Default Probability Prediction by Explainable Machine Learning for Small and Medium-Sized Enterprises

Aalborg University Business School

Economic and Business Administration

Master of Finance

4th Semester

Due to: 3rd June 2024

Supervisor

Professor Cesario Mateus

Student

Mehdi Shadpour

20221764

Number of pages: 55
Number of stokes: 85330

Thanks to:

The entire faculty of finance at AAU
especially my supervisor Prof. Mateus whose support
made this research to be done

and

Joseph & Patryk
for standing alongside, working incessantly and
for what we learned and experienced together

Table of Contents

Abstract

In this paper it is attempted to assess the credit risk of small and medium-sized enterprises (SMEs) via explainable machine learning through different machine learning models. Although SMEs are vital for economic growth and so lending loan to them, their credit worthiness evaluation is a burdensome and complex task. Nowadays, Artificial Intelligence (AI) is vastly used by finance sector to provide accurate models especially for credit risk of loan borrowers. By the same trend in this research, the three groups of SMEs in USA which have received the aid program during Covid pandemic (Payroll Protection Program) were examined for forecasting their default probability.

Both machine learning algorithm models and explainable AI (XAI) were hired to construct a precise model for predicting each SME's default probability. Despite returning accurate result by using machine learning models such as eXtreme Gradient Boost (XGB), Logistic Regression (LR), Support Vector Machine (SVM), these models cannot explain the importance of each feature in the default prediction decision. However, by using XAI such as Shap and LIME, besides more accurate results, the importance and effect of each parameter is observable and can be employed to interpret the model's decision. The research findings proved that XAI models are more accurate, transparent and comprehensive which could assist the financial decision makers to efficiently predict SMEs default probability.

Introduction

Determining the credit risk of small and medium size businesses, has never been an easy task. There are plenty of reasons for this difficult burden such as not having access to the background of them, no access to the historical financial data of them in usual sources, not enough and organized financial statement, less and sometimes no availability of their business data, etc. which in fact that is an endless list. On the other hand, these entities are playing a vital role in the economic systems which is not negligible, hence funding them in time is vital as well. The prepollent of the engine of production, creativity, innovation and breaking the monopoly markets, comes from them.

Everyday many startups come to existence and for implementing their idea they seek loans from financial institutions. Besides them, many small businesses from the past need loan to pass the recession and come back to the productivity. As the result, the number of potential borrowers is enormous. The correct credit worthiness assessment system which can evaluate these fund seekers fast and precisely, is vital for the lenders too. Governments, banks and the other financial institutions which are supposed to pump credit into this sector, all facing limited resources that by not reliable assessment can jeopardize their own existence. Many default borrowers would easily lead to bankruptcy or liquidation of the lender itself. Therefore, there is a tough competition especially between banks for using a reliable model to evaluate the credit risk of these borrowers. If they can apply such a system, not only they reduce the number of defaulted customers but also will manage to make a lucrative business as the small firms' growth rate can be unbelievably high, and hence overall economic growth. Considering all what has been said, makes the decision for small businesses loan approval a breath-taking task.

So far, many financial institutes have developed different models to assist the credit assessment of SMEs. With the emerge of Artificial Intelligence (AI), many of these institutes have invested heavily to create a reliable model for such task. In the literature review, we will see some recent research that are conducted in this area. The number of similar jobs in the commercial sector is estimated much more than academic sector.

Here the machine learning is used to assess SMEs who received Paycheck Protection Program (PPP) loan in the United States during Covid pandemic period. The federal government provided credit for small businesses, an incentive scheme, to pay their employees payroll and retain their jobs. It was reported that 89.6 million jobs were retained by the program. The data provided by Small Businesses Administration (SBA) in US, a government agency, is divided PPP loans into three categories: the PPP loans below $150,000, the PPP loans from $150,000 to $1,000,000 and finally $1,000,000 to $10,000,000 which were lent by financial institutions across all states (data and its dictionary is available in: https://www.pandemicoversight.gov).

During pandemic period, businesses faced adversity and needed support by governments to keep themselves upright. As mentioned before the necessity of their existence, made the governments across the globe to allocate budget to them for passing this period while most of the harshly affected by lockdowns as well. According to the SBA in the US report published in October 2023, totally 11.5M loans were given which accounts for $792.6B for all states that led to retain 89.6M jobs. The loans were guaranteed by the Department of Treasury and asked the businesses to keep up to 8 weeks of payroll costs including benefits by small businesses to retain their jobs (home.treasury.gov). If a business managed to do so, i.e. retain the jobs until a certain period, then there was an opportunity for it to request on forgiveness of the loan. Maybe at the time it looked easy duty, but now we know that the pandemic extension made it so hard, and many businesses could not survive. On the other hand, for federal government $792B amount to lend again and again to retain the jobs was impossible too. As the result many of these small businesses default on the loans which are presented on the following parts.

Seven kinds of machine learning algorithm were hired to find out which one(s) can forecast whether the business defaults or pays the loan back. Extreme Gradient Boosting, Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Trees, Random Forest and K-Nearest Neighborhood are the algorithms which are explained in the analytic section. XGB and LR showed the highest accuracy to predict default situation and were used in XAI for model interpretation and actual results evaluation.

The Covid pandemic period has gone, but the assessed data and the results could be useful in the similar situation in the future as crisis rise again and authorities should optimize spending of limited budget to save the jobs and businesses. In the following sections we will see how

these three groups managed their loan in the end. The loan status is the variable that is taken as the dependent variable and has two objects, Charged Off or Paid in Full. These are equal to default or not-default. The other variables of each group are considered as the independent variables, affect the loan status registered by the business at the end. In the next section several research in this area is presented (Please note that throughout the entire text, it is attempted to be clear and quality of the content is prioritized to its quantity. For further information in any section, please refer to mentioned supplementary materials).

Literature Review

Although using AI in different sectors to optimize processes is yesterday news, this trend very recently has accelerated thanks to the new technology and advanced processors such more powerful Graphic Unit Processor (GPU) and stronger Central Unit processing (CPU) which have provided analysis of big data ubiquitously. In the recent years countless number of researchers and companies published their findings by AI models in the financial area. Here we review several of them.

Yang Lu et.al. created a novel framework of credit risk features for SMEs by using optimization algorithms and seven machine learning classifiers. As they claim, though there is no universal model for SMEs default risk as they vary a lot, their model improves the prediction of the SMEs default risk (Lu, Yang. June 2022. A novel framework of credit risk feature selection for SMEs during industry 4.0. Springer Nature). They suggest that improving their model will benefit SMEs to know about their weaknesses in the assessment process by financial creditors and empower themselves to enhance them. In an emprical study Kryzanowski, revealed that the usual ratios asked by lenders such as business age and total asset of the firm are not significant in loan default by the firms (Kryzanowski. 1985. Small businesses debt finance: An emprical investigation of default risk). Lan H. Nguyen and Megumi Sagara in their article Credit Risk Database for SMEs financial Inclusion, exploited machine learning to process extremely large body data transaction for SMEs in Japan. They assert that the best indicators of the risk default probability are cash balance and cash outflow related to repayment. In

addition they claim that their machine learning model outperforms logistic models both for short and longterm default risk prediction. They believe their model can assess credit risk SMEs without financial statements (Nguyan, Lan. April 2020. Asian development bank). Paolo Giudici et al. in their research 'Artificial Intelligence Risk Measurment', proposed the first Key AI Risk Indicator (KAIRI) model. They considered the set of four principles (Sustainability, Accuracy, Fairness, Explainability), required by European regulatgory instututes, to develop model for measuring AI risk for constructing a framework to effectively measure AI risk and thereby promoting a safe and thrustworthy AI in finance. Galluci et.al. published their article by which they made a model to predict the SMEs default risk by assessing 973 Italian firms. They employed a Bayesian approach to predict the SMEs' default status (Galluci, Carmen. 2023. Financial ratios, corporate governance and bank-firm information. Journal of Management and Governance).

Bitteto et.al. developed two models for estimate small businesses default risk in Italy. A classic parametric approach and a nonparametric approach by using machine learning historical random forest model (HRF). After comparing the results of the two models, they claim that the HRF model outperform the parametric model and can successfully estimate the SMEs default risk (Bitteto, Alessandro. 2023. Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses. Socio economic planning science. Elsevier). In another research by Chelagat, SMEs loans in the Nairobi were assessed. He suggests that there is a relationship between the age of the business and interset rates with the default risk. Poor credit analysis, economic situation and repayment period are the parameters that contribute to loan default for SMEs (Chelagat, Kibosia Naomi. 2012. Determinants of Loan Default for SMEs amongst commercial Banks in Kenya. University of Kenyia). In the other work, Mtenda and Sibanda used logistic regression models to detect and explain the determinants of default probabilty for unaudited and audited SMEs under distresses condition in Zimbawe.In their study they found out that the DP factors for unaudited and audited firms are not the same. Also they said that the macroeconmic condition has an important effect on the rate of SMEs default (Ranganati Matenda, Frank. Sibanda, Mabutho. 2022. Determinants of default probability for audited and unaudited SMEs under stressed condition in Zimbabwe. Economies. Switzerland).

Bussmann et.al. in their research examined 15000 small and medium size businesses by machine learning to measure their credit risk. They claimed that their model which is using Shapley Values can divide both risky and not risky borrowers into group by their similiar financial characteristics and explain the credit risk score which can predict the SMEs future behaviour (Bussemann, Niklas. Gidici, Paolo. 2020. Explainable Machine Learning in Risk Management. Computational economies). In an intensive literature review for SMEs, Ciampi et.al. reviewed the default risk prediction for SMEs over the 34 years period, from 1986 to 2019. They state that using modern analytical techniques such as artificial intelligence will improve the modelized default risk prediction for SMEs (Ciampi, Francesco. Giannozi, Alessandro. 2021. Rethinking SMEs default prediction: A systematic litrature review and future perspective. Scientometric). Gogas and Papadimitriou, researched the usage and its extention of AI and machine learning in finance and economics. They named the models observed in the literature which are vastly used AI models such as Shapley Values and Gradient Boosting to predict default risk (Gogas, Priklis. Papadimtriou, Theophilos. 2021. Machine learning in economics and finance. Computational economics).

Aniceto et.al. by collecting the loan performance from commercial banks in Brazil, evaluate the performance of different machine learning models. They findings suggest that Random Forest and Adaboost have more accuracy compared to the other models. They also stated that Support Vector Machine (SVM) is performing poorly with both linear and non-linea kernel (Aniceto, Masia Cardoso. barboroza, Flavio. June 2020. Machine learning predictivity applied to consumers creditworthiness, Future business journal). In the other work by Huang et.al. the authors analyzed 1.8million loan transactions of Chinese leading online banking firms and suggested that Fintech firms can handle big data appropriately to create prediction models by machine learning to help the financial institutions avoid huge losses (Hyuang, Yiping. Zhang, Longmai. Sep 2020. Fintech credit risk assessment for SMEs: Evidence from China, IMF working paper). Kyeong and Shin developed a two stages regression model based on Bayesian approach. Their model enhanced the performance of credit scoring model and interpretabilty of that, according to the authors (Kyeong, Sunghyon. Shin, Jinho. 2022. Two stage credit scoring using Bayesian approach, Journal of big data). Heng and Subramanian in the recent work focused on how XAI can improve the credit worthiness of machine learning models. They also review some of the available software in the market for cedit assessment. They believe,

although some complex models can performe an accurate prediction, lack of interpretability for user is a weakness for them. XIA moves these models to the interpretable and shed a light over them, which is demanded crucially by the lagurators when they check the fairness of the financial instututes in bestowing the loans to the customers (Sheng Heng, Yi. Subramanian, Preethi. Oct 2023. A systematic review of machine learning and explainable artificial intelligence in credit risk modeling. Springer). Rudin and Shaposhnik developed a predictive credit risk by construction a global models based on the local models for each specific observation. Unlike the usual models which explain the local solutions, their model was constructed to explain the past data globally consistently. They designed multiple algorithms to extract discrete and continuous dataset to study the theoritical properties of datasets (Rudin, Sinthia. Shaposhnik, Yaron. 2023. Globally consistent rull based summary explainations for machine learning models: Applications to credit risk evaluation. Journal of machine learning research). There also many similar works either interdisciplinary or specific fields which have developed same approach to predict the outcome. For example, Feng and Shen used machine learning models to develop the prediction of Shizophernia by genes. As they say Programmed Cell Death (PCD) play a role in immune system and cause many diseases, especially Shizophernia. They used machine learning models to obsere this indicator of the disease (Feng, Yu. Shen, Zhing. 2023. Machine learning based predictive models and drug prediction for Shizophernia in multiple programmed cell death patterns. National library of medicine. USA). Biecek and Burzykowski write in their book that today for predicitve models there is no lack of data, niether algorithms, nor models, but lack of expolaration, explaination and examination (Biecek, Przemyslaw. Burzykowski, Tomasz. 2021. Explanatory model analysis: Explore, explain and examine predictive models. 1st edition. CRC press).

Jammalamadaka and his colleague worked on the German credit cards dataset to develop a fair credit scoring model which gives response to the expectations. By using machine learning modeling approach they tried to reduce the share of features in missclassification. They claimed that XGBoost model decreases mismatch and improves fairness and accuracy of scoring system. For both age and gender legal permission levels, the variable in the credit scoring system, AI assisted them in their model to find the optimum threshold of fiarness metrics (Jammalamadaka, Krishna. 2023. Responsible AI in automated credit scoring system. AI and ethics). In another study, Madarres et.al. investigated deep machine learning (DPM) in credit

risk scoring. They believed that lack of model interpretability is a hinderance for financial institutes to to use DPM. In order to shed a light in this field they assessed neural networks usage in DPM (Modarres, Ceena. Louie, Melissa. 2018. Towards expalinable deep learning for credit lending: A case study. Capital One). In an intensive series of studies, Zoynul Abdin et.al. looked into application of machine learning in different financial sectors. These studies included, financial risk management, corporate bankruptcy prediction, portfolio management, and stock price prediction. They employed machine learning to tacke business risk and uncertainty which are the main concerns of financial institutes (Zoynul Abedin, Mohammad. Hassan, Kabir. 2021. Essential of machine learning in finance and accounting, Routledge publications).

Methodology

Credit risk assessment is essential for the financial institutions to accept the borrower request and lend a loan. The main attention for the lender is to know the probability of default (PD) on the loan. PD is defined as the inability of the borrower to payback a portion or the entire borrowed loan.

By using logistic regression model, probability of default is calculated by:

PD=1/(1+e^(-z))

Where:
PD: Probability of default
e: The Euler number
z: The linear characteristics of the borrower entity and the correspond coefficients

And for z:

$$z = \alpha + \sum_{j=1}^{n} \beta_i \, x_j$$

and here alpha is the intercept and betas are the regression coefficient of each element for each entity showed by x. Once the intercept and coefficients are estimated we can find the probability of default.

