAR, CODE_GENDER, OCCUPATION_TYPE, NAME_INCOME_TYPE	AVS stands for automatic variable selection and is used to remove redundant variables from the input data.
K-nearest Neighbor	Classification	K=160	K represents the number of neighbors that will be examined to find the classification of a data point.
Random Forest	Classification	ntree=500, mtry=12	Ntree refers to the number of trees in the model. Mtry refers to the number of variables that are chosen at random for each split.

Results

Exploratory Data Analysis

The boxplot displayed in Figure 1 splits the annual income for the clients based on their risk level. Both plots are heavily skewed to the right making the median a better measure of center. The median annual income for the low and high-risk clients appears to be around $180,000. Numerous outliers are present to the right of the box plots, and the small size of the boxes indicates little dispersion between the data points.

There is no significant difference between the annual income level between the different risk levels of the clients. The months balance boxplot by risk level shown on the right in Figure 2 displays major differences. High risk clients have a lower median month's balance than low risk clients at about 12 and 25 respectively. The high-risk box plot is skewed to the right with several outliers located between about 41 to 62 months. For the low-risk box plot, the data is normally distributed with a larger amount of variation than the high-risk box plot. There are no visible outliers present in the plot for low-risk clients.

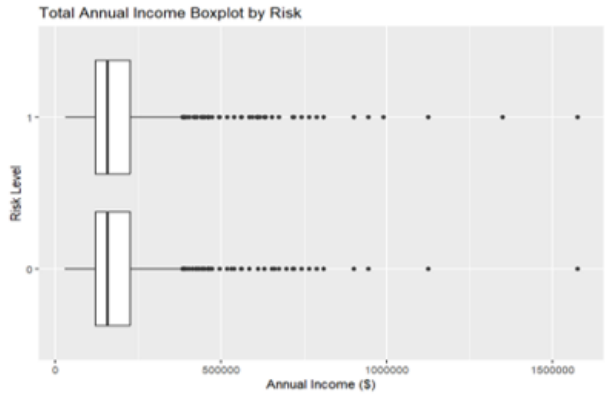

Figure 1: Box plot of annual income by risk level

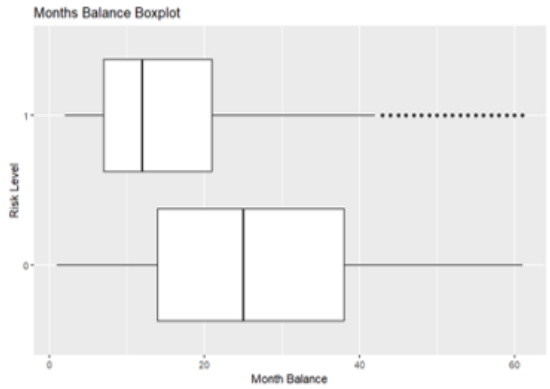

Figure 2: Box plot of months balance by risk level

Figure 3 shows the bar plot of high and low risk clients in the data set. This target variable is balanced as there is almost an equal amount for both classes. There is a total of 18,708 people that are considered low risk, and 17,749 people that are considered high risk. Figure 4 represents the amount of people associated with each type of family status. Civil marriage, married, separated, single/not married, and widowed are the five values that an individual in the data set can be categorized to. Married family status contains the largest amount of people at about 25,000, and widowed contains the smallest amount of people at about 1,500. Single/not married is the second largest type at only about 5,000, which is significantly less than the married status.

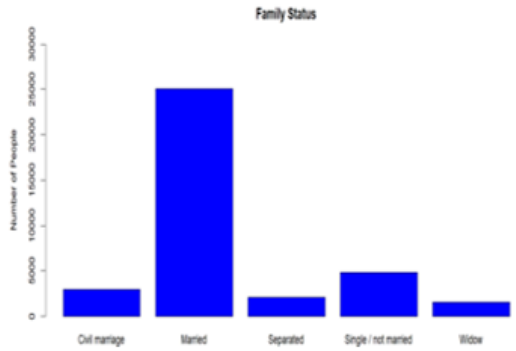

Figure 3: Distribution of risk level in the data set

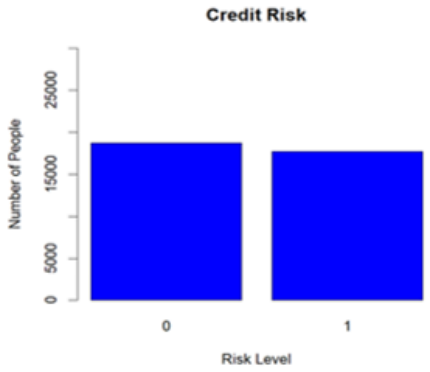

Figure 4: Amount of people related to each family status

Figure 5 shows the screenshot of the summary statistics from the final data set for analysis. The mean value for annual income is $186,686. However, there appears to be several outliers present in the data implying skew. The minimum annual income is $27,000, and the maximum is $1,575,000. Using the median gives a better measure of center due to the outliers. The median value for annual income is $1,575,000. The average age of the clients is about 43.74 years old, indicating that most of them are middle-aged. For years employed, there appears to be a wide gap between the minimum and maximum, so the median will again be the better measure of center for this column. Median years employed is about 4.25 years. Most clients in the data set do not have children, and the average number of family members is about 2 people.

```
> summary(num2)
    no_loan       total_past_due      less_90          over_90         past_due_diff         months           RISK         CODE_GENDER
 Min.   : 0.000   Min.   : 0.00   Min.   : 0.000   Min.   : 0.0000   Min.   :-61.000   Min.   : 1.00   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 0.000   1st Qu.: 3.00   1st Qu.: 3.000   1st Qu.: 0.0000   1st Qu.: -8.000   1st Qu.: 9.00   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 1.000   Median : 6.00   Median : 6.000   Median : 0.0000   Median : -1.000   Median :18.00   Median :0.0000   Median :0.0000
 Mean   : 4.003   Mean   : 8.29   Mean   : 8.234   Mean   : 0.0556   Mean   :  0.749   Mean   :21.33   Mean   :0.4868   Mean   :0.3299
 3rd Qu.: 3.000   3rd Qu.:11.00   3rd Qu.:11.000   3rd Qu.: 0.0000   3rd Qu.:  8.000   3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :61.000   Max.   :61.00   Max.   :61.000   Max.   :48.0000   Max.   : 59.000   Max.   :61.00   Max.   :1.0000   Max.   :1.0000
  FLAG_OWN_CAR    FLAG_OWN_REALTY   CNT_CHILDREN     AMT_INCOME_TOTAL  FLAG_WORK_PHONE    FLAG_PHONE      FLAG_EMAIL      CNT_FAM_MEMBERS
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.0000   Min.   :  27000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   : 1.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 121500   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.: 2.000
 Median :0.0000   Median :1.0000   Median : 0.0000   Median : 157500   Median :0.0000   Median :0.0000   Median :0.00000   Median : 2.000
 Mean   :0.3797   Mean   :0.6722   Mean   : 0.4303   Mean   : 186686   Mean   :0.2255   Mean   :0.2948   Mean   :0.08972   Mean   : 2.198
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 225000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.: 3.000
 Max.   :1.0000   Max.   :1.0000   Max.   :19.0000   Max.   :1575000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :20.000
      AGE         UNEMPLOYED      YEARS_EMPLOYED
 Min.   :20.50   Min.   :0.0000   Min.   : 0.000
 1st Qu.:34.12   1st Qu.:0.0000   1st Qu.: 1.117
 Median :42.61   Median :0.0000   Median : 4.249
 Mean   :43.74   Mean   :0.1683   Mean   : 6.024
 3rd Qu.:53.22   3rd Qu.:0.0000   3rd Qu.: 8.633
 Max.   :68.86   Max.   :1.0000   Max.   :43.021
>
```

Figure 5: Summary statistics of the numerical columns in the data set.

Figure 6 presents a correlation matrix that indiacates that most of the variables have weak correlation coefficient with one another. Exceptions include CNT_CHILDREN and CNT_FAM_MEMBERS indicating a strong, positive relationship with one another (0.89). AGE and UNEMPLOYED have a moderate, positive relationship as their correlation coefficient was 0.62. UNEMPLOYED and YEARS_EMPLOYED have a moderate to weak, negative relationship with one another at about -0.42. The positive correlations explain that as one of the variables increases, the other one increases as well. However, the negative correlations indicate an inverse relationship because as one increases, the other will decrease.

```
> cor(num)
                    months         RISK  CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY  CNT_CHILDREN AMT_INCOME_TOTAL FLAG_WORK_PHONE   FLAG_PHONE
months           1.000000000 -0.349645304  0.008995695  0.034323601   -0.0384951729 -0.0043029914     0.02598785     0.02153644  0.019301921
RISK            -0.349645304  1.000000000 -0.007039893 -0.002875586    0.0081034006 -0.0086242858     0.01604460    -0.01206217  0.004135836
CODE_GENDER      0.008995695 -0.007039893  1.000000000  0.361378903   -0.0507584201  0.0776903994     0.19780496     0.06499354 -0.026832878
FLAG_OWN_CAR     0.034323601 -0.002875586  0.361378903  1.000000000   -0.0151846115  0.1058386592     0.21550604     0.02164383 -0.014019190
FLAG_OWN_REALTY -0.038495173  0.008103401 -0.050758420 -0.015184611    1.0000000000 -0.0005748801     0.03271882    -0.20773159 -0.066600842
CNT_CHILDREN    -0.004302991 -0.008624286  0.077690399  0.105838659   -0.0005748801  1.0000000000     0.03369093     0.04809087 -0.016290567
AMT_INCOME_TOTAL 0.025987855  0.016044601  0.197804960  0.215506044    0.0327188228  0.0336909255     1.00000000    -0.03774611  0.017244699
FLAG_WORK_PHONE  0.021536436 -0.012062173  0.064993539  0.021643833   -0.2077315914  0.0480908743    -0.03774611     1.00000000  0.311644305
FLAG_PHONE       0.019301921  0.004135836 -0.026832878 -0.014019190   -0.0666008420 -0.0162905668     0.01724470     0.31164431  1.000000000
FLAG_EMAIL       0.009775573  0.001060329 -0.003284197  0.021749989    0.0521943401  0.0159596128     0.08668149    -0.03483799  0.010454902
CNT_FAM_MEMBERS  0.016295149 -0.008929426  0.110781946  0.151814364   -0.0057229236  0.8891141560     0.02374994     0.06452691 -0.004221315
AGE              0.051005883  0.014352879 -0.202352235 -0.157143970    0.1298381243 -0.3393569322    -0.06790835    -0.17905357  0.028659340
UNEMPLOYED      -0.014269835  0.008535834 -0.175261057 -0.157496003    0.0931085882 -0.2303173379    -0.16834431    -0.24272986 -0.006541084
YEARS_EMPLOYED   0.075025488 -0.002926927 -0.031731478  0.006243702   -0.0336459079  0.0433576106     0.08713033     0.11132721  0.043025230
                 FLAG_EMAIL CNT_FAM_MEMBERS         AGE   UNEMPLOYED YEARS_EMPLOYED
months          0.009775573     0.016295149  0.05100588 -0.014269835    0.075025488
RISK            0.001060329    -0.008929426  0.01435288  0.008535834   -0.002926927
CODE_GENDER    -0.003284197     0.110781946 -0.20235223 -0.175261057   -0.031731478
FLAG_OWN_CAR    0.021749989     0.151814364 -0.15714397 -0.157496003    0.006243702
FLAG_OWN_REALTY 0.052194340    -0.005722924  0.12983812  0.093108588   -0.033645908
CNT_CHILDREN    0.015959613     0.889114156 -0.33935693 -0.230317338    0.043357611
AMT_INCOME_TOTAL 0.086681492    0.023749941 -0.06790835 -0.168344308    0.087130327
FLAG_WORK_PHONE -0.034837986    0.064526914 -0.17905357 -0.242729858    0.111327206
FLAG_PHONE      0.010454902    -0.004221315  0.02865934 -0.006541084    0.043025230
FLAG_EMAIL      1.000000000     0.014619114 -0.10562536 -0.086316036   -0.002167800
CNT_FAM_MEMBERS 0.014619114     1.000000000 -0.30402031 -0.221924990    0.054587259
AGE            -0.105625364    -0.304020313  1.00000000  0.620345157   -0.023496858
UNEMPLOYED     -0.086316036    -0.221924990  0.62034516  1.000000000   -0.418175552
YEARS_EMPLOYED -0.002167800     0.054587259 -0.02349686 -0.418175552    1.000000000
>
```

Figure 6: Correlation matrix of numerical variables in the data set.

Performance Metrics

Accuracy = (TP + TN)/ (TP + FP + FN + TN)

Sensitivity = TP/ (TP +FN)

Specificity = TN/ (TN + FP)

The three supervised machine learning algorithms performed in this study were evaluated using confusion matrices. Confusion matrices visualize the predictions and results of classification problems. It shows the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). True positives represent the number of correct positive predictions. True negatives represent the number of correct negative predictions. False positives represent the number of incorrect positive predictions. False negatives

represent the number of incorrect negative predictions. Using these values, accuracy, sensitivity, and specificity were calculated for each model to measure their performance. Accuracy determines the number of overall correct predictions the model made. Sensitivity relates to determining how well the model can predict positive cases while specificity relates to determining how well the model can predict negative cases.

Logistic Regression

Training and testing sets were created to measure the model's performance. 70% of the data went into formulating the training set, and the remaining 30% went into the testing set. The logistic regression was created on the training set, and summary statistics of the model displayed the statistical significance of each variable. Statistical significance was indicated by p-values that were closest to zero. Variables showing the highest statistical significance included values from the following:

AGE
NAME_INCOME_TYPEPensioner
AMT_TOTAL_INCOME
MONTHS.

Other key values associated with credit risk included the following:

FLAG_OWN_REALTY
CNT_CHILDREN
NAME_HOUSING_TYPEHouse/apartment
NAME_HOUSING_TYPEMunicipal apartment
NAME_HOUSING_TYPEOffice apartment
NAME_HOUSING_TYPERented apartment
CNT_FAM_MEMBERS

From the analysis, age, the annual income a client makes, the months balance, and if they are a pensioner have the highest importance in determining credit risk. Factors such as owning apartment property and the amount of children and family members in a household play a role in predicting credit risk as well.

To determine the level of accuracy of the logistic model, confusion matrices were made for both the training and testing sets to compare levels of accuracy, precision, sensitivity, and specificity. Figure 7 displays the visualization for the training matrix and Figure 8 displays the visualization for the testing matrix. The training set produced an accuracy level of 66.89% showing that the model was moderately effective at making correct predictions. The sensitivity level was 62.52% and specificity level was 71.50% meaning that the model was better at predicting negative cases over positive cases by about 10%. The testing set produced an accuracy level of 66.84%, sensitivity level of 62.49%, and specificity level of 71.42% validating that the model has a moderate to low level of prediction accuracy.

ROC curves can help further assess performance of classification models. Figure 9 shows the ROC-AUC curve that was created for the logistic regression model. The calculated area under the curve 66.95%, which indicates that the binary factor in the model, RISK, has poor accuracy level at distinguishing between the low risk and high-risk clients in the data. In the graph, the closer ROC curve approaches the top left corner, the better the model. This model is closer to the baseline indicating it poor performance at deciphering between the two classes. To improve the model, AVS was used to determine what variables could be removed to lower the AIC and increase accuracy. The FLAG_EMAIL, FLAG_PHONE, FLAG_WORK_PHONE, FLAG_OWN_CAR, CODE_GENDER, OCCUPATION_TYPE, and

NAME_INCOME_TYPE columns were removed from the model reducing the AIC from 31.9433.33 to 31931.83. Using AVS on the model to not impact the model significantly as it only increased the accuracy by about 0.15%.

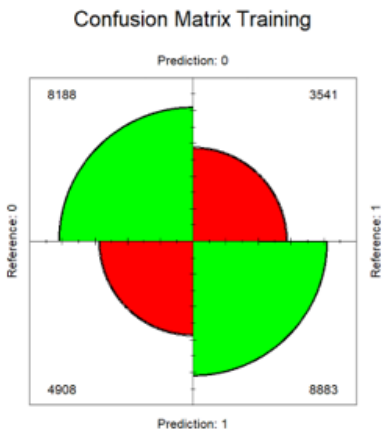

Figure 7: Confusion Matrix Training

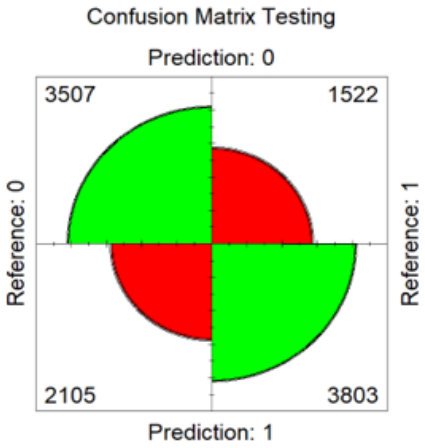

Figure 8 Confusion Matrix Testing

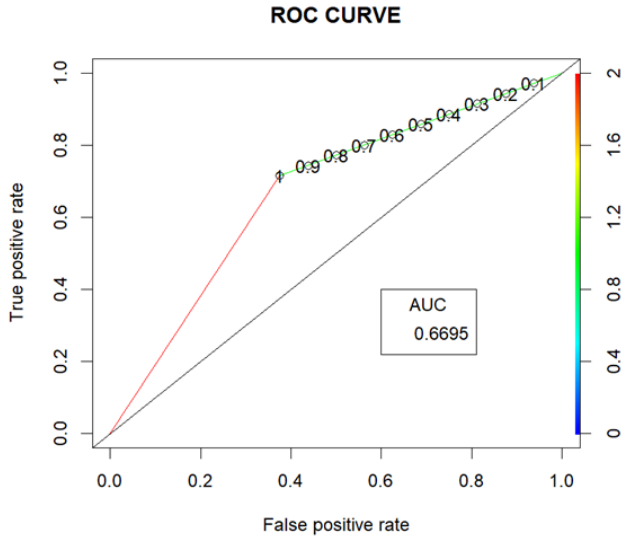

Figure 9: ROC curve of logistic regression to assess performance of the model.

K-Nearest Neighbor

Training and testing sets were created with and without the target variable, RISK, in preparation for this model. Optimal k is determined by taking the square root of the number of observations. For the initial model, the k value was selected at 159 as the number of observations in the training set was 25,520. Based off the value of 159, the accuracy levels of models with k values from 139 to 161 were calculated to determine which value would produce the best model. A k value of 160 resulted in the highest accuracy level at 65.54%. The final model was created with this value.

Figure 10 displays the first six rows of a data frame created using the predicted values from the KNN algorithm compared to the actual values from the testing set. The predicted values correctly matched 4 out

of the 6 first actual values for the data. However, a confusion matrix was created, which is shown in Figure 11, to determine the performance of the model. Balanced accuracy for the algorithm was 65.55%, which indicates that the model had poor performance. Sensitivity and specificity were calculated and yielded values of 61.14% and 69.97%. Like the previous model, the KNN algorithm predicted the negative cases more accurately than the positive cases by about 9%. The logistic model produced a slightly higher accuracy than this KNN algorithm by only about 1%.

	pr2	target_test
1	1	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	1

Figure 10: Data Predicted Vs. Test Data

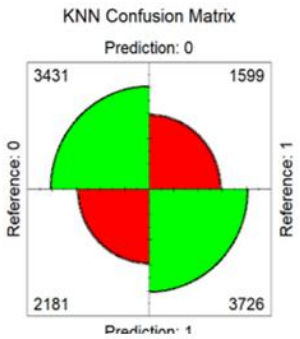

Figure 11: KNN Confusion Matrix with k of 160

Random Forest

Like the logistic model, the training data set was used to create the model testing the same variables. The forest was created with 500 trees. At each split, four variables were tried, and the out of bag (OOB) estimate of error produced was 25.17%. The OOB error measure is calculated by determining the probability that a prediction is wrong in the data set. According to the confusion matrix created in the model, there was a 22.19% misclassification error for the false positives, and a 28.31% misclassification error for the false negatives. The model was then tuned to improve accuracy and prediction quality. Mtry is the value for the number of variables selected at each split. Tuning the random forest showed that the optimal mtry value was 12 variables. Figure 12 displays the OOB error for each of the mtry values. The mtry value that had the lowest OOB error in the graph was 12 at about 0.0000216%. Using the selected mtry value, new models were created on the training and testing sets. The new training model had an OOB estimate of error of 25.61%. The projected misclassification errors for the false positives were 23.14% and false negatives was 28.22%.

To discover the important variables from the model, mean decrease accuracy and mean decrease Gini plots were visualized (Figure 13). Mean decrease accuracy explains how much of a decrease the accuracy of the model would be if a variable was removed from the model. Mean decrease Gini determines variable importance using the Gini Impurity Index. Gini Impurity Index calculates how the features should be separated in the data set to formulate all the decision trees. Months had the highest value in both graphs with a mean decrease accuracy of about 500, and a mean decrease Gini of about 3,500 showing that it was the most important variable in the model. Other key variables that proved to be of high importance were AGE, AMT_INCOME_TOTAL, and YEARS_EMPLOYED. UNEMPLOYED was the least important variable in the model as it produced mean decrease accuracy and mean decrease Gini values a little over zero.

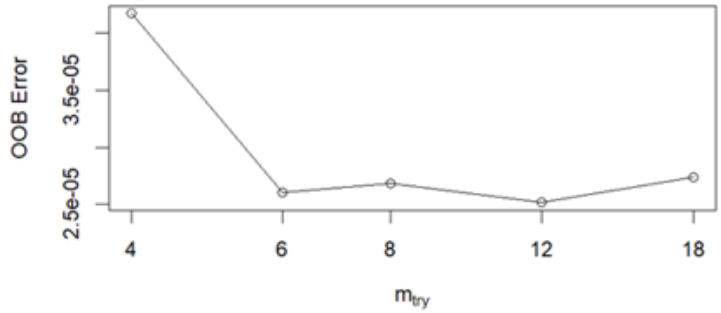

Figure 12: OOB error associated with each ntry value

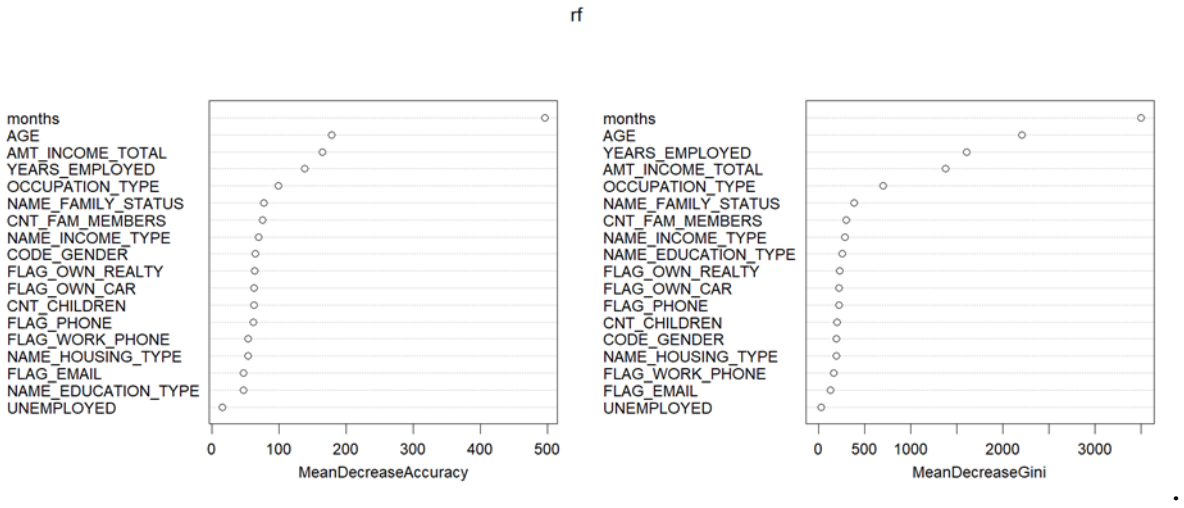

Figure 13: Mean Decrease Accuracy and Mean Decrease Gini Decision tree outcome plots

Through analysis of the random forest, specific factors were able to be determined for identifying clients as high or low risk. Key variables proving to have the high importance in the model were MONTHS, AGE, AMT_INCOME_TOTAL, and YEARS_EMPLOYED. Validation of the model was determined using confusion matrices like the previous logistic model. Confusion matrices shown in Figure 10 for both the training and testing sets were conducted to compare accuracy, precision, sensitivity, and specificity levels. The training set confusion matrix yielded an accuracy level of 97.20%. Compared to the logistic model, the random forest produced around a 30% higher accuracy.

Sensitivity and specificity values were calculated at 95.17% and 99.13%. This model was slightly better at predicting the negative cases, or the high-risk clients, over the positive cases. The balanced accuracy was calculated at 97.23% showing high prediction accuracy between each individual class. Validation on the testing set produced the sensitivity and specificity values of 99.13% and 98.25%. The accuracy of the model was slightly higher than on the testing set as it was 98.70%. Due to the high values for accuracy, sensitivity, and specificity of the random forest model, it can be accepted and is the best fit model in the study.

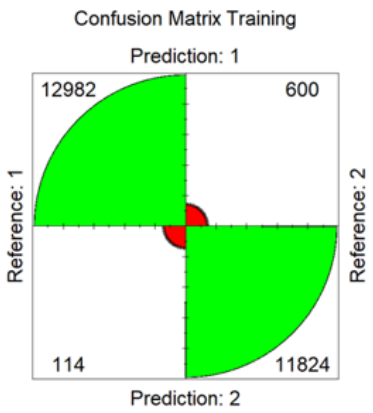

Figure 14: Random Forest Confusion Matrix Training

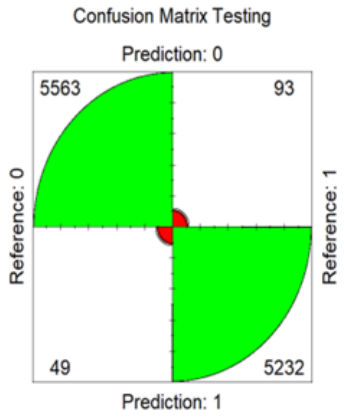

Figure 15: Random Forest Confusion Matrix Testing

Model Comparison

Table 2: Evaluation of model performance using metrics of accuracy, sensitivity, and specificity

	Accuracy	Sensitivity	Specificity	False Negatives
Random Forest	97.20%	95.17%	99.13%	49
Logistic Regression	66.89%	62.57%	71.50%	2,105
K-Nearest Neighbor	65.44%	61.14%	69.97%	2,181

.

Based on the accuracy, sensitivity, and specificity levels produced by each model, we can see in Table 2 that the random forest model outperformed the others. The random forest's accuracy was about 30% higher than the logistic regression and about 32% higher than the KNN model. Accuracy determines how close the estimated value is to the original value. Its sensitivity level was about 33% higher than the logistic regression and about 34% higher than the KNN model. Sensitivity refers to the true positives, so it assesses the total number of correct positive predictions divided by the overall total number of positive values. Its specificity level was about 28% higher than the logistic regression and about 29% higher than the KNN model.

Specificity refers to the true negatives, so it assesses the total number of correct negative predictions divided by the overall total number of negative values. It is important to not just base model performance solely on accuracy but on other measures as well. The random forest model ended up producing the lowest number of false negatives in this study. This value dealt with the clients that were not predicted to default on their loans but did. This is one of the most important values to consider as companies will greatly suffer from accepting these applicants. With only 49 false negatives for the random forest model compared to 2,105 for logistic regression and 2,181 for k-nearest neighbor, it proved that the random forest model was the best fit for this research.

Implications and Future Work

Credit risk is a prominent issue in the world, especially financial institutions as they have the necessity to minimize loan approval for high-risk clients. Without this process, institutions will lose massive amounts of capital by approving loans to individuals that lack the willingness or wherewithal to pay back loans. Machine and Deep learning algorithms provide faster results to classifying clients into one of low and high-risk categories. By assessing the specific characteristics and socio-economic features of individuals,

meaningful insights for determining creditworthiness can be obtained. Therefore, financial institutions' business goals can be properly aligned for growth and better services to consumers while minimizing risks the type of risks that bankrupted financial institutions in the past. Future research should focus on improving the performance of the random forest model. Shallow and deep learning techniques can be performed to compare performances with the random forest in a bid to keep minimizing the approval of loans to high-risk clients. Other questions pertinent to future work should also include: How should we minimize false negative and false positives? To what extent can we tolerate either, while we maintain the efficacy of the model?

Limitations

A major limitation of this study is data size. With only a total of 36,457 unique IDs from a single institution in the data set, the potential for bias and lack of variability can potentially skew results and impact the model viability. Analysis on a larger sample from a variety of financial institutions can potentially eliminate bias. Another limitation of this study is that individuals that possess the characteristics of a low-risk client can still default on their loan and high-risk clients can still repay their loan. Demographic factors found in this study cannot solely determine the risk level of an individual. These should be considered in combination with the individual's payment history and other credit data to provide more accurate results.

Conclusions

Initial hypotheses suggested that education, income level, and occupation are dependent on credit risk as they are key factors for assessing the level of a customer's credit risk. As credit default risk deals with the customer's ability to pay back their loan, higher income individuals could have an easier time repaying their loans over lower income individuals. Albanesi et al. (2017) found that income was one of the most important factors for credit scores as the two have a strong positive relationship. Higher education such as bachelors', masters', or doctorate degrees are required for many high paying careers like doctors or lawyers. Thus, it is predicted that higher risk individuals will have less education and lower paying professions than lower risk individuals.

Three different models were created to classify individuals into low or high-risk categories based on demographic factors and credit history. The logistic model yielded accuracy levels below the accepted range at 66.87% and an area under the curve of 66.95%. It was able to predict negative cases better than positive cases by about 10 points according to the sensitivity and specificity values calculated. The random forest model proved to have high prediction accuracy as the accuracy level produced was 97.20%. High sensitivity and specificity levels were associated with the model as well.

Based on the results, the random forest model is the best fit model for analysis due to the above average validation. Important variables gathered for classifying risk levels were AGE, YEARS_EMPLOYED, MONTHS, and AMT_INCOME_TOTAL. However, through exploratory analysis, the average annual income level between high and low risk clients was roughly the same at about $180,000. Certain demographic factors had influence on risk level such as age, housing type, occupation, and education. Chi-squared hypothesis tests revealed that occupation and education were both dependent on risk level due to the low p-values and high chi squared values. The null hypothesis of the risk variable being independent of occupation type and education was rejected as their p-values were below the significance level of 0.05.

This proved that two out of the three initial hypotheses were correct as only income level was independent of risk level. Further research will be needed on certain demographic factors to verify the relationship with risk level. The findings in this research are beneficial to financial institutions in understanding the best

features and models that can reliably predict clients' willingness or ability to pay loans back, thereby minimizing losses that are often associated with high-risk default customers.

References

Addo, P., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and Deep Learning Models. *Risks*, *6*(2), 38. https://doi.org/10.3390/risks6020038

Albanesi, S., De Giorgi, G., & Nosal, J. (2017). *Credit Growth and the Financial Crisis: A New Narrative*. https://doi.org/10.3386/w23740

Arya, S., Eckel, C., & Wichman, C. (2013). Anatomy of the credit score. *Journal of Economic Behavior & Organization*, *95*, 175–185. https://doi.org/10.1016/j.jebo.2011.05.005

Barbaglia, L., Manzan, S., & Tosetti, E. (2021). Forecasting loan default in Europe with machine learning. *Journal of Financial Econometrics*, *21*(2), 569–596. https://doi.org/10.1093/jjfinec/nbab010

Baser, F., Koc, O., & Selcuk-Kestel, A. S. (2023). Credit risk evaluation using clustering based fuzzy classification method. *Expert Systems with Applications*, *223*, 119882.

Egan, J. (2022, July 18). *What's the most important factor of your credit score?* Experian. Retrieved February 16, 2023, from https://www.experian.com/blogs/ask-experian/what-factor-has-the-biggest-impact-on-credit-score/

Jumaa, M., Saqib, M., & Attar, A. (2023). Improving credit risk assessment through deep learning-based consumer loan default prediction model. *International Journal of Finance & Banking Studies (2147-4486)*, *12*(1), 85–92. https://doi.org/10.20525/ijfbs.v12i1.2579

Krivorotov, G. (2023). Machine learning-based profit modeling for credit card underwriting - implications for credit risk. *Journal of Banking & Finance*, *149*, 106785. https://doi.org/10.1016/j.jbankfin.2023.106785

Li, Y. (2019). Credit Risk Prediction Based on Machine Learning Methods. *2019 14th International Conference on Computer Science & Education (ICCSE)*, 1011-1013.

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and Random Forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012042. https://doi.org/10.1088/1757-899x/1022/1/012042

Moradi, S., & Mokhatab Rafiei, F. (2019). A dynamic credit risk assessment model with data mining techniques: Evidence from Iranian banks. *Financial Innovation*, *5*(1). https://doi.org/10.1186/s40854-019-0121-9

Rao, C., Liu, Y. & Goh, M. Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model. *Complex Intell. Syst.* **9**, 1391–1414 (2023). https://doi.org/10.1007/s40747-022-00854-y

Schulz, M. (2023, February 23). *2023 credit card debt statistics*. LendingTree. Retrieved February 26, 2023, from https://www.lendingtree.com/credit-cards/credit-card-debt-statistics/

Vyas, P., Reisslein, M., Rimal, B. P., Vyas, G., Basyal, G. P., & Muzumdar, P. (2021). Automated classification of societal sentiments on Twitter with machine learning. IEEE Transactions on Technology and Society, 3(2), 100-110.

Vyas, P., Vyas, G., & Dhiman, G. (2023). RUemo—The Classification Framework for Russia-Ukraine War-Related Societal Emotions on Twitter through Machine Learning. Algorithms, 16(2), 69.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A comparative assessment of Credit Risk Model based on machine learning ——a case study of Bank Loan Data. *Procedia Computer Science*, *174*, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503–513. https://doi.org/10.1016/j.procs.2019.12.017