Josh Cubero

Loan Default Risk Predictor –

A Machine Learning-Based Approach

MSBA 307

Golden Gate University

Abstract

Credit risk is at the forefront of many firms in the financial services industry, which accounts for approximately 20-25% of the total world economy (Investopedia 2022). The global financial services sector is estimated to have reached $22.5 trillion USD in 2021, an increase of approximately 9.9%, which was hampered by the global pandemic. Accounting for such a large portion of the global economy, it's no surprise that the financial services industry was at the epicenter of the 2008 Financial Crisis, which has been analyzed and even portrayed in several Hollywood films.

An underlying cause of the 2008 Financial Crisis caused by a series of failures from insufficient government regulations to risky behavior by industry insiders (Insider, 2022). However, while debatable, at the heart of the financial crisis was the massive number of credit defaults. While companies such as Experian, Transunion, and Equifax report credit worthiness that enable lenders to make lending decisions, how does a company continue to monitor credit worthiness after issuing credit to the borrower? The purpose of this research is to conduct a feasibility study by creating a machine learning model that will monitor specified live data on borrowers and return a probability that the borrower could default. Ultimately, this model would be the engine that powered the application and would require APIs and pipelines to link the model to the data and the front-end user interface.

Problem

The credit rating companies Standard & Poor's and Experian collaborate monthly on a credit default index that reports default rates. These indices compile an aggregate measure of either increases or decreases in personal credit defaults, where the index increased two basis points in the month of June 2022. The defaults increase represented the seventh consecutive month of an increase in consumer credit defaults. Furthermore, bank card defaults increased by six basis points to 2.55%, while the auto loan default rate increased to 0.62%, and the mortgage default rate increased to 0.38%.

While Standard & Poor's and Experian have collaborated to monitor defaults, many banks today do not monitor borrowers' financial health. According to Moody's Analytics, banks have made little effort to monitor and predict defaults in advance. Moody's Analytics states that companies are under such great pressure to produce new accounts that much of the budget focuses on technologies that enable quick approvals and on-boarding. Banks, however, should consider protecting their investments through borrower monitoring. The purpose of this paper is to prove that firms can quickly and easily use Python programming language to implement a cost-effective borrower monitoring model.

Literature Review

The literature review is titled: *Modeling Money Attitudes to Predict Loan Default*, by Sunil Bhardwaj and Kaushik Bhattacharjee. The intent of the paper was to classify those who defaulted and those who have not defaulted. The interesting aspect of the literature review is that the attributes are qualitative as opposed to quantitative. Bhardwaj and Bhattacharjee sought out to prove that lenders can use personality traits, money attitude, and income dimensions to create a model that predicts the likelihood of borrower default.

Bhardwaj and Bhattacharjee stated that while it is broadly accepted that there is a need to predict potential defaulters, the models in use can be improved. Currently, many predictive default models use various quantitative and qualitative indicators to assess default. Bhardwaj and

Bhattacharjee argue that current models are flawed because of a reliance on widely used financial indicators. The authors argue that a default model should classify borrowers based on personality traits.

Bhardwaj and Bhattacharjee first used an adapted survey to collect data on borrowers that have defaulted and those that have not defaulted. The questions in the survey determined respondents' attitudes towards money. The survey identified four factors that would be used as inputs to the model: Power-Prestige, Retention Time, Distrust, and Anxiety. The data had been collected from a multinational corporation bank in India. The sample size is 501 respondents categorized by Gender, Age, and Annual Family Income.

Bhardwaj and Bhattacharjee study found that respondents who scored high on loan use, meaning, the respondent frequently used loans, tended to default at a higher rate. Additionally, the literature found that those that viewed high income as important, and those with anxiety toward money defaulted at a higher rate. Lastly, bank officials at the subject location found that higher income clients defaulted at a higher rate. Bhardwaj and Bhattacharjee concluded that anxiety towards money and income are most closely associated with default. The authors did note, however, that there are some limitations to their study. Most notably, the difficulty reaching the bank's entire population set. Next, the authors noted that some of the fields in the model could become obsolete with economic cycles. Lastly, the authors acknowledge sample size limitations and the effect on margin of error.

While this study, conducted by a team of researchers in India may pass the legal litmus test, one would be wise to use their approach with caution in the United States of America. The US has consumer protection laws such as the Equal Credit Opportunity Act, Fair Credit Reporting Act, and the Consumer Credit Protection Act that govern the way creditors assess credit worthiness. Granting or denying credit based on borrowers' attitudes towards finance could create unnecessary risk of litigation due to perceived discrimination.

Next, the literature review relies solely on data received from respondents to a survey. There are some advantages to survey research such as costs, practicality, and speed of collecting data, there are some disadvantages that would outweigh the costs. According to vittana.org, the number one disadvantage of survey research is the risk that respondents won't answer truthfully. Next, survey questions could get lost in translation, with respondents interpreting questions differently. Lastly, vittana.org states that surveys simply don't offer personalized questions.

Bhardwaj and Bhattacharjee's work is forward thinking and an example of thought leadership, the legal risks involved with assessing credit based on qualitative survey data outweigh any benefit received. This study will build on credit assessment fields commonly used in the US to create the machine learning foundation of a continuous credit health monitoring application.

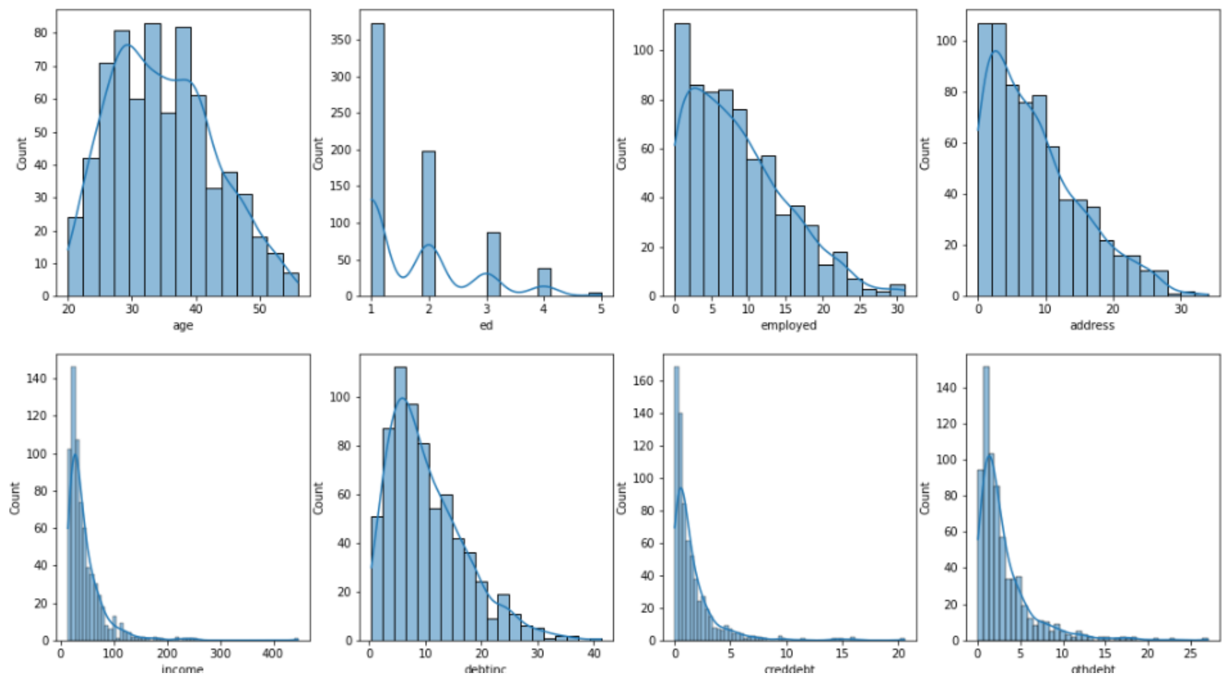
Dataset

The data set is comprised of eight independent variables and one binary dependent variable. Independent variables are:

1. Age
2. Education Level
3. Years Employed
4. Years at Address
5. Annual Income
6. Debt to Income Ratio
7. Credit Limit to Debt Ratio
8. Other Debt

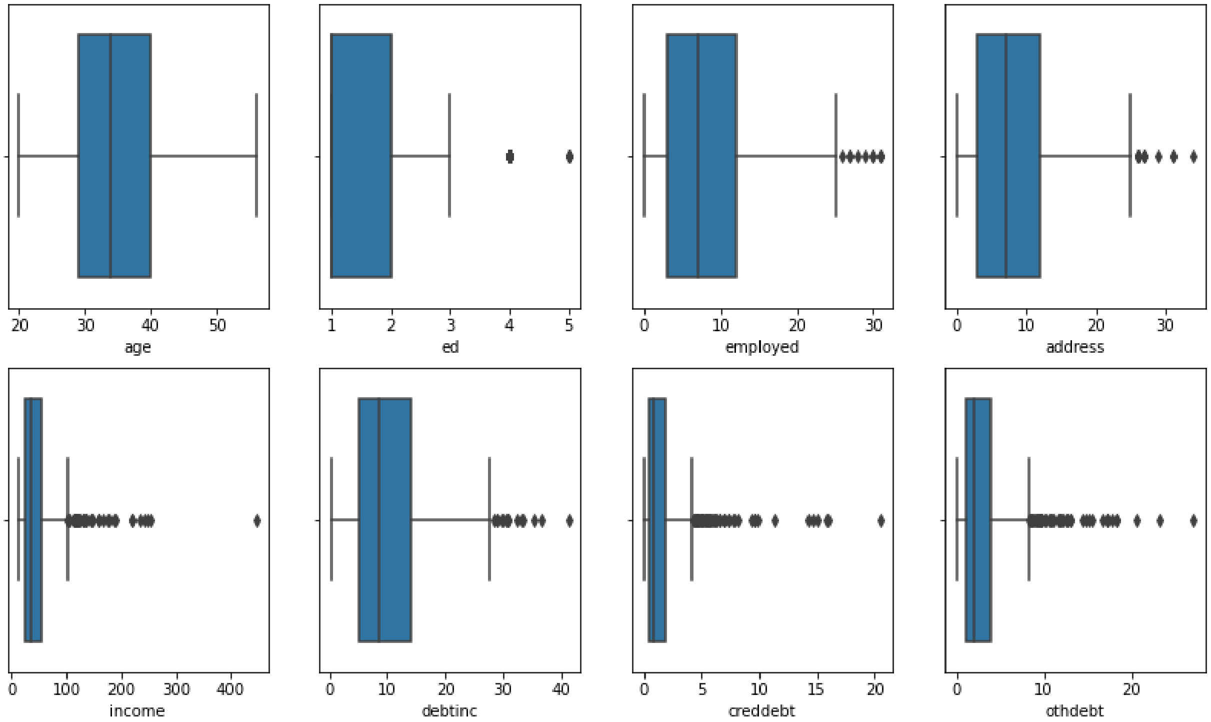
The dependent variable is Default, annotated by 1 or 0. The data is located on www.kaggle.com, and depicts various quantitative personal and financial metrics commonly used to assess a borrower's credit worthiness.

The exploratory data analysis started with a series of histograms depicting the distribution of the data.



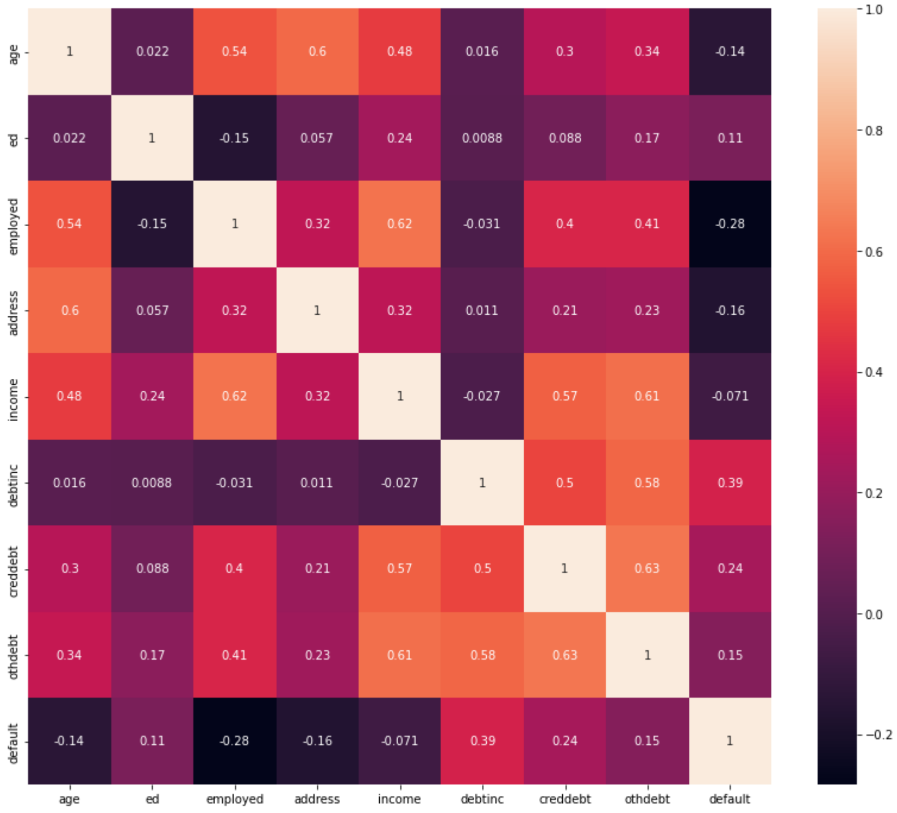
The histogram of the data set reveals that all eight of the independent variables are not of a normal distribution. Dependent variable Age is the closest to a normal distribution, but still falls short. Since the data is not normally distributed, one would need to employ a non-linear model.

Next, the boxplot depicts the dataset's quartiles, mean, and outliers.



Upon inspection of the dataset, it's clear that there are significant outliers in many of the fields. Optimally, one could eliminate the outliers, however the dataset is just 700 observations. With a dataset this size, it would not be advantageous to eliminate the outliers. The better option would be to scale the data.

The last item in the exploratory data analysis is the correlation matrix. The purpose of this correlation matrix is to gain a better understanding of the relationship between the independent variables and the dependent variable.



	age	ed	employed	address	income	debtinc	creddebt	othdebt	default
age	1	0.022	0.54	0.6	0.48	0.016	0.3	0.34	-0.14
ed	0.022	1	-0.15	0.057	0.24	0.0088	0.088	0.17	0.11
employed	0.54	-0.15	1	0.32	0.62	-0.031	0.4	0.41	-0.28
address	0.6	0.057	0.32	1	0.32	0.011	0.21	0.23	-0.16
income	0.48	0.24	0.62	0.32	1	-0.027	0.57	0.61	-0.071
debtinc	0.016	0.0088	-0.031	0.011	-0.027	1	0.5	0.58	0.39
creddebt	0.3	0.088	0.4	0.21	0.57	0.5	1	0.63	0.24
othdebt	0.34	0.17	0.41	0.23	0.61	0.58	0.63	1	0.15
default	-0.14	0.11	-0.28	-0.16	-0.071	0.39	0.24	0.15	1

The heatmap above depicts the correlation between all variables in the dataset. Of note, there's a negative correlation between default and Age, Years Employed, Years at Address, and Income. The negative correlation just described indicates that as Age, Years Employed, Years at Address, and Income increase, borrower Default decreases. An interesting correlation is that of Default vs Debt to Income Ratio. There exists a .39 correlation coefficient between Default and Debt to Income Ratio, the strongest in the entire dataset.

Methodology

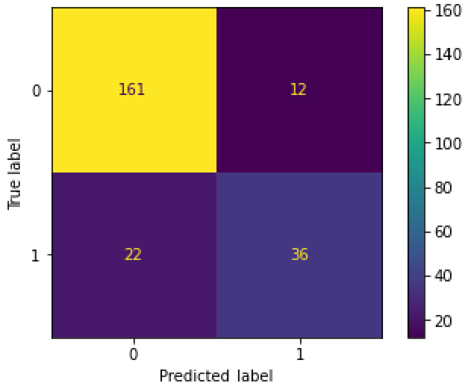
This study is a machine learning based approach to building a bank loan default predictive model. The Python programming language and Jupyter Notebooks is the platform of choice for analyzing and building the predictive model. Exploratory data analysis revealed interesting insights and guided the types of models used in the study.

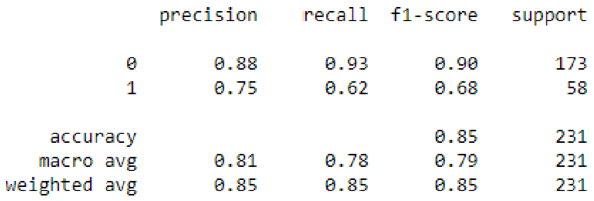
Next, various machine learning models were created with the idea of homing in on the most effective model. This process would start with the least complex model, then increase in complexity. Logistic Regression would be the first model employed, followed by a decision tree classifier, then a K-nearest neighbors model. The study then increased in complexity by implementing a K-fold cross-validation that included a support vector classifier, random forest classifier, adaboost classifier, and an entropy tree. Employing K-fold cross-validation enabled efficient comparison of multiple models.

Analysis

Logistic Regression

The first machine learning algorithm experiment conducted involved the use of Logistic Regression. The model was trained using train test split and evaluated using a confusion matrix.



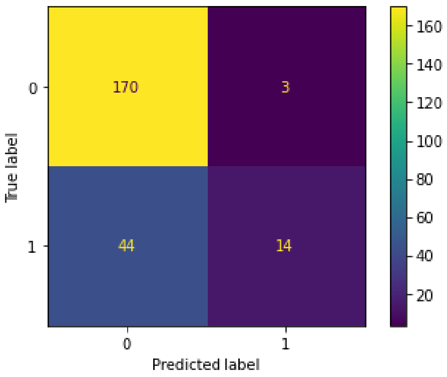


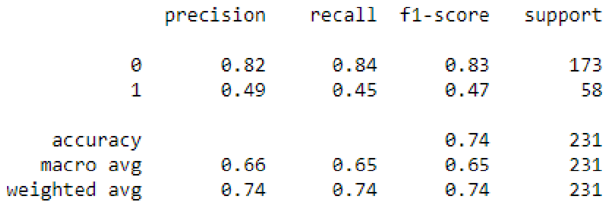
```
              precision    recall  f1-score   support

           0       0.88      0.93      0.90       173
           1       0.75      0.62      0.68        58

    accuracy                           0.85       231
   macro avg       0.81      0.78      0.79       231
weighted avg       0.85      0.85      0.85       231
```

The logistic regression model correctly predicted 161 true negatives, and 36 true positives, however returned 22 false negatives and eight false positives. Additionally, the model returned true precision of 0.75 and true recall of 0.93 and overall accuracy score of 0.85.

K-Nearest Neighbors

Next, the K-Nearest Neighbor (KNN) algorithm was fitted to the data. The KNN model returned the following confusion matrix and scores.



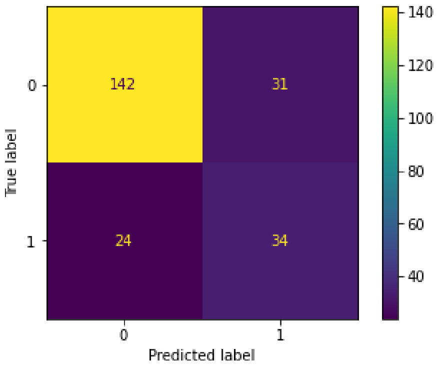


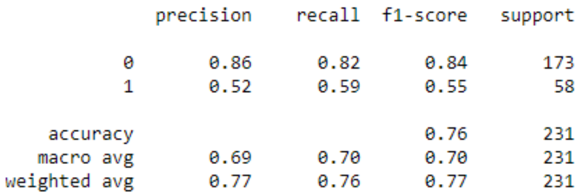
```
              precision    recall  f1-score   support

           0       0.82      0.84      0.83       173
           1       0.49      0.45      0.47        58

    accuracy                           0.74       231
   macro avg       0.66      0.65      0.65       231
weighted avg       0.74      0.74      0.74       231
```

The KNN model returned 44 false negatives and three false positives, while returning 14 true positives and 170 true negatives. This KNN model returned a negative precision of 0.82 and true precision of 0.49. The model's recall returned 0.84 negative recall and 0.45 positive recall. The model returned an overall accuracy of 0.74.

Decision Tree Classifier

The decision tree classifier model was fitted to the data, and the following visuals depict the confusion matrix and classification report.



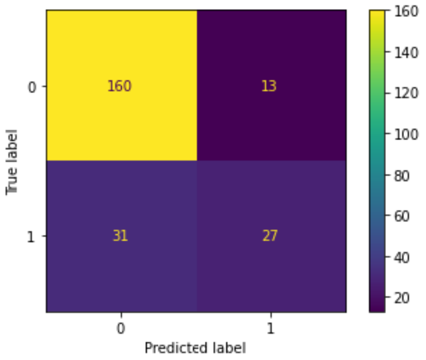


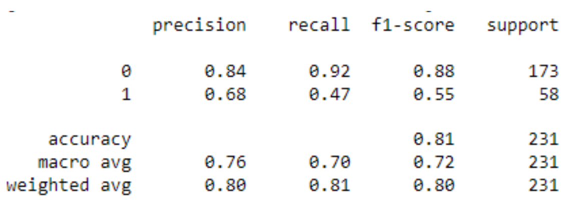
```
              precision    recall  f1-score   support

           0       0.86      0.82      0.84       173
           1       0.52      0.59      0.55        58

    accuracy                           0.76       231
   macro avg       0.69      0.70      0.70       231
weighted avg       0.77      0.76      0.77       231
```

The decision tree classifier was set to a maximum depth of four and returned a false precision of 0.86 and true precision of 0.52. The model returned a false recall of 0.82 and true recall of 0.59, and an overall 0.76 accuracy score.

AdaBoost Classifier

The adaboost classifier model was next implemented with the following confusion matrix and classification report.

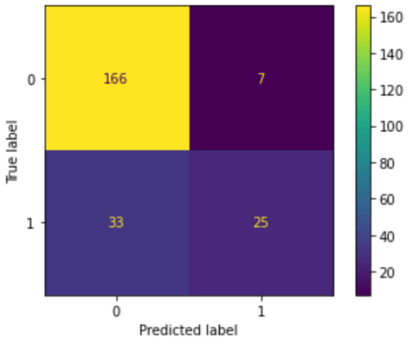


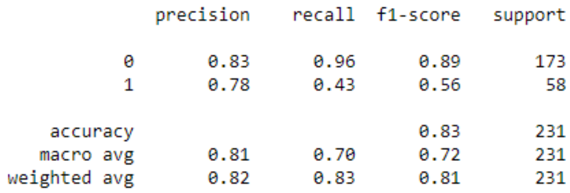

```
              precision    recall  f1-score   support

           0       0.84      0.92      0.88       173
           1       0.68      0.47      0.55        58

    accuracy                           0.81       231
   macro avg       0.76      0.70      0.72       231
weighted avg       0.80      0.81      0.80       231
```

The adaboost classifier returned false precision of 0.84, true precision of 0.68, false recall of 0.92, true recall of 0.47, and an overall accuracy score of 0.81.

Random Forest Classifier

The last model implemented in this study was the random forest classifier, which returned the following confusion matrix and classification report.




```
              precision    recall  f1-score   support

           0       0.83      0.96      0.89       173
           1       0.78      0.43      0.56        58

    accuracy                           0.83       231
   macro avg       0.81      0.70      0.72       231
weighted avg       0.82      0.83      0.81       231
```

The random forest classifier returned false precision of 0.83, true precision 0.78, false recall 0.96, true recall 0.43, and an overall accuracy score of 0.83.

Analysis of Results

This analysis will focus primarily on each model's precision, recall, and overall accuracy scores. The best performing model in this study was the logistic regression model. The logistic regression model outperformed all other models in the study with an overall accuracy score of 0.85. The logistic regression model had strong precision and recall scores as well. Although the logistic regression model's false recall score of 0.93 was not the highest of all models tested, the 0.62 true recall score was the highest. Recall is a very important score because it calculates the percentage of true positives. Depending on the model's use case, recall could be a determinant factor in the reliability of the model.

The least performing model in the study was the K-Nearest Neighbors model. The KNN model did not fit the data well and returned a true recall score of 0.36. This recall score would be far too low to implement with this data set, as tossing a coin would return a greater likelihood of prediction.

Conclusion

This paper conducted a literature review of research conducted on data derived from a multi-national based in India. The literature review proposed a new model for predicting creditworthiness that entailed qualitative factors derived from a survey. This study countered the idea of utilizing survey and qualitative data due to weaknesses in surveys and US consumer protection laws. Next, this study created a model that proved 85% accurate from data that was obtained from a public website. The model in this study would likely be strengthened even further with the addition of private customer data only accessible by the given financial institution such as credit score. Considering these factors, this paper proves that there is no need for creditors to away from quantitative data to qualitative data in default prediction models. Firms, however, should adopt a model like the model in this study and use machine learning to monitor customer data and predict the likelihood of default.

References

Bhardwaj, S., & Bhattacharjee, K. (2010). Modeling Money Attitudes to Predict Loan Default. IUP Journal of Bank Management, 9(1/2), 12–20.

Credit Risk Analysis. (n.d.). Www.kaggle.com. Retrieved May 16, 2022, from https://www.kaggle.com/datasets/karanagarwal/bankloans?select=bankloans.csv

Fair Lending. (2019, April 6). Www.occ.treas.gov. https://www.occ.treas.gov/topics/consumers-and-communities/consumer-protection/fair-lending/index-fair-lending.html

Indices, S. D. J. (n.d.). *S&P/Experian Consumer Credit Default Indices Show Third Straight Increase In Composite Rate In February 2022.* Www.prnewswire.com. Retrieved May 16, 2022, from https://www.prnewswire.com/news-releases/spexperian-consumer-credit-default-indices-show-third-straight-increase-in-composite-rate-in-february-2022-301503065.html

Board of Governors of the Federal Reserve System (US). (1985, January 1). *Delinquency Rate on All Loans, All Commercial Banks*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DRALACBN

What percentage of the global economy is comprised of the financial services sector? (2020). Investopedia. https://www.investopedia.com/ask/answers/030515/what-percentage-global-economy-comprised-financial-services-sector.asp

Kimberly Amadeo. (2014). *2008 Financial Crisis Timeline*. The Balance. https://www.thebalance.com/2008-financial-crisis-timeline-3305540

Field, A. (n.d.). *What caused the Great Recession? Understanding the key factors that led to one of the worst economic downturns in US history*. Business Insider. https://www.businessinsider.com/personal-finance/what-caused-the-great-recession#:~:text=The%20collapse%20of%20the%20housing

Redefining loan monitoring through an integrated solution. (n.d.). Www.moodysanalytics.com. https://www.moodysanalytics.com/articles/2018/redefining-loan-monitoring

Gaille, L. (2020, January 27). *20 Advantages and Disadvantages of Survey Research*. Vittana. https://vittana.org/20-advantages-and-disadvantages-of-survey-research