Machine Learning

Machine learning is all kind of methods that computer uses to predict a trend or improve a model based on the input data. Usually, a set of large data is needed to get a reliable output. For machine learning purpose, the input data is divided by the operator into training and test set. The training set is a portion that will feed to the computer to observe the patterns and the test set is the part that computer is applying the prediction model for checking whether its predictions comply with the actual data. More match the predicted data by computer to the actual data, more accurate the model would be. Machine learning models can be divided into two general categories: not interpretable and interpretable. Models like deep learning and gradient boosting are considered to be not interpretable (Blackbox AI). All processes, calculations and the importance of factors in blackbox models are done by computers behind the scenes and human being cannot observe them. It is vague for human, in which direction model elements got more weight and why they got that weight. This is problematic for the financial decision makers when they want to present their credit scoring decision to the potential borrowers, regulators and authorities.

On the other hand, the interpretable models provide a good provision and understanding for human. Shap (given from Shapley Values) and LIME are two packages which visualize the result with the importance of each parameter in the final model. Shapley values model, which is named in the honor of Lloyd Shapley and its work for the Game Theory, defines the contribution of each player (factor/feature) in the game (process or model). For example, if we have three persons with three impaired gloves (person one has only the left hand glove, person 2 has also the left hand and person 3 has the right hand of the pair) and want to know the contribution of each for making a pair of gloves then we have: person 1 contribution 25%, person 2 contribution 25% and person 3 contribution 50% as he could make a pair with person 1 or 2. SHAP package in Python programming language, which stands for Shapley Additive exPlanations, is an interpretability method based on Shapley values and was introduced

by Lundberg and Lee (2017) to explain individual predictions of any machine learning model. When it comes to explaining complex model such as ensemble methods or deep networks, usually simpler local explanation models that are an interpretable approximation of the original model are used. In SHAP, this explanation model is represented by a linear model — an additive feature attribution method — or just the summation of present features in the coalition game. SHAP also offers alternatives to estimating Shapley Values (c3.ai/glossary/data-science/shapley-values). The general formula introduced by Lloyd Shapley in 1951 is:

$$\varphi_m(v) = \frac{1}{p}\sum_{s}\frac{[v(s \cup \{m\}) - v(s)]}{\binom{p-1}{k(s)}}\ , m = 1,2,3, \dots, p$$

where:

$\varphi_m = Shapley\ Value$

$\varphi_m(v) = the\ fair\ share\ of\ each\ member\ in\ the\ team$

$m = team\ members$

$s = team\ subset$

$k(s) = the\ size\ of\ subset$

$v(s) = value\ achieved\ by\ subteam$

and $v(s \cup \{m\}) = the\ realized\ value\ after\ m\ joined\ S$

The total individual value is equal to the team value:

$$\sum_{m=1}^{p} \varphi_m(v) = v(T)$$

The other package LIME, Local Interpretable Model-agnostic Explanation, provides the interpretable of each factor role in the local final model. In LIME model aim is to minimize the 'loss function' while we are looking for the locally explainable model for the blackbox model f( ) around the instance of the interest factor. The loss function will be as below:

$$\hat{g} = \arg\min_{g \in \mathcal{G}} L\{f, g, v(\underline{x})\} + \Omega(g)$$

where model g belongs to class $\mathcal{G}$, and $v(\underline{x})$, defines a neighborhood of the $(\underline{x})$ in which approximation is seen, L is the function of the discrepancy of models f( ) and g( ) in the neighborhood of $v(\underline{x})$ and $\Omega(g)$ is the penalty for the complexity of model g( ). The penalty is more when the model g( ) is more complicated than class $\mathcal{G}$ (ema.drwhy.ai/LIME.html).

Both packages are model agnostic which means that regardless of the kind model it can be applied to it and can study the underlying of the model's structure without assuming that it can be accurately described by of the model. By this, the bias interpretation will be avoided.

Models Accuracy

In finance industry a little deviation in calculation could lead to huge and irreversible losses. Hence the accuracy of any model even with small decimals is perceived as an improvement. In this research different models and their accuracies were checked. The number of false predictions for both defaulted and not defaulted businesses are calculated and shown by figures in the relevant parts. This approach is a vivid approach, number of false predicted default companies which did not default and the number of false predicted of not-default firms who were defaulted in the actual dataset. Also, the number of correct predictions belong to default predicted businesses who defaulted in real, not default predicted companies and in fact they paid in full. Then the numbers in each category and its total numbers are shown in the heatmaps. Finally, the models with the most accuracy percentage were selected to feed the Shap and LIME packages to interpret and explain each feature share in the decision to default probability prediction by models.

Data

All data in this paper are taken from Small Businesses Administration (SBA) organization in USA. The datasets are pretty huge (categorized into Mega database), to some extent that even in this website for loans below $150K, datasets are divided into 12 files (each database includes 900000 firms for a few states as for example it cannot be presented all in once in an excel

sheet). After downloading the entire data, there were many missing data which removed from the database to not drive to miscalculations. In addition, the firms marked as Exemption4 replaced by Charged off (as default firms) due to definition by what is considered as Exemption4 by SBA as well as making the dataset coherent. Hereby the firms coherently are categorized as Charged Off (default) or Paid in Full (Not Default). Moreover, many variables such as the names of the companies, the loan guarantee percentage (that stands 100% for all companies), not using zip codes but cities and states and so forth, were removed from database to avoid deviation in the machine learning calculations. The dictionary of the data is available in the appendix section.

Paycheck Protection Program (PPP)

The Paycheck Protection Program (PPP) was a part of the Covid pandemic relief aid program, was approved by the Congress to help small businesses retain their jobs and continue working during the pandemic time. This program initially granted SBA to spend money on eligible small businesses. The rules of the program were clear, at least 60% of the loan must be paid for the payroll. The rest could be spent on other expenses of the business such as mortgage interest, liability interest, utilities, etc. If the business managed to retain their jobs for at least two months (8 weeks), depends on field of the business, it could apply for the loan forgiveness (the application must be handed in less than 10 months after receiving the loan). There was no need for personal guarantee or personal collateral for the loan, the interest rate was 1% annually, non-compounding, non-adjustable, and lenders could rely on the borrowers certificate to assess their eligibility required by delegated authorities. The lending happened in the two periods in both 2020 and 2021. If a firm had missed the first period in each year, it could withdraw the loan in the second announced period in the same year. Some businesses were entitled to receive the loan more than once (www.sba.gov/funding-programs/loans/covid-19-relief-options/paycheck-protection-program).

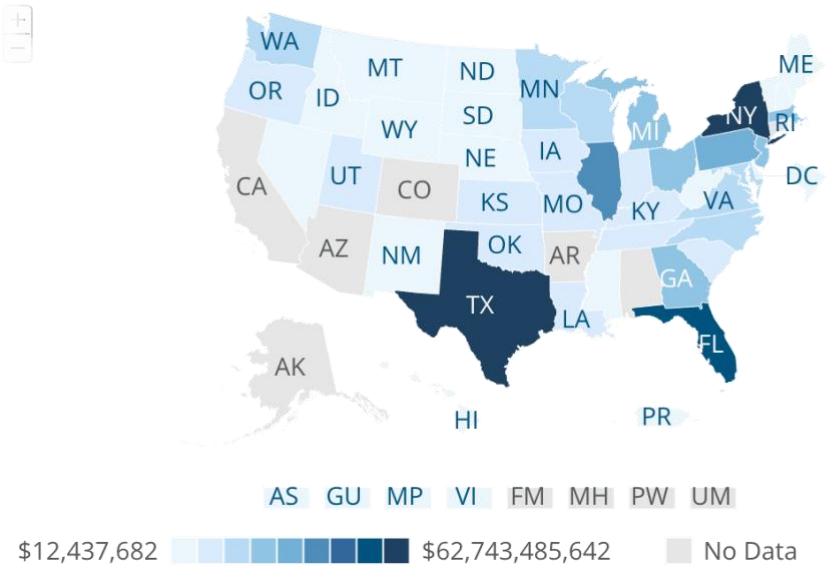


Figure 1. Total spent PPP loans by state

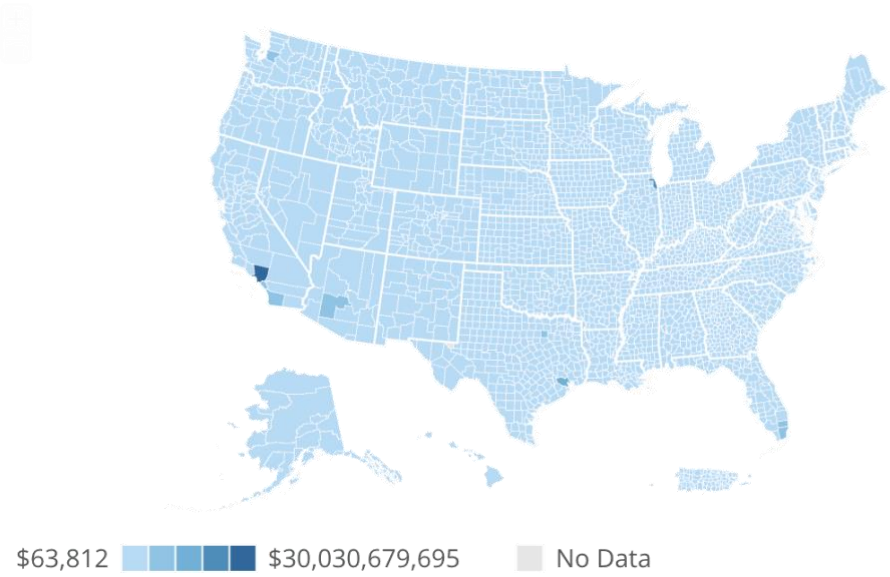


Figure 2. Total spent PPP loans by county

SBA in US has a definition for what is small business, the standard definition is the number of employees less than 500, and the annual income up to $7.5M. However, there are some exceptions, that mostly are defined by the industry in which the business works. North American Industry Classification System (NAICS) codes provide details on each industry and the norms for small or not small businesses.

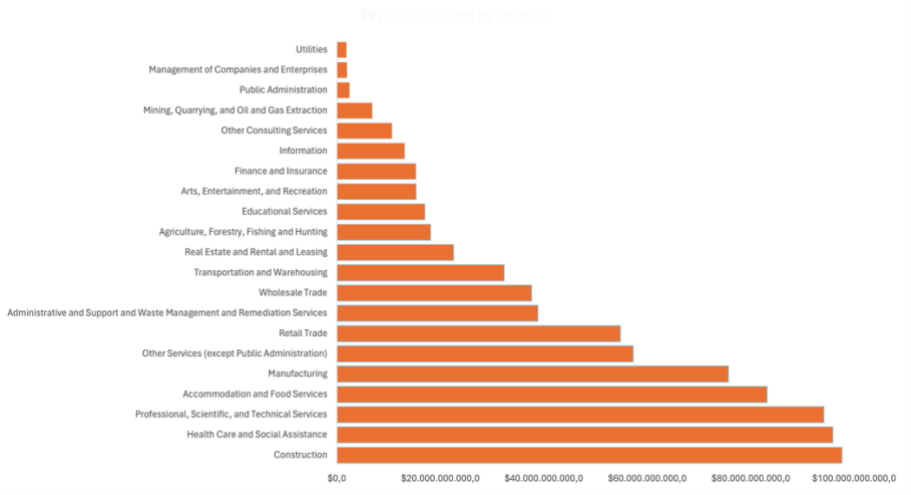


Figure 3. Amount of PPP loans lent to borrowers by industry

Totally, according to the SBA, 11.5M loans were lent which accounts for $792.6B in sum amount. To visualize the scope, some figures are shown in the following:

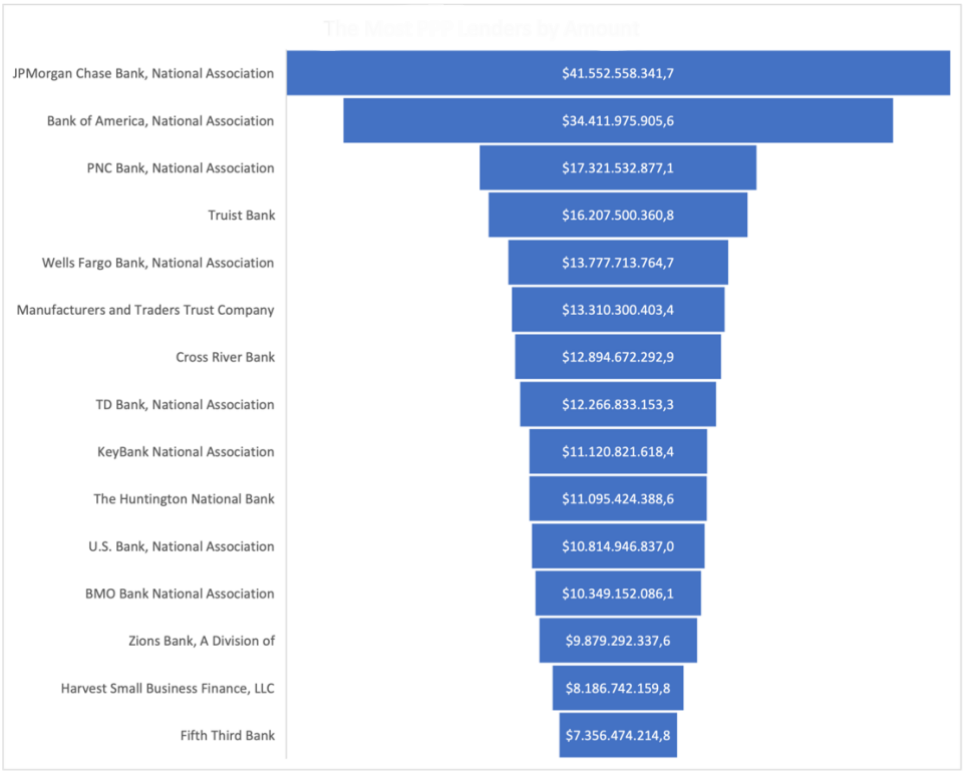


Figure 4. The Most PPP Loan Lender by Amount

Results

For analysis, the total available datasets were investigated by the three categories which are accessible in PPP oversight website. The businesses who borrowed up to $150K (the most businesses, almost 11M borrowers are registered in this category), the firm which borrowed from $150K to $1M and the firms who borrowed from $1M to $10M.

To perform a fair comparison and dig more into which businesses defaulted the loan under the similar condition, the narrowest range of loan amount in the largest group was probed (i.e. up to $150K borrowers). It revealed that a huge portion of the borrowers needed $21000 dollar (precisely $20800) across the US to keep the business upright. Therefore, these borrowers were chosen to get analyzed and discover a pattern for influence of each parameter to success or failure to repay the loan and from now forward, the smallest group is confined from $0~150K to $20,8K (still groups are mentioned with the loan amount of borrowed to avoid confusion).

Dataset Overall View

The variables are not exactly the same in all three groups. For example, in $1M to $10M, there is no term in the database. However, the most variables are the same in all groups. In the mentioned dataset that is for bigger companies, the average of employees of the firm is 176, and the average loan amount equals to $2.12M. In the dataset there are 100,000 companies which many of them have 500 employees. As expected, the rate of default among these businesses is not high and stands for 3.06% of the total loans.

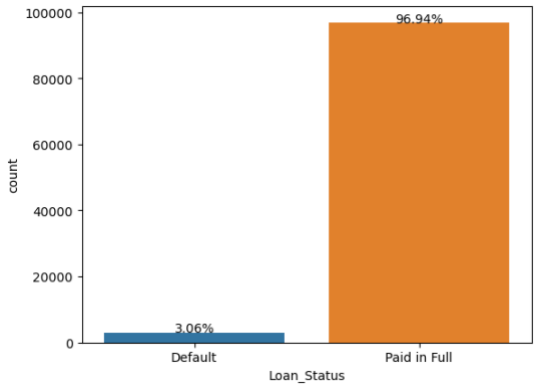


Figure 5. Default Rate for $1 to $10M Borrowers

For the firm which borrowed from $150K up to $1M, after consolidating the dataset, the number of the firms confined to 130,572. The average of employees, the term of borrowing, and the loan amount, in this group are reported 37, 40 months, and $324.27K, respectively. The default rate in this category is slightly more than former one and reaches to 3.56% of all the loans. The most borrowers here agreed to payback the loan in 24 months. In the following figure, the variables and their correlations for this group is depicted (please note that term and processing method high correlation is due to encoding processing method into numbers 0 and 1 by computer and does not mean that number of loan term installments goes hand in hand with first or second group of borrowers).

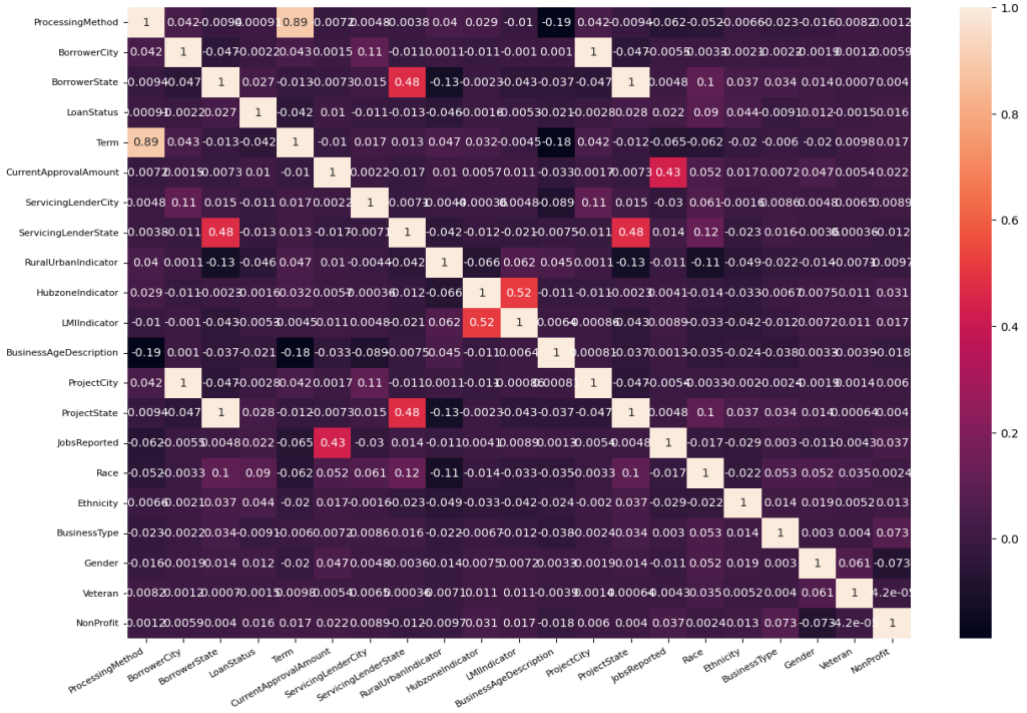


Figure 6. The features and their correlations for the $150K~1M PPP loan borrowers

The scenario for the borrowers who got up to $150K was a bit different. Firstly, the rate of default is quite high and gets to 9.07%. The following table shows the primary statistics about this group.

Table 1. The primary statistics of the borrowers up to $150K

	Term	**Approval Amount ($)**	**Jobs Reported**
count	1038845	1038845	1038845
mean	49.60	39937.55	5.13
std	15.97	30812.28	9.98

min	6.0	20000.0	0.0
25%	24.0	20832.0	1.0
50%	60.0	20833.33	1.0
75%	60.0	48587.5	7.0
max	180.0	149999	500.0

As mentioned before in this category, all $20.8K amount loan borrowers were selected to assess the default probability be machine learning. The default rate for this group of borrowers is much more than the other groups:

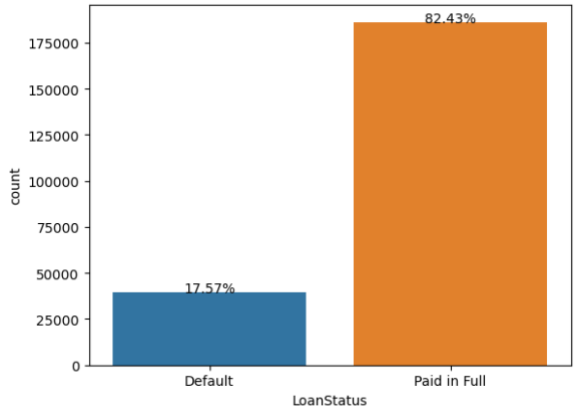


Figure 7. Default rate among $20.8K PPP loan borrowers in across the USA

By a glance at the default rate charts for all three group, we can vividly see that smaller businesses get, default rate of the loan gets more.

Data Analytics

Machine Learning Algorithms

Extreme Gradient Boosting (XGB)

As mentioned in the literature part, eXtreme Gradient Boosting or briefly XGB is a powerful tool to predict or improve models. Here also XGB was one the main model which hired for separate analysis and XAI models analysis as well. Amongst all machine learning algorithms, XGB showed one the highest accurate prediction ability for all three group of datasets. For

$1~10M the accuracy of XGB was 96.72%. According to this model, the city which borrowers are coming from, and the loan amount were the most important features to predict the firm default risk while how long is business have been to existence and race of the business owner were the least. Table 2 shows that variables and their importance in this group:

Table 2. Features and their importance to predict the firm's default by XGB

Feature	**Importance%**
BorrowerCity	16.389245
Loan_Amount	16.286812
Lender_Name	15.057618
Jobs_Reported	13.674776
Industry_Detailed	13.085787
Borrower_State	8.271447
Industry	7.195903
Business_Type	3.457106
Date_Approved	2.304738
Gender	1.587708
Business_Age	1.434059
Race	1.254802

In the second group of data, XGB algorithm accuracy hits 98.23% and by this method, how much is borrowed and which city the borrower comes from, are the most important features. The location of the lender, in fact the bank, is mattered to predict the default or not default businesses for the algorithm here. If the company is nonprofit or for profit, located in the rural or urban area, the owner is veteran or not, his/her ethnicity, cannot say much about the default prediction. The whole features and their importances are depicted in the following table:

Table 3. Features importance in XGB predicting for 150K plus borrowers

Feature	Importance%
Approval Amount	17.81
Lender City	13.61
Borrowe City	13.61
Jobs Reported	13.25
Term	8.06
Borrower State	7.34
Lender State	6.84
Project City	4.08
Business Type	3.20
Race	3.20
Processing Method	1.79
LMI Indicator	1.38
Gender	1.10
Hubzone Indicator	1.07
Ethnicity	0.85
BusinessAge	0.74
Veteran	0.55
Project State	0.52
RuralUrbanI ndicator	0.52
NonProfit	0.38

So far, XGB demonstrated high accuracy in prediction, however, for the last group the rate is not as high as it was for the other groups. For the smallest borrowers the algorithm presented 87.2% precision.

Table 4. Features importance for $20.8K borrowers predicted by XGB

Feature	Importance%
Date Approved	3.095421
Gender	3.044257
Hub-zone Indicator	2.327961
LMI Indicator	2.072141

Rural Urban Indicator	1.637247
Project City	1.407009
Ethnicity	1.279100
Processing Method	1.151190
Jobs Reported	0.844206
Veteran	0.588386
Project State	0.460476
Approval Amount	0.383730
Business Age	0.230238

From the table we can see that, when the loan was bestowed by the bank, has the most priority for the algorithm to predict the future status of the loan as default or not. Then, the business owner gender and next is the Hub-zone indicator which based on the SBA loan dictionary belong to the region with priority to be supported by the federal government. Yet in this table the percentages and differences are fairly meager amounts, i.e. we cannot rely on these to make a remarkable prediction on the loan status. In the upcoming part, where we see the features importance ranked by XAI, we will more talk about the ranking and effectiveness of the features in the prediction.

K-Nearest Neighbors Algorithm (KNN)

The K-nearest neighbors algorithm, or KNN, is a non-parametric, supervised learning method. It classifies or predicts the grouping of a data point based on its proximity to neighboring points. KNN is a versatile tool widely used in machine learning for various classification and regression tasks.

The abbreviation KNN stands for K-Nearest Neighbour, is a supervised machine learning algorithm. It can be used to solve both classification and regression problem statements.

The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol ‘K’ (analyticsvidhya.com/blog).

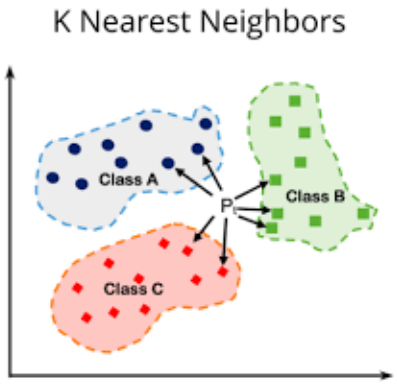


Figure 8. Schematic of KNN algorithm to classify P in the closest group

For the $1~10M borrowers, the KNN algorithem showed 96.44% accuracy. As it is shown in the graph, the accuracy of the model drops at the second knot but rises again in the third one (the trend is almost the same for all three groups of borrowers, from the third knot up to 7 of the neighborhood the accuracy rises).

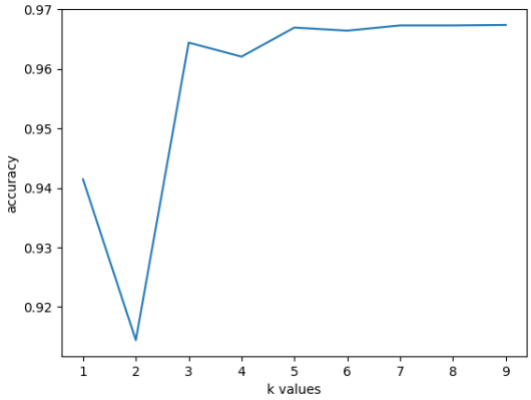


Figure 9. KNN graph, up to 10 knots for $1to10M PPP loan borrowers

In the second group KNN algorithm reached to 95.82% and for the last group of borrowers, $20.8K, it was 80.37% precision.

Logistic Regression (LR)

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The

model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

By assessing the database of $1to10M loan borrowers via LR, the algorithem could 96.73% correctly predict the default situation of the firms.Also by showing the following heatmap, it is observable that the rate of false prediction by the model is so low (816 instances falsely predicted to not defaulted but they have registered as defaulted in the real data, 0=Default & 1=Paid in Full).

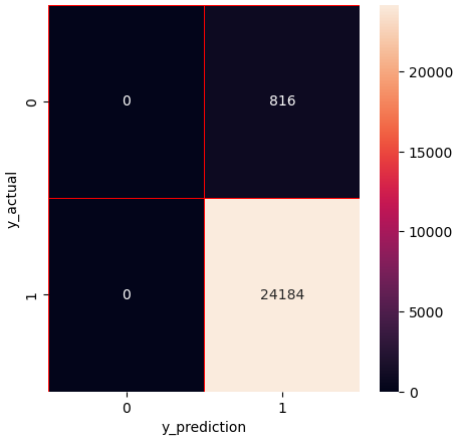


Figure 10. Logistic regression prediction heatmap for the $1~10M borrowers

The algorithm accuracy for the second group revealed 96.29% and 82.34% for the $20.8K borrowers. For the last group, the algorithm is not capable to predict if the business will default accuratly. There are 9971 enterprises which have predicted to default the loan but in the real data they have not paid the loan back.

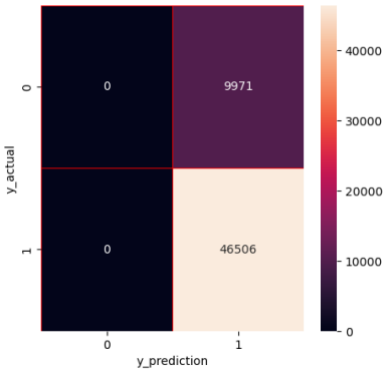


Figure 11. Logistic regression prediction heatmap for the $20.8K borrowers

Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model (scikit-learn.org).

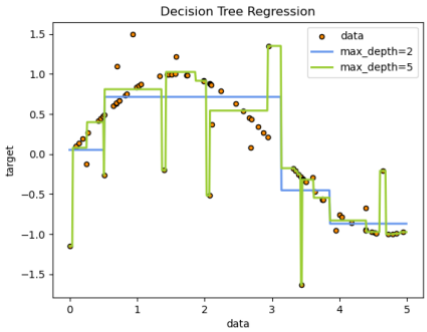


Figure 12. Example of the decision trees in data investigating

For the biggest group of borrowers, DTs had 93.5% accuracy and its feature importance is:

Table 5. The importance of each feature for the $1to10M group by KNN

Feature	Importance
Loan_Amount	0.186534
Borrower_City	0.167240
Jobs_Reported	0.153017
Lender_Name	0.142947
Industry_Detailed	0.112839
Borrower_State	0.070846
Industry	0.059671
Business_Type	0.041015
Gender	0.024435
Race	0.016818
Date_Approved	0.014905
Business_Age	0.009732

DTs algorithm accuracy for the $150K ~ 1M and $20.8K borrowers were 96.79% and 79.26%, recpectively. DTs model for the last group gives the priority to 'Term' amongst all the feature for affecting the prediction of default situation.

Random Forest

Random forest is a flexible, easy-to-use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most-used algorithms, due to its simplicity and diversity (it can be used for both classification and regression tasks). One of the well-known example of using random forest is the prediction for e-mails to distinguish as spam or not (builtin.com).

This algorithem showed 96.73% accuracy for predicting in the $1~10M group, and there was a slight difference in the feature importance compared to LR and DTs.

Table 6. Feature importance by RF algorithm for $1~10M borrowers

Feature	Percentage
Loan_Amount	16.83
Jobs_Reported	16.76
Borrower_City	14.96
Lender_Name	13.54
Industry_Detailed	11.80
Borrower_State	8.02
Industry	6.52
Business_Type	3.94
Gender	2.67
Race	1.99
Date_Approved	1.62
Business_Age	1.29

Support Vector Machine (SVM)

Support Vector Machine (SVM) is another machine learning algorithem which is vastly used by expert. The algorithem's aim is to find a hyperplane that distinctly classify data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence (towardsdatascience.com).

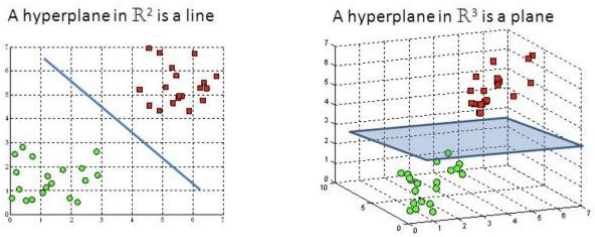


Figure 13. SVM finds a hyperplane to distinct the data points

SVM also could show 96.73% accuracy to predict the default and not-default firms in the $1~10M group. For the second group the rate was 96.29% and for the $20.8K it was 82.34%.

Naïve Bayes

Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes. A popular example in statistics and machine learning literature to demonstrate this concept is medical testing. For instance, imagine there is an individual, named Jane, who takes a test to determine if she has diabetes. Let's say that the overall probability having diabetes is 5%; this would be our prior probability. However, if she obtains a positive result from her test, the prior probability is updated to account for this additional information, and it then becomes our posterior probability. From then forward, the probability is renewd to create the final model (ibm.com).

For the $1~10M, the accuracy of NB algorithem is observed as 95.88% which not very promising. The following chart shows the result for different algorithem accuracy for the biggest group of PPP-loan borrowers:

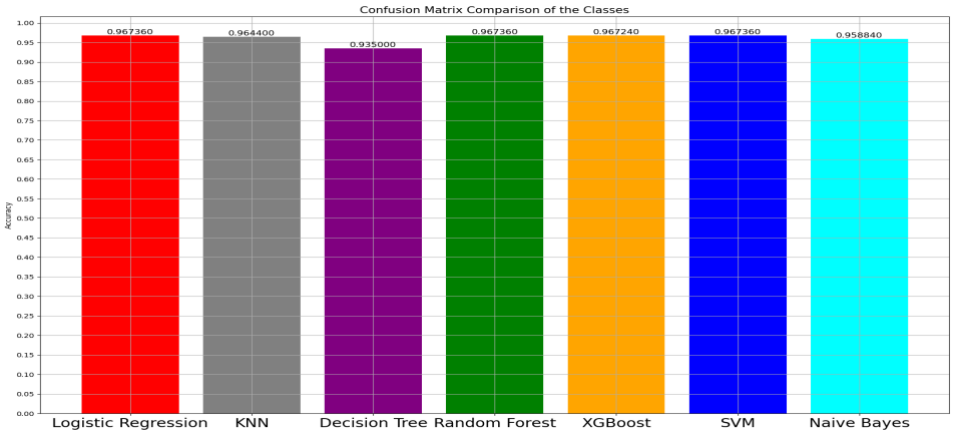


Figure 14. Different algorithems accuracy to predict deafeult status of $1~10M borrowers

The NB algorithm for the second group got 96.02% accuracy and the comparison of all the 7 used algorithm is shown in the following chart.

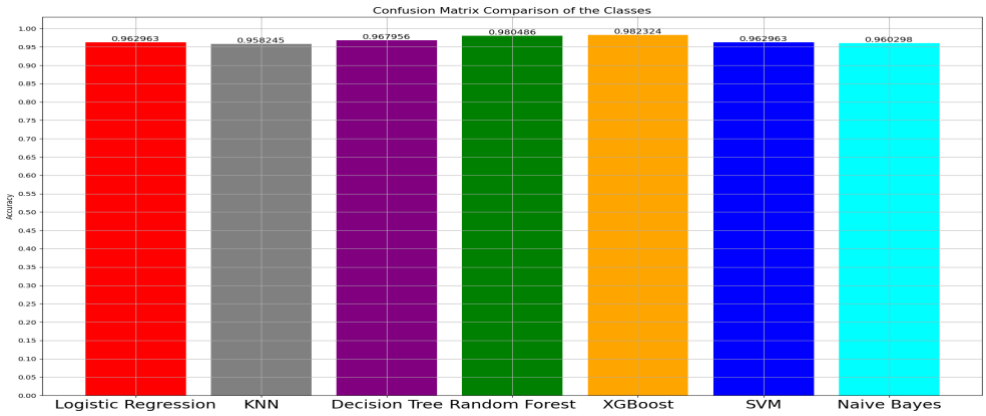


Figure 15. Different algorithems accuracy to predict deafeult pattern of $150K to $1M borrowers

The last group which has been the most difficult group of the borrowers, could only get to 85% accuracy via using NB algorithm and by this final algorithm we can see the comparison bar chart of different algorithms and their correspond precision to predict then default condition of the businesses in the following figure. (the gap to predict this group behavior compared to the two others is clearly visible here).

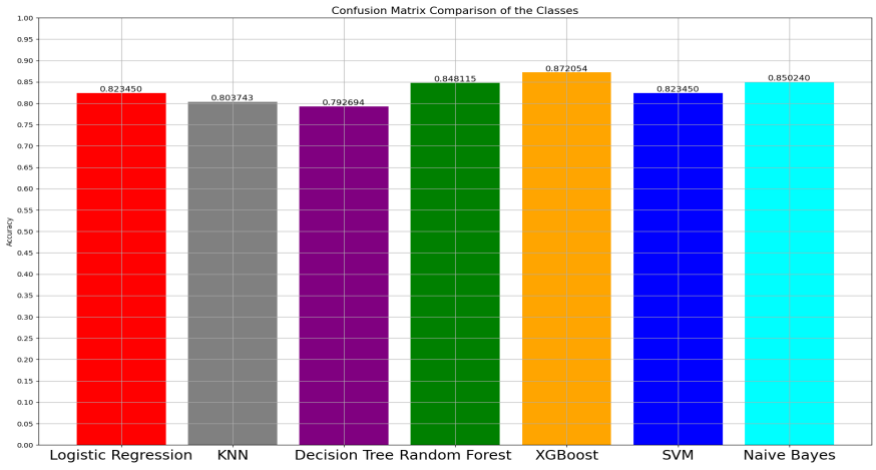


Figure 16. Bar chart for accuracy of default prediction from different algorithms for $20.8K borrowers

LIME Analytics

Local Interpretable Model-agnostic Explanations or shortly say LIME, can explain the complicated decision by different algorithm in an understandable way. LIME will provide many locally examples to address the importance of features and their weights in the prediction decision made by complex algorithm such as Random Forest, Neural Networks and Logistic Regression. As XGB was the most accurate algorithm to predict the default condition of the business in all three groups of the borrowers, the model provided by this algorithm used to feed the LIME package as input. According to what is mentioned before LIME will give local solution and is able to provide insight how the importance of each factor to the final decision will vary in different parts of the dataset. LIME by checking the capability of the algorithem in diverse points, will tell how much the model can predict the conditions. The sum of all conditions eqaul to 1 which is for 100% as the whole. For instance if the model in the locally chosen area comes by 0.8 and 0.2, that means it can be able to predict 80% of the firms at this point who would default, then the ability of the model for predicting the firms which not default would be 20%. This story is also true in reverse way. In each local solution, LIME also sorts the importance of each feature in the outcome of the model. In the next part after presenting the results, how LIME works would be understood better.

Starting with the biggest borrower group in our selected SMEs database, the first scenario claims that XGB algorithm probability to predict loan status is 88% for non-default firm and 12% for default firm. It translates as in the locally chosen chunk of the dataset, XGB can forecast 88% correctly that a firm will not default and only 12% correctly that the firm will default. For such a prediction we most consider that the number of the jobs reported by the firm should be less than 93 (the firm with maximun 93 employees) is the most important feature to help to predict that a firm will default. However the importance of this feature is only 5%. On the other word, there is 5% chance that if the borrower has less than 96 employees, it will default. Loan amount feature is less than $1.29M while the contribute of this feature to predict the firm will not default is 3%. Furthermore it is stated that if the borrower has received the loan in the second quarter of the 2020, there is 2% probability that it defaults.

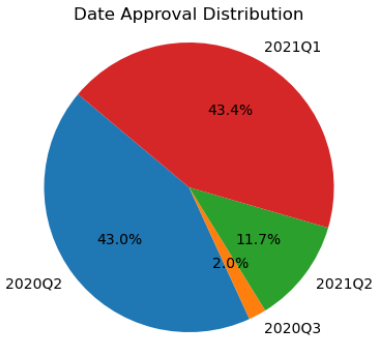


Figure 17. Distribution of the four priods of lending the $1~10M PPP loans for defaulted firms

Here gender and the fund who lent the loan (usually a bank) do not play any role to predict if the loan will default or not. On the contrary of the common sense for the old businesses will stand more, how long the business has established and worked, does not affect the prediction as well. It means the new commers and the old businesses can not tell much if they borrow loan, one group can be pointed with more probability to default. Industry sector in which the firm works in it has a only 1% effect on the prdiction if the firm will default (note that only default and not for pay in full). For example we can say that there is 2% probability that if the firm operates in the food services and accomodation, or constrauction, the firm will default on the loan. Maybe at the first glance it seems not very struggling to predict that during Covid pandemic that hotels, restaurants and big construction companies would default, however, the model findings show that these industries can take into account for only 2% of the default prediction (i.e. there is still 98% parobability for the firms involving in this industry which are not gonna default). Also it is astonishing to say that those firms who are dealing in professional, scientific and technical services are far ahead of the firms operating in art, entertainment and recreation services, to default the loan during Covid pandemic.

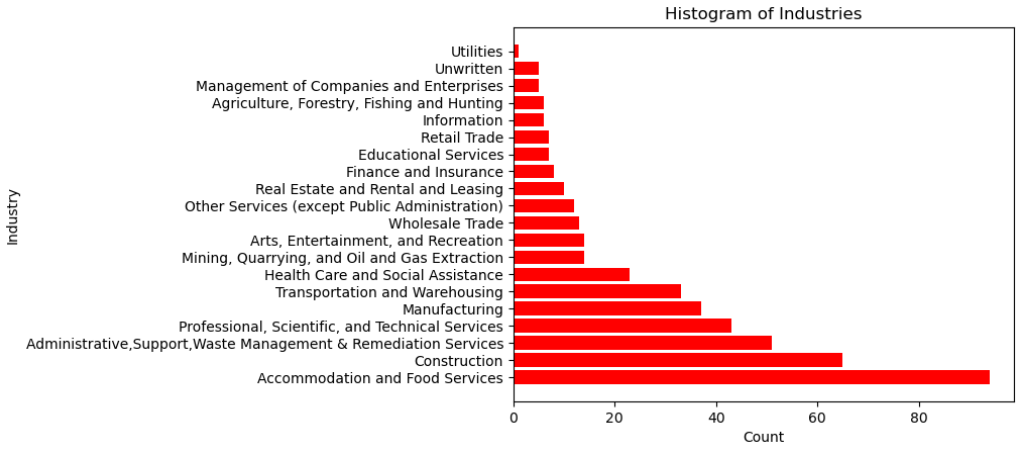


Figure 18. The most defaulted number of firms by industry for $1~10M borrowers

If we move forward to the other point of data, then race and gender would get a bit more colourful in prediction. Male owned businesses defaulted the loans a bit less than female owned businesses.

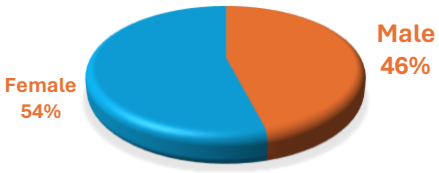


Figure 19. Female/Male owned business defaulted loan $1~10M distribution

And for the race feature when the defaulted businesses investigated there distribution is such as below.

Table 7. Defaulted rate of $1to10M borrowers by the owner race

White	Asian	American Indian	Native Hawaiian	Black African
2,56%	5,11%	4,65%	9,38%	8,73%

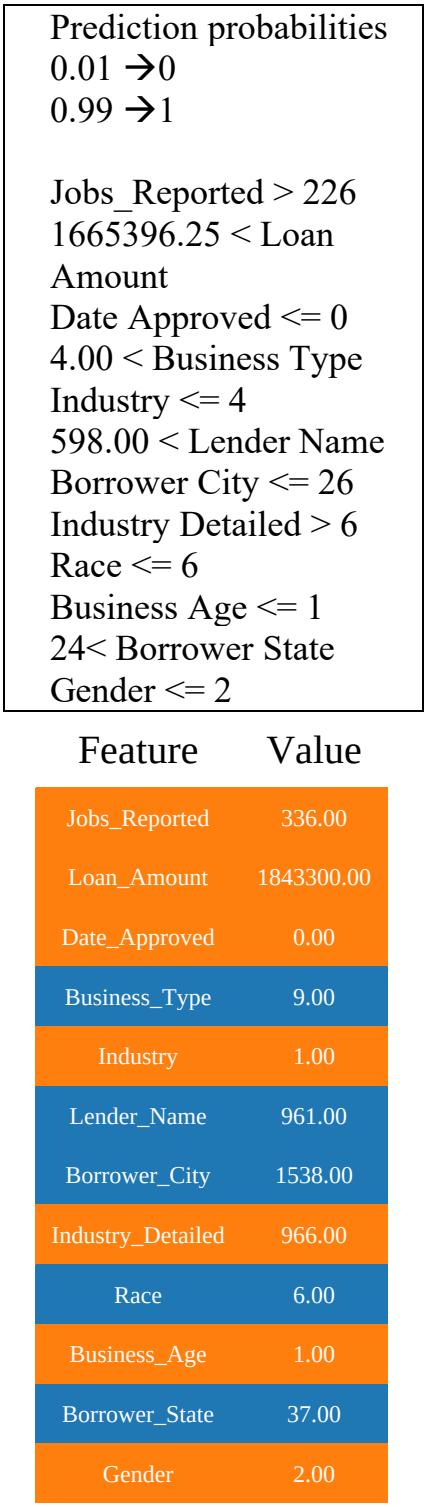


Figure 20. An example of LIME package output for feature importance ($1~10M)

For the second group of borrowers, LIME package by using XGB, when the model can predict the default firms by 2% and not default firms by 98%, it is plausible that 17% of the early borrowers will lead to not default. After this feature the by looking at the number of employees in the firm, it can be said that for those firms which have 17 or less personnel,

there is 5% probability to default. Rural or urban borrowers, gender, race and ethnicity have nothing to say about default or not default probability prediction here. With the borrowers of $200K, we can consider 5% probability to be sure that they do not default (note that it does not mean that 95% will probably deafult).Moving forward to another point and concentrating on the prediction of default, 98% default and 2% not default ability to forecast, there is 18% probability of default that late comers to default the loan in both 2020 and 2021 (who showed up in the bank in the second round of lending, announced for those eligible firms who missed the first round). Then we can look at then a certain number of cities with 8%, short term borrowers borrowers of less than 2 years commited to payback the loan with 6%, and firms with less than 17 employees with 2%, are plausible to default, recpectively. Changing the position to other point of data for getting the prediction accuracy in this group, the result would be not far from each other. That means we can expect the same feature, more or less can give us the similar importance percentage to predict the probabilty of default or fully payback the loan by the firms. Early or late commers, the amount of the loan, the lenders city, term of the loan and number of staff are the most important, while gender, ethnicity, race, urban or rural area, low document indicator and prior zones to receive the aid package have the least importance on prediction.

For the last group of borrowers, $20.8K in this research, no matter at which point of data is looked, the predition could not get close to 100% in any direction (Unlike the other groups while in some points of data the model could reach to 1 or 0.99 for not default probability). The first local optimize solution by LIME using XGB as input model, the scenario with predicting probability with 80% for the firms which will not default and 20% who will default, the most important feature can be looking at where the borrowers come from and where they got their loans from. Banks in states of Pensilvania, New Jersey, and Arizona by the order had the most defaulted rates.

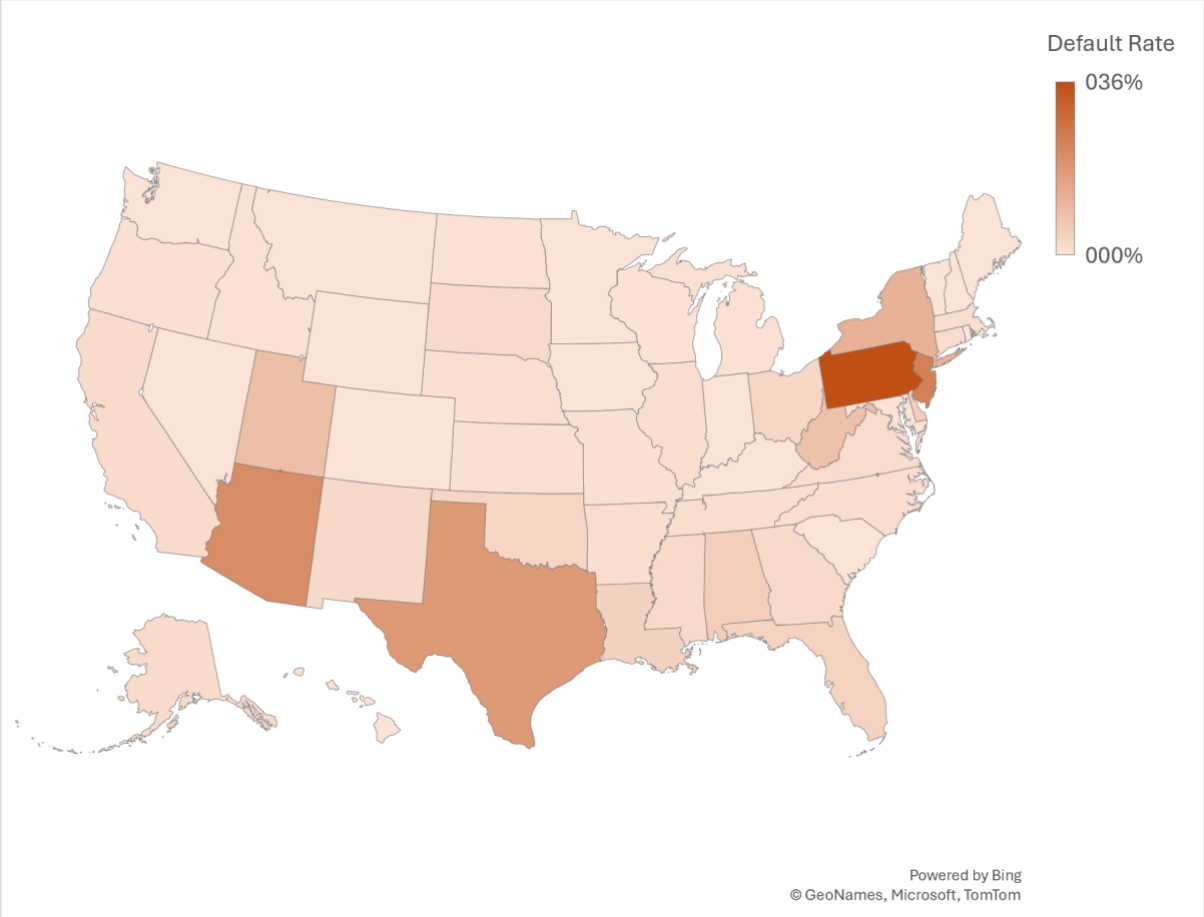


Figure 21. The default rate for the $20.8K lenders (banks) by state

It should be mentioned that the states default rate from the side of borrowers can be differ from the lender side. If we scrutinize data more and go further for small districts, take cities into account, specific cities at lender side had 17% of loans registered as default. Milwaukee and Detroit had the most registered rate of default by 26.63% and 25.54% at the borrower side. The late borrowers are again more plausible to default by the rate of 4% prediction.

Moving to another point of data, by probability to predict default firms by 11% and pay in full by 89%, aside the above-mentioned feature, if we know that the borrowers are not veteran, then there is 4% probability to put them into default category at the time of giving them the loan. Solo-partnership companies in this group of borrowers has the most rate of default by 19.5%, this feature, the type of business, can help by 3% probability to say if the business will default.

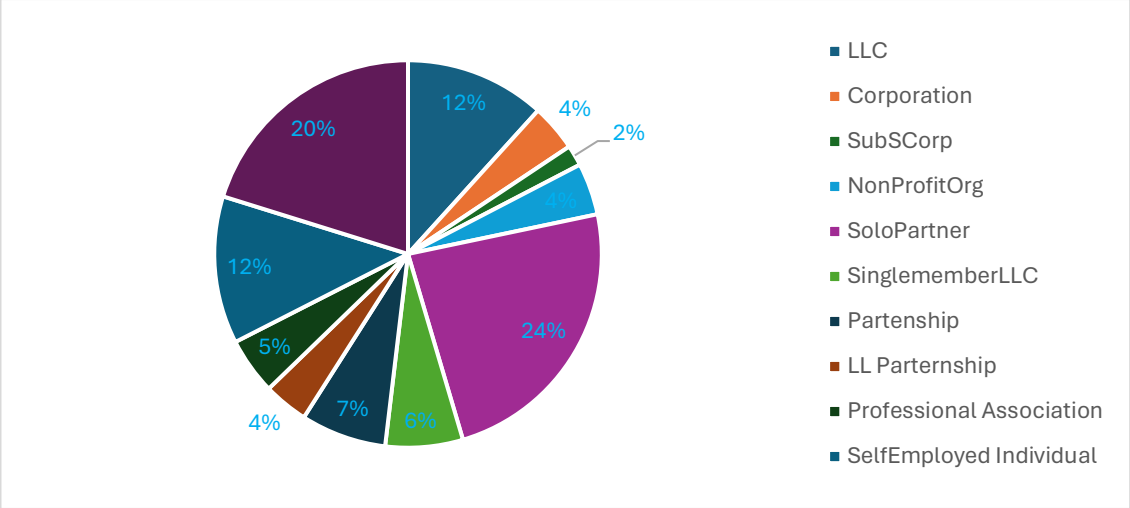


Figure 22. The default rate by business type for $20.8K borrowers

Also, it is intriguing that the start-ups had no default in this group. Most businesses who defaulted were existing more than two years (The complete results are attached in the appendix).

Shap analytics by XGBoost as the reference model

Shap, as mentioned before stands for Shapley Values, provides comprehensive understanding of how each feature affects the machine learning model output. In this research this Python package was intensively used in many ways. Besides that, a handful of models employed as the input of the package, and they result examined with the actual data. To divide dataset for training the machine and applying the model for prediction, 75% of data was allocated as training set and the rest 25% as the test set. Observing all the models results, it revealed that the most effective results come from XGB and Logistic Regression (LR). The performance of the XGB, surprisingly showed highly performance. The game theory by which the contribution of each player in the final result can be explained, brings an impressive option on the table for financial decision makers. In the following, the result of applying Shap on the models XGB and LR are presented. Let us begin with the biggest group of borrowers in our dataset.

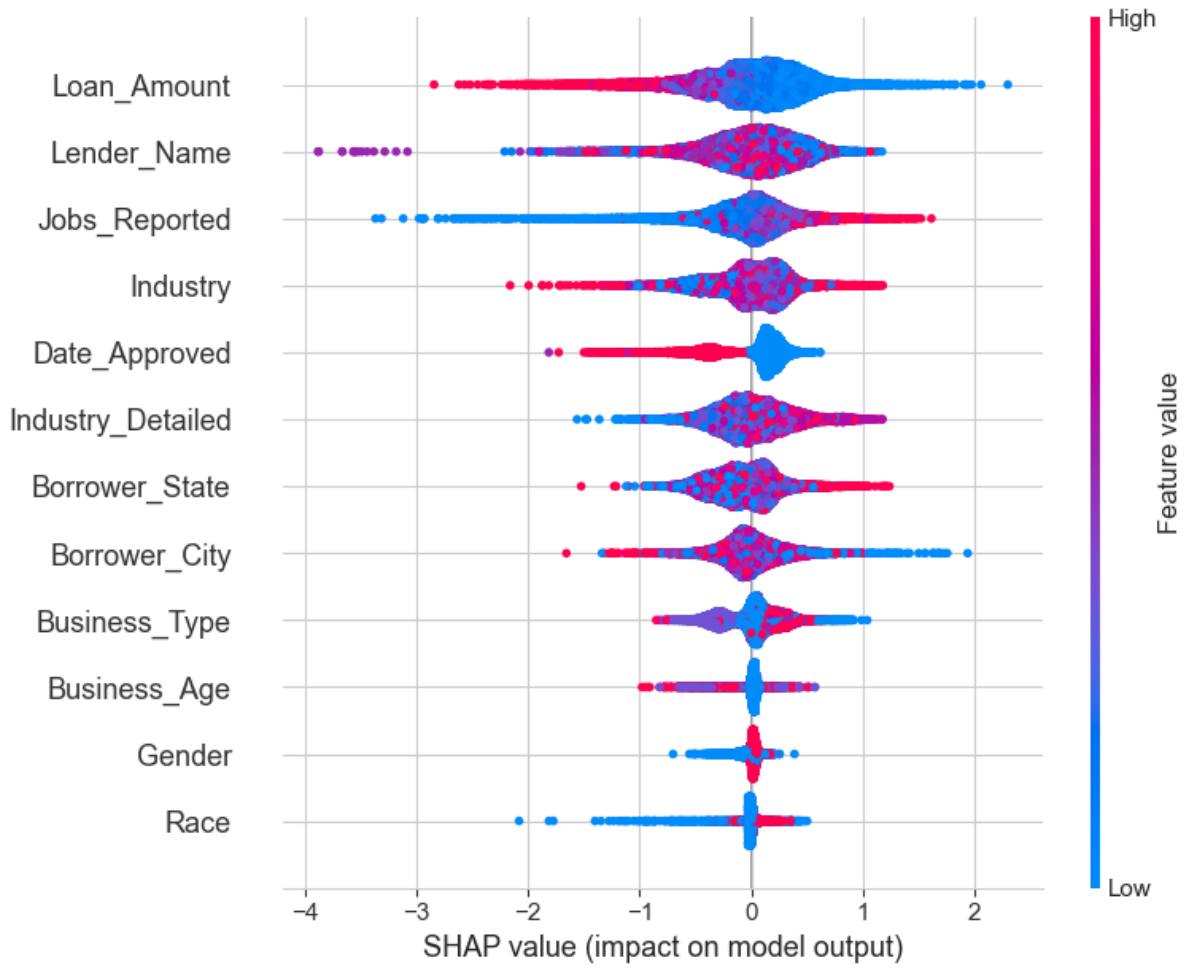


Figure 23. The importance and impact of each feature in the XGB model prediction ($1~10M)

The graph as said before, is the output of Shap which was fed by XGB model. This means that the graph shows the share of each feature importance to predict if a firm defaults or does not default the loan. If we get back to the XGB result section, we see some changes in the order of this graph and what we had in the feature importance table. There are many reasons for the mismatch. First of all, a model like XGBoost, splits the data many times to find the optimal solution. The number of splits sometimes can be very high. During this process the role of each individual feature in the final prediction can be lost. Secondly, the model looks at the frequency of the feature, which is used for the prediction, which may not be the highlighted feature in all predictions. Also, model has a holistic view for the entire dataset, while in many predictions the introduced feature as an important may have no effect at all. On the hand, Shapley Values, provides both locally and globally solutions (feature importance in this research). Hence, the results of that could be considered for explaining what happens in any part of data or for its entire. Moreover, Shap package,

assess the importance of each feature in all predictions of the model, therefore the provided output tells how much each feature really matters in the final prediction. Finally, Shap offers understandable and explainable results that could help humans to get a fairly well insight of why model came to such a decision.

Getting back to the graph, it is observable that there are features which are ascending based on their importance; So, they sorted from low to high. On the x-axis there is two side of minus and plus, which the features are stretched in this spectrum with different concentration. The positive side means that the feature plays a positive role in the prediction. Higher the number and higher the bulkiness, more important the feature is for predicting the situation (here default or payback the loan by borrower). On the other hand, the negative side demonstrates the negative impact of the feature in the predictions. More negative the number is, less impact and even rarely distracting effect on the final prediction. Also features are shown by two colors of red and blue and their combination. The red part means that there features lead to predict the higher side of the dataset; In this paper as we are looking at the default or not default condition, the red color represents the paid in full businesses. By another mean, if the feature has more red part, it means that it has more impact on the model for predicting that the business will not default. On the contrary the blue color is for the opposite condition (here default). So, if the feature is colored in blue on the spectrum, then it has an impact on the default prediction.

By looking at the graph for borrowers in $1~10M, we see that loan amount is the most important feature in the prediction. This feature has pretty clear distinction between colors. A concentration in the 0 point is visible, and then move to sides it gets narrower and narrower. The blue color stretches itself for this feature on the positive side, this means that loan amount has a share (noticeable) in predicting which firm will default. On the other side, the red color goes toward negative side, and it means that loan amount has negative effect on predicting that the firm will payback the loan. The next feature is the lender name or literally the bank which has lent the loan. The colors red and blue are so mixed here that we could not be certain to say that this feature affects positively or negatively to predict default or not default firms. Looking at the actual data shows that

there are certain lenders who had the most and who had the least defaults. However, as the concentration in this feature is not high (as the nature of it, there many banks in the USA and have many branches in different locations), we cannot point certainly some banks as the future default lenders.

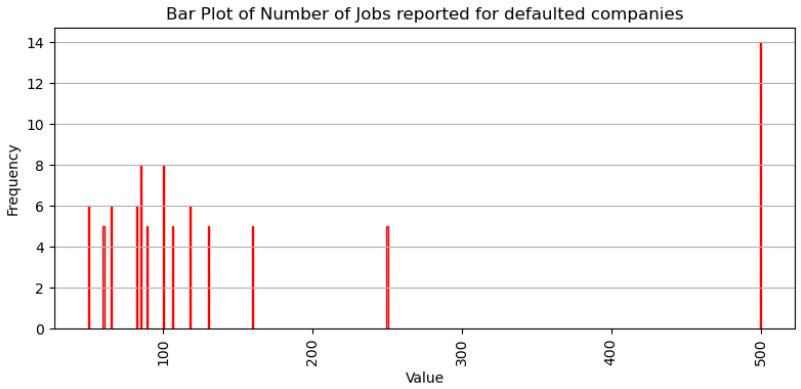


Figure 24. Businesses defaulted frequency by certain number of employees

The next important feature is the number of the employees. It has quite clear distribution with red in the positive side and blue in the negative side. Then we can guess that some businesses with a certain number of staff will not default. Here we observe that most defaulted companies have around 50 to 150 personnel (though the firms reported with 500 employees have the most default rate, the total number of defaulted in the mention range is clearly higher). One reason for explaining this could be when small firms are in the transition phase to get into medium category, they are more vulnerable while they experience more turbulences during transition phase which leads them to default. After these ones, the firms with 500 employees have frequently defaulted while due to pandemic they had to be more cautious about social distancing and had many days off, yet the payment of such a number of staff is huge overdue that business cannot handle it without operating.

Next important feature for the XGB model to forecast loan status is, in which industry the businesses are operating. According to the graph for a few industries we can consider more probability for them to not default the loan, but to say which industries will be more eligible to default, there is no strong evidence. The below chart shows which industries had the most numbers of default.

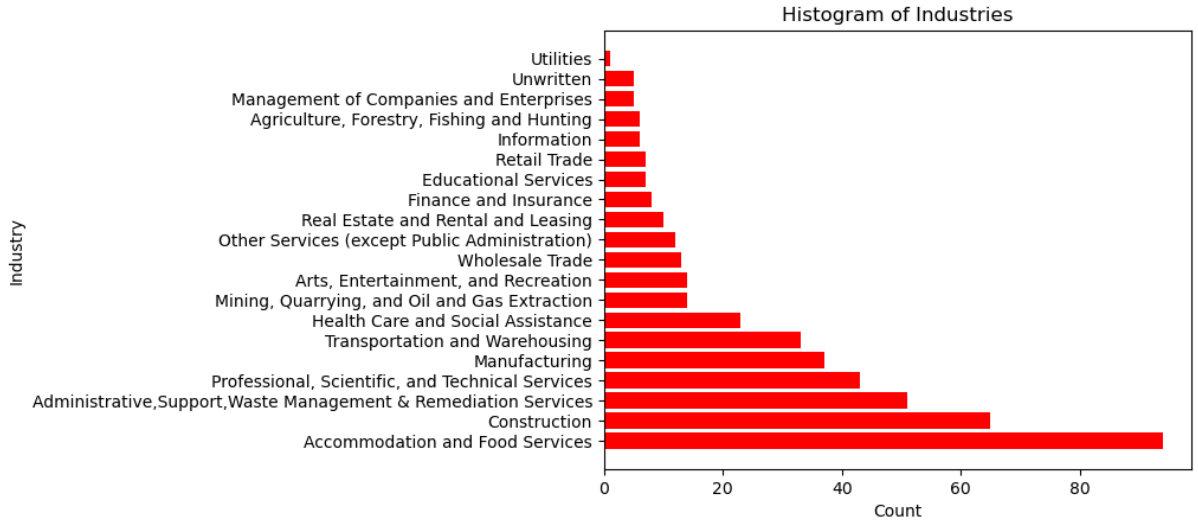


Figure 25. The industries by number of defaulted firms for $1~10M borrowers

Although the graph is fairly matched with our perception with Covid pandemic period, that restaurants and big construction companies came to halt for operation, data analysis shows that the highest rate amongst all borrowers in this group belongs to the firms involving the management of companies and enterprises by 1.3% (0.2% more than accommodation and food industry). It is vividly visible that this feature should locate in the middle as not having much to contribute to default or not default prediction. The date or the period in which the loan dispatched to the company again has a quite clear effect, late comers are more entitled to default. The blue chunk at the beginning of the graph tells that we can consider a positive correlation between certain periods of spending the loan with the probability of default. From which place the borrowers come, has small impact on prediction of default condition. Apparently specific states are good and specific cities are not good at not defaulting. The graph shows a narrow red stretched line in borrowers state and a narrowed blue line for their cities along the positive side. Business age, gender and race are the least important features in this predictive model, though the race of the borrowers shows a bit of default makers probability. Native Hawaiian or other Pacific Islander by 9.37% had the most rate of the default that could be regarded as the nature of the place and the businesses which based on tourism industry and hurt during pandemic, even though they were the smallest group of the borrowers.

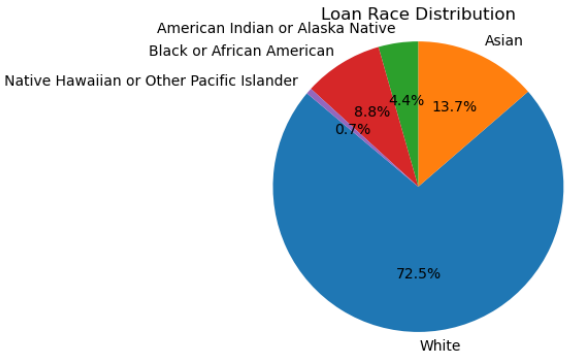


Figure 26. The loan distribution for $1to10M borrowers by their race

Moreover, although the female owned companies defaulted 12 times more than male owned companies, this feature could mislead the prediction if we take it seriously into account. And finally, the last feature we investigate in the Shap graph is the business type. It is located in the top of the last quarter of features importance. Clearly when we talk about medium size businesses it envisages Corporation or Limited Liability Company in our mind which have also the most numbers of defaulted firms. Despite that, the highest rate is for Solo Partnership which 12 out of 714 companies registered default in this category which accounts for 1.67% of the borrowers with the same business type.

Then we have the second group of borrowers who received between $150K to $1M loan. The Shapley values graph for this group is presented in the following.

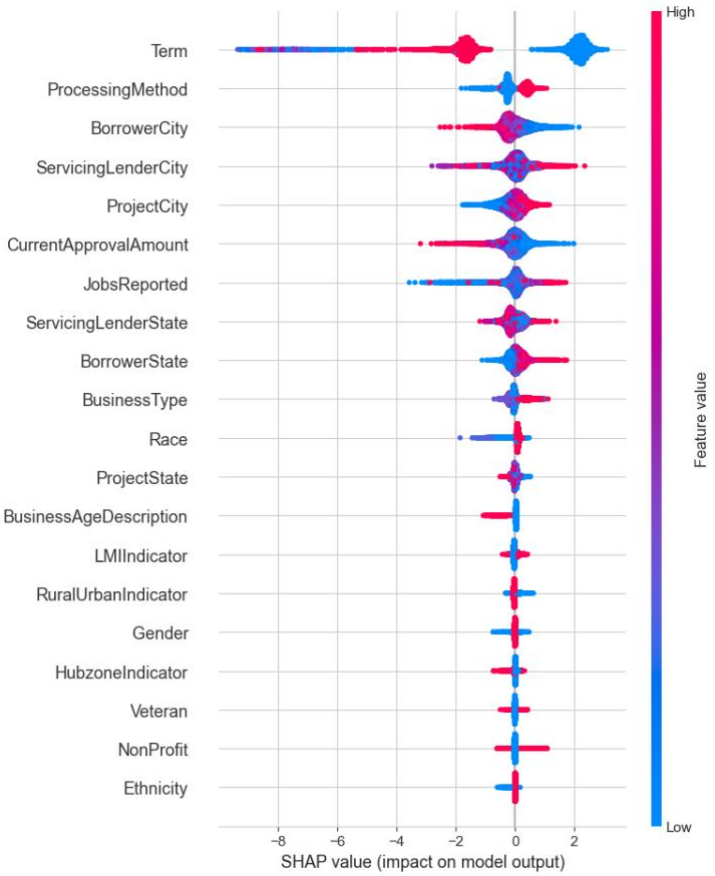


Figure 27. Shap feature importance graph by using XGB model for $150K~1M borrowers

Now the scenario is different for feature importance when the share of each feature in the outcome is assessed by the Game Theory. For $150K+ loan borrowers the most reliable variable to make a forecast on the loan status is, how long a borrower agreed to payback the loan. With certain terms (in this research based on month/s), we cannot confidently but with fair probability predict that the loan will default (as in the data review section could be seen many 33 to 38 months borrowers defaulted). Then who comes first and who comes late gives a chance to a group that the default low rate is less than the other. Here the red line on the positive side of the axis means that the fully paid group could be guessed by when they come to borrow the loan. After that there are certain cities that the borrowers will probably default on the loan, on the other hand there are certain cities that their banks have a probability to not register default. The project city is the same as the borrower or lender city and therefore was not investigated deeper. To some extent we can say a loan will default by looking at the amount and is not going to default by looking at number of employees in this group. States are after these features got settled. It sounds the city is a better way for guess on the loan status not the whole state. And both borrowers and lenders

are good to payback the loan in some states. There is also a small probability to default given to loan if we know the race of the business owner. Gender of the owner, ethnicity, low documents program and prior zones have not much to say here. Even an extended narrow red line on the positive direction for the non-profit feature (we know that the default rate of non-profit businesses is very low), the model has not put it on top. A reason is that the numbers of these businesses are not enough to deduct any conclusion.

For the last group of the borrowers, $20.8K, the Shap values graph for features share in the model prediction is shown in below.

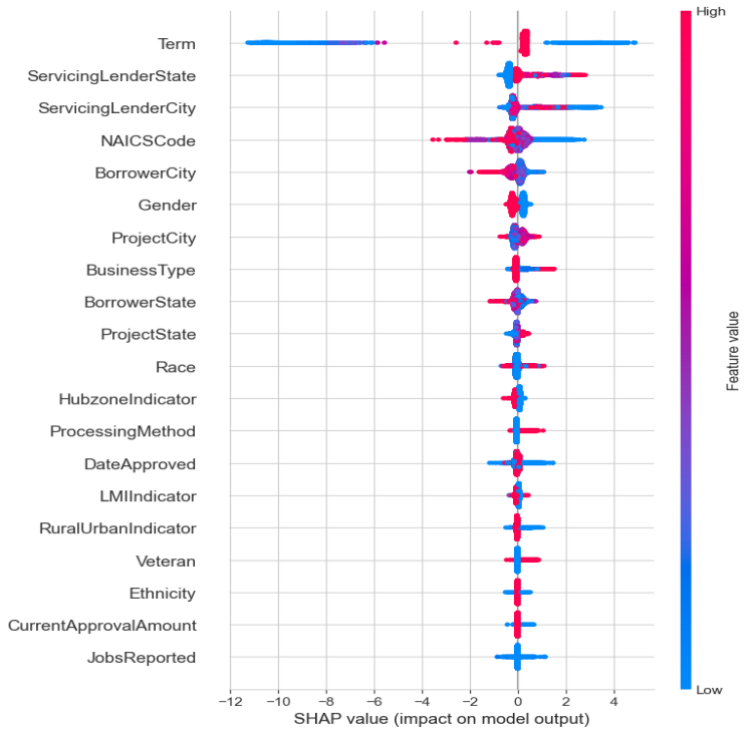


Figure 28. The Shap values graph for default prediction by XGB ($20,8K borrowers)

Like the former group, again term matters as the most. Model suggests that the best bet could be on how long the borrowers are going to payback the loan. Certain months are susceptible to default. Lenders location by state and city are the next important features (as we will see in the actual data investigation part, banks of some cities have a huge rate of default). Which business the borrowers involves in is the next. For certain activities the default prediction could be an educated guess. Borrower city, gender, race and the rest features are presented afterwards. The least important feature is the number of jobs

reported. Although the default probability by certain number of employees goes up, it goes down on at the other side. That is noticeable, the model could only employ this feature to predict the default ones and apparently the pattern of not defaulted ones had no contribution to the prediction.

Shap analytics by Logistic Regression as the reference model

Let us start with the features importance graph provided by Shap when it was fed by LR model as the input.

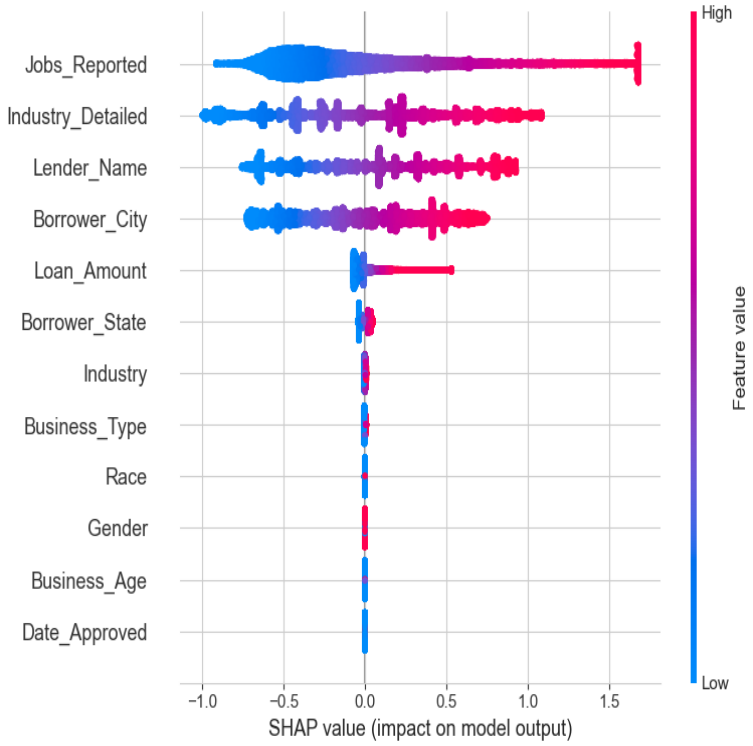


Figure 29. Shapley Values for the contribution of features to predict by LR model

As it is observable, there are some changes in the ranking of the feature importance between the XGB model graph and the LR model graph. LR model asserts that, looking at the number of employees of the firm has the most priority if we want to know about the probability of default or not. The model has extended a red line in the positive side the graph and at the end there is concentration. It means the companies with a certain amount of the staff have fair chance to not default. The second important feature is that the

company industry of the operating in detail. According to the Shapley Values graph what subindustry takes the model into account is not very informative to predict companies as default makers. On the hand there are particular subindustry sections that can be seen as safe to lend them the loan. The following histogram shows which subindustries recorded the most default number of firms.

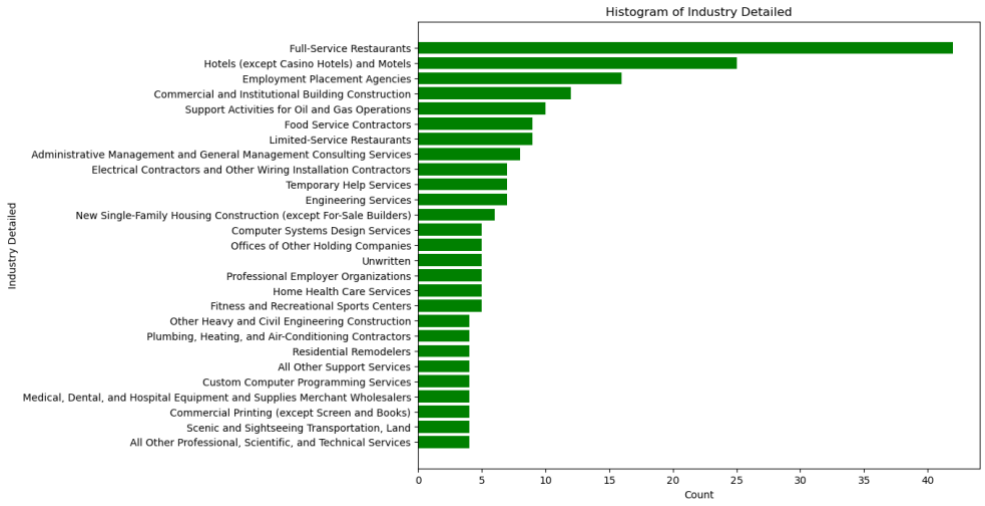


Figure 30. Subindustry defaulted numbers for the $1~10M borrowers

In this model, industry, the business type (Corp., LLC, etc.), the owner race, the owner gender, and how long the business has been in the existence, literally have no effect on the prediction.

Now we look for a single instance by using Shap package to create a Force graph. Plotting the probability default prediction by using ‘force’ makes a graph to envisage feature importance as an instance. The force graph shows that for the model to predict the probability of default, number of employees is reliable feature. As it is shown in the following, we can probe each single instance in the model’s prediction:

Figure 31. XGBoost Force graph for the borrower of $1~10M PPP loans

Interpreting the graph, we see that if the borrower has 46 employees, there is relative high probability (compare to looking at all other features) that it defaults the loan. Next feature is the industry, i.e. if the borrowers deal with transportation and warehousing (the code 17 is the encoded this industry digits by computer), again we consider more probability that it defaults. The bank as the lender which Comerica Bank (code=2414), the probability of default is an option on the table. One way to explain this could be not checking the borrowers enough by this bank staff, or the bank customers and thereby borrowers are certain type which are not very committed to payback. Or simply we can say that Comerica operates in an area with vulnerable business. Nevertheless, for any reason, Comerica bank has registered a high default rate. On the side of the spectrum, we see that the borrowers of the $1.285M loan, are kind of borrowers who likely not default (probability is not very high. The other features are playing small roles in the default condition prediction by XGB algorithm.

Figure 32. Logistic Regression graph for single instance in default prediction ($1~10M borrowers)

This time we get the graph when our input model is Logistic Regression (LR). Here, the firms with 46 personnel are likely going to default (as the same as XGB prediction model). Then LR suggests that the businesses involve in home healthcare services (which code 223 stands for) are the ones with default probability. We have the same loan amount such as XGB model equals to $1.285M, but this time LR gives this amount borrowers a chance of default. Comerica bank is considered as a not default probability to record default loan; This is opposite as XGB suggestion, though, the weight of probability to register default loans by XGB was low but LR give more weight to this for registering not default. We can conclude that this bank has a noticeable amount of default registered, yet the fully paid back loan registered are considerably more that.

In the second group of the borrowers, Logistic Regression model has a more trimmed pattern.

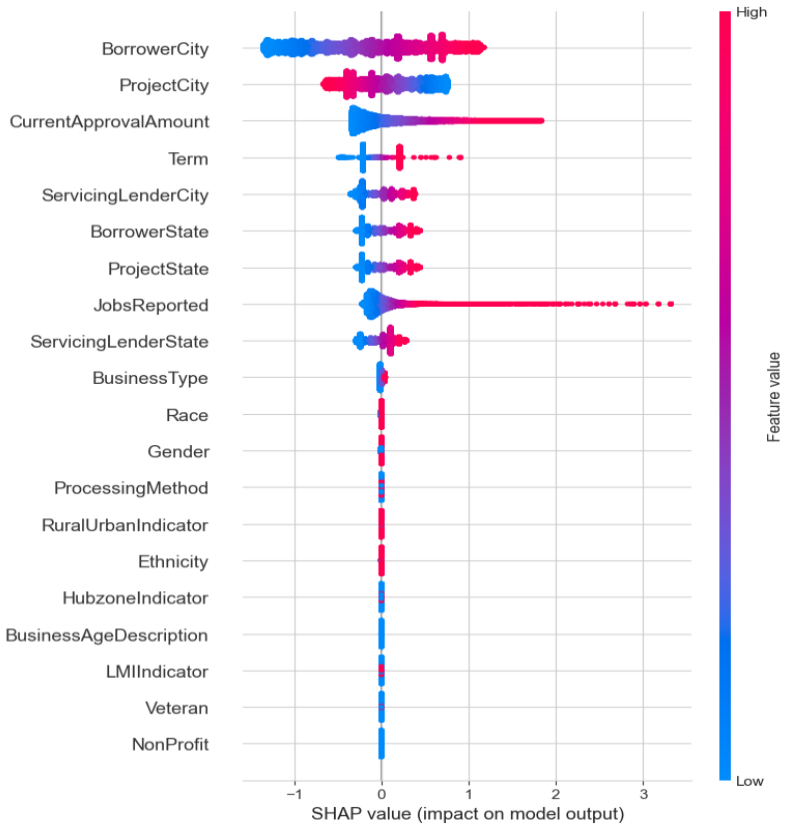


Figure 33. LR graph by Shap for feature importance to prediction (+150K borrowers)

This model uses the city of borrowers (mainly the project city) as the most important feature for default and not default prediction. After that a certain amount of loan have good chance to say they will not be defaulted by the borrowers. It applies for the time span of the installments too. Then we have the lender city followed by the borrower state that with a low probability for fully be paid. Number of employees is another feature that can calm the lender down to some extent. The rest literally have no contribution for prediction by LR model.

Figure 34. Force plot for features importance of $150K~1M borrowers

The force plot shows that, 5 years loans are eligible to be considered as default. Borrower's city also could arise suspicion about receiving it back (Las Vegas as the highest rate of default). The early comers to borrow (processing method=1) likely will payback the loan. The other features have small shares in the model prediction.

Finally for the borrowers of $20.8K, Shap package showed the importance features for Logistic Regression as below.

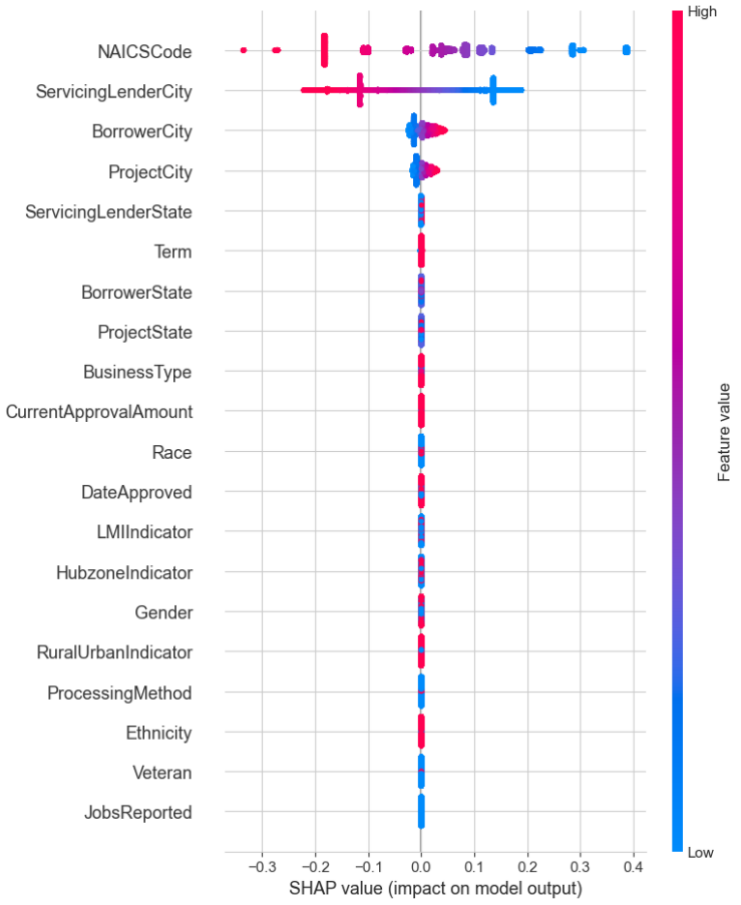


Figure 35. Feature importance graph by Shap for LR model ($20.8K borrowers)

Apparently, what these small businesses are doing, is the matter of the default. For some businesses (such as barber shops or beauty salons) the probability to default the loan is not negligible (not difficult to guess even during the pandemic time when the loans were lent to them). There are also some candidate cities at the loan lenders side, who are not expect that they can put the money in a good hand (these banks cities are mentioned in the actual data investigation section). Borrower city (almost the same as project city) is the next, yet

with very low magnitude. Rest of the features do not play any role in the default probability by the model.

For this group the instance graph via Shap using XGB is as the following.

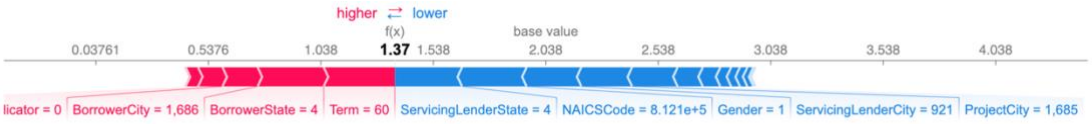


Figure 36. Shap instance plot for $20.8K borrowers in the default probability prediction

The graph complies with the above explanations for the feature importance. Besides what has been said above, the model suggests that we can give a probability for the 5 years borrowers and borrowers of South Dakota (encoded as 4 by computer) to pay fully back the loan.

Actual data investigation

Datasets for all three group of borrowers, were assessed separately. States could be mapped to find out pattern for the defaulted borrowers. Scanning these maps showed the expected result as populated and crowded area got recorded more loan defaults compare to the other areas. The state such as New York, New Jersey and California are highlighted in the maps here. Also, the touristic resorts such Hawaii and Florida had high rates of defaulted borrowers as well. However, this not all the story, in some instances we observe different trends. Firstly, there is yawning gap between groups of the borrowers in up to $150K and from $1M up to $10M. In the former group the average amount of the loan lent is $21000 but for the latter one it equals to $1.2M (these firms received almost 60times bigger aid amount). In addition, the former borrowers are mainly sole businesses or individual contractors while in the latter group most businesses are Corporations and LLCs. Therefore, they behave differently on the loan status. There were also some similarities, such as interesting fact that no startup was registered as default in all three

groups of borrowers. Some of the facts about these groups are presented in the following parts.

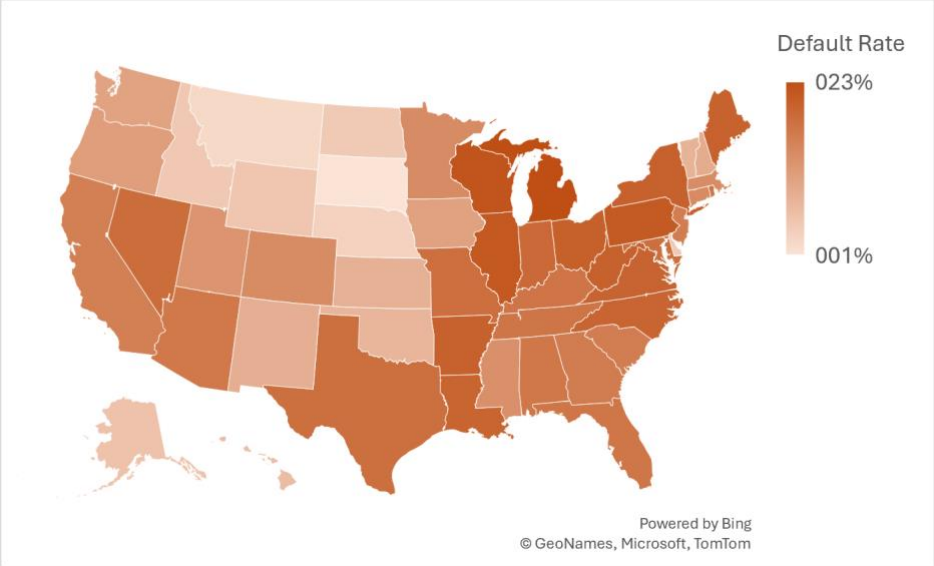


Figure 37. $20.8K borrowers default rate by state

The map reveals that Michigan has the most rate of default amongst all states for this group of the borrowers. Generally, say, the east part of US is deeper shaded compared to the other parts. In the west, Nevada is highlighted by having high default rate. An interesting observable fact here is that the rich state of Maryland has pretty high rate of default 17.87%, while the poor state of Mississippi has one the lowest default rate 12.98%. This tendency less or more exists in the visualized data map. It seems when the loan is guaranteed by federal government, the poor are more committed to pay it back. The other surprising fact about this data map is, it is different from the lender point of view.

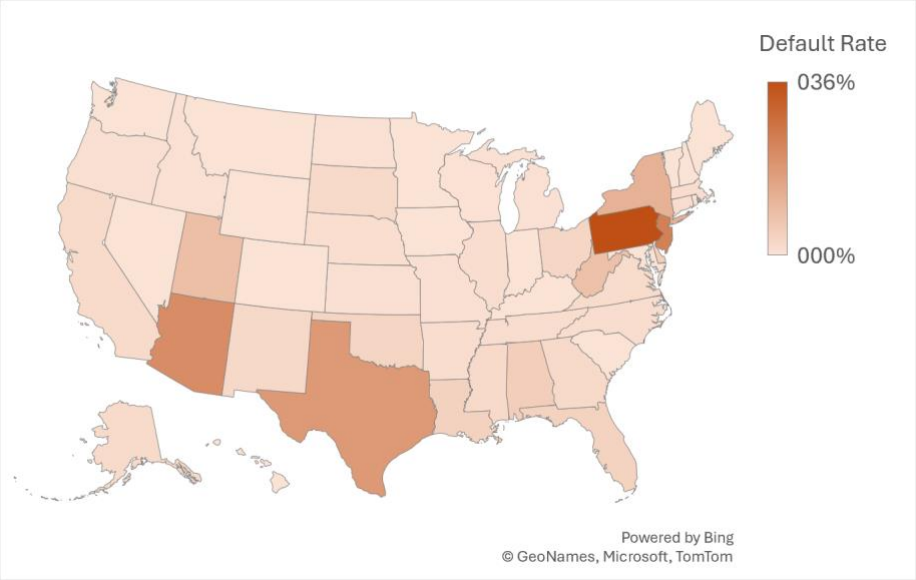


Figure 38. The lenders of $20.8k loan default rate by states

Apparently, many borrowers who received the loan from one state but are registered in another state by the official address. Here the default rate of Pennsylvania is high (36.27%), which has the default rate of 20.94% at the side of borrowers. Among the cities, Malvern in Pennsylvania has the astonishing default rate of 68% percent at the lender side, but when it comes to the borrower side, Detroit in Michigan and Milwaukee in Wisconsin by 25.5% and 26.6% are on the top. In this group of the borrowers, whenever businesses collectively stood up together to run the business such LLC or Partnership Co., the default rate was low and whenever the business got on only one person's shoulder such as self-employed or independent contractor, the default showed high rate. There was not a significant difference between the male and female owned business to default in this group. Non-profit business defaulted 9 times proportionately more than for profit ones. Another conspicuous fact about this group of borrowers was when they become 11 or 18 people in the business, a quarter of these businesses will default. In the following table which businesses registered the most default is depicted.

Table 8. Default rate for $20.8K borrowers by business

Field of business	Default Rate
Beauty Salons	20,30%
Residential Remodelers	22,98%
Barber Shops	30,47%
All Other Personal Services	20,16%

Specialized Freight (except Used Goods) Trucking, Long-Distance	12,77%
All Other Miscellaneous Store Retailers (except Tobacco Stores)	18,32%
Independent Artists, Writers, and Performers	18,04%
Taxi Service	15,46%
Offices of Real Estate Agents and Brokers	12,70%
Caterers	18,92%
General Freight Trucking, Local	16,10%
Janitorial Services	16,75%
Landscaping Services	23,90%
All Other Personal Services	19,55%
Local Messengers and Local Delivery	17,39%
Wholesale Trade Agents and Brokers	15,46%
Home Health Care Services	13,74%
Child Day Care Services	13,79%
Women's Clothing Stores	17,92%

For the second group of the borrowers a meaningful difference between male or female owner, veteran or non-veteran owner, for profit or non-profit business, is not observed. Also, low document loan (LMI-Indicator) and coming from high priority or not (Hub Zone Indicator) did not show a gap for default borrowers. In this category Las Vegas followed by New York City by 9.38% and 9.05% registered the highest rate of default. A bigger picture when we look at state of the borrower is drawn in below.

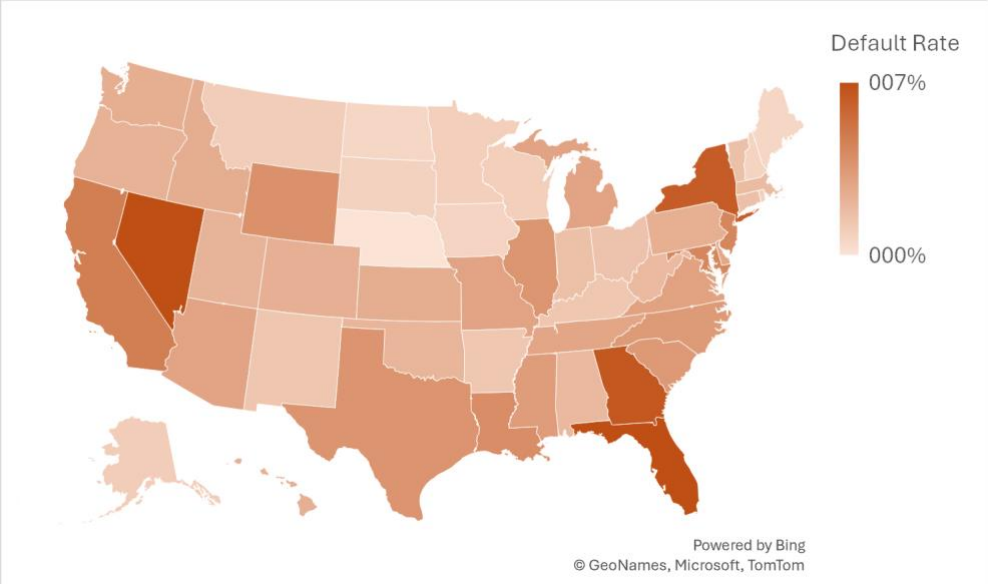


Figure 39. The default rate for $150K~1M loan borrowers by state

Generally, the rate of the default for this group of the borrowers is much lower than previous one. Again, the northern states show resilience during pandemic and the rate of default is low. In the center of the country, Nebraska has the lowest rate of default. This state has almost 2 million population and stands in the middle by ranking of the average annual income, yet these characteristics have nothing much to deduct any result for the low rate of default on loan for the other states. Figure 40 shows the default rate on the lender side for this group.

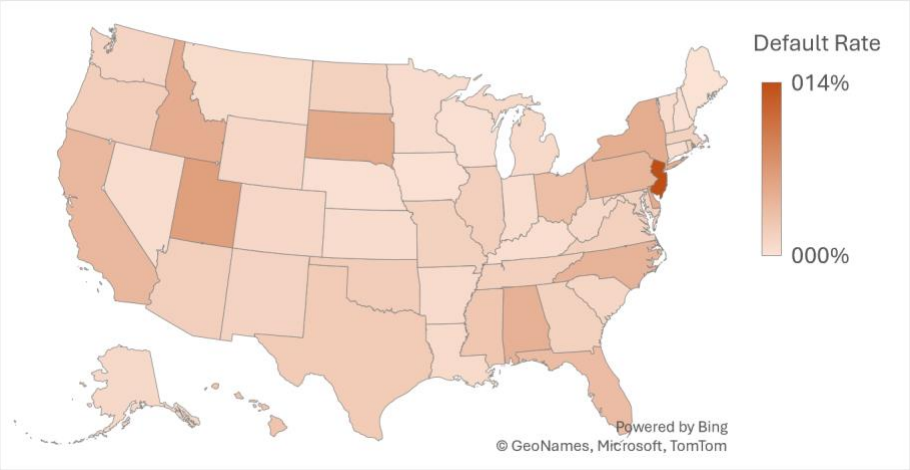


Figure 40. The default rate for the lenders of $150~1M loan by state

New Jersey has registered far ahead of any other states defaulted loans (13.63%); The next state is Nevada by almost half of New Jersey. Most lenders in the states recorded a fair rate of defaulted loan. Apparently, the aid-program has done well for these businesses to keep themselves upright during the pandemic. In this group the self-employed individual borrowers defaulted 37.93% of the loans. This rate has a yawning gap by the any other business type. This is what we observed also in the $20.8K borrowers, individuals default more than collectives. Loan borrowers with 3 to 4 years had the highest rate of default while 2 years borrowers showed promising results and paid their loan back. There 230 firms with 10 employees defaulted on the loan which accounts for 7.8% default for all the peers and is the highest rate compared to any other number of jobs reported by the businesses.

For the big amount borrowers, the pattern and shades for default rate is pretty deviated from the former group.

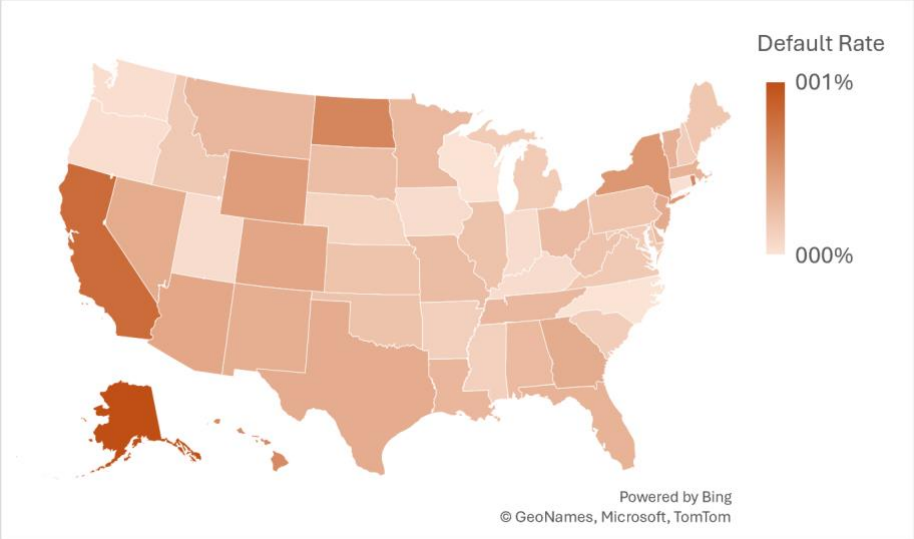


Figure 41. Default rate for the $1~10M borrowers by state

Here the state of Alaska which had a very low default rate amongst the other groups, stands on the top. Although 1.24% still is not high, when it comes to million dollars then it can be noticeable. California is behind Alaska by 0.23% less and North Dakota is the third from the top. At the lender side, unfortunately dataset does not include this variable to visualize the default rate registered by the banks. Borrowers from the city 'Fullerton' in California had the highest rate of default by 8.2%. In this group almost 1 out of 1000 companies with 100 employees defaulted the loan (highest rate). The table of which type businesses had the most default rate is in the below.

Table 9. Default rate by business type for $1~10M borrowers

LLC	Corporation	Sub S Corp	Non-Profit Organization	Solo Partner	Partenship	LL Parternship	Professional Association
0,58%	0,52%	0,38%	0,01%	1,68%	0,46%	0,16%	0,33%

Other findings

All three group of borrowers were looked through different lenses. The political orientation of each governor and state observed. As the PPP loans were dispatched in both 2020 and 2021, there was a guess that the presidential power transition from republican to democrats could perhaps affect the loans and defaulted rates. By looking at the maps and following the results, no political interest was seen amongst the states. It can be confidently said that the power transition did not affect the aid program as well. The population of the states, and the average

income of a person in the states had no bold impact on the default rate of the loan as well. The geographically position of the states had an impact in the loan default rate, which does make sense when we look back the places where dense touristic resorts and many lockdowns during pandemic happened. By moving toward coastlines, the probability of loan default by businesses increases.

Research Limitations

Like any other research, time bonded and deadlines for investigations are a matter of hinderance. There could be done much more with data analysis and using other models and packages to scrutinize the findings could be done in the bigger time framework. Yet as a Master thesis and finish the analysis in the right time, there was no more room to continue for further analytics. Another bold issue was the capacity of the computer to handling the datasets. Processing the models with Shap package to get the interpretable results and more comprehensive outcomes is a time taking as well as complicating task. The personal computer, which was used for such a purpose, could not handle all data at ones sometimes. For example, for the small borrowers of up to $150K, using the chunk of concentrated of data for the borrower and making a few cuts was mandatory to get the research done (though it made a quite fair comparison in this group for the businesses). Besides that, to train, test and get the result of Shap codes, each time the author had to wait, sometimes up to an hour to get the outcome. Many times, a small change and/or adding a variable to observe the impacts on the results demanded to be enormously meek. By improving the technology every day and introducing more powerful configurations in computer such advanced CPUs and GPUs, these investigations will get settled faster and more reliable.

At the beginning of the semester, the author had a huge hope that could get a database from a financial institution or related organization to perform domestic research. As living and studying in Aalborg, Denmark, correspondences were done by banks, Danmarks Statistics, Danmarks Nationalbank, yet all the efforts did not lead anywhere. Also, in a bigger scope European Central Bank did not decline the request to provide a dataset for the research, but dataset was not very informative to conduct valid research; Further contact with this

organization was on the road to nowhere as well. Copenhagen Business School openly accept the researcher to use its database, however the needed data had not been available anymore, due to the termination the contract between CBS and Moody (the company which works with risk credit and credit scoring). SMEs area is already not a low hanging fruit to study and investigate. Moody and its peers such as Experian, are collecting these data from financial institutes to sell for exorbitant prices (at least smallest package they offer is out of the researcher's affordability) and it seems that this market need to be monitored by authorities.

To know about their behavior and tackle the challenges and remove their barriers ahead of SMEs, there is a strong need for more cooperation between the commercial banks and also governmental financial institutes with universities. Without them, innovation and production will damage and thereby the cornerstones of any economy. If the limited resources will not allocate properly to the creditworthy small businesses, devastating consequences must be expected.

Even the available data which were found after surfing hundreds of webpages, and examined here, is not completed in the SBA website yet (at the time of writing this paper). For example, the borrowers of $1~10M dataset has few variables than the other two groups of data. As an instance the variable 'Term' which showed interesting results for default probability, is not presented for this group. Hope in the future the path for oncoming researchers is paved and they only put their efforts on getting robust results not get distressed on collecting data. There is huge space to probe and dig more into this field of study. Data for SMEs and the loans borrowed by them in the normal condition and under booming economic situation could increase the reliability of models. Here we looked at only US SMEs PPP loan borrowers during pandemic. Analyzing the data of these borrowers, for the loan without federal government guarantee could teach us other aspects of adversities and prosperities for this group. Also looking at the other geographical places could be intriguing too. Even the presented data, which is available as mentioned, can be examined in other ways. There are plenty of models and programs by which the behavior of the SMEs could be assessed. There is no flawless research or model, as this one is not either; Therefore, the findings of the other could improve the results or models that are explained here. Many issues are still obscure in these datasets, for example one can focus on the correlation of the term and the amount of the loan with more detail to discover new areas on the loan status (as these two features appeared important in this research).

If EU wants to shade its role in the economy of the world, publishing such a dataset to be studied looks in a dire need at the moment. The bottom line is that SMEs deserve more considerations, and the availability of their data is a necessity.

Conclusion

In this research the dataset of three groups of SME loan borrowers in the United States under an aid-program were assessed by machine learning algorithms, followed by explainable machine learning (XAI) methods. The findings showed that XGBoosting and Logistic Regression have the most accuracy of default probability prediction. XGB precision was slightly better than LR in the experiments. It revealed that for all three groups of the loan borrowers, number of employees, industry in which they operate and the amount of money that they borrowed are important features for default probability prediction. The interesting important feature here was the term of the loan. Less than one-year borrowers, were plausible to default the loan but if the loan term was two years, most of the loan got fully paid back. The trend between three to five years changed the direction again where many loans defaulted. It seems when SMEs have the optimistic approach (while it does not come true) or pessimistic approach toward passing the crisis period, here Covid19 pandemic, they will lose their appetite to keep running the business.

The XAI outcomes when the input model was XGB showed high reliable results. Besides revealing the important features, these methods provide remarkable insight about how much each variable can affect each side of prediction (default or fully payback). Shap package result presented more validity and reliability compared to LIME package when the model outcomes were checked by the actual data.

References

1. Bitteto, A. (2023). Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses. Socio economic planning science. Elsevier
2. Lu, Y. (2022). A novel framework of credit risk feature selection for SMEs during industry 4.0. Springer Nature
3. Bitteto, A. (2023). Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses. Socio economic planning science. Elsevier
4. Zoynul Abedin, & M. Hassan, K. (2021). Essential of machine learning in finance and accounting, Routledge publications
5. Nguyan, L. April (2020). Asian development bank
6. Galluci, C. (2023). Financial ratios, corporate governance and bank-firm information. Journal of Management and Governance
7. Chelagat, K N. (2012). Determinants of Loan Default for SMEs amongst commercial Banks in Kenya. University of Kenyia
8. Ranganati Matenda, & F. Sibanda, M. (2022). Determinants of default probability for audited and unaudited SMEs under stressed condition in Zimbabwe. Economies. Switzerland
9. Bussemann, N. & Gidici, P. (2020). Explainable Machine Learning in Risk Management. Computational economies
10. Ciampi, F. & Giannozi, A. (2021). Rethinking SMEs default prediction: A systematic litrature review and future perspective. Scientometric
11. Gogas, P. & Papadimtriou, T. (2021). Machine learning in economics and finance. Computational economics
12. Aniceto, M C. & Barboroza, F. June (2020). Machine learning predictivity applied to consumers creditworthiness, Future business journal
13. Hyuang, Y. & Zhang, L. (2020). Fintech credit risk assessment for SMEs: Evidence from China, IMF working paper
14. Kyeong, S. & Shin, J. (2022). Two stage credit scoring using Bayesian approach, Journal of big data

15. Sheng H, Y. & Subramanian, P. (2023). A systematic review of machine learning and explainable artificial intelligence in credit risk modeling. Springer

16. Rudin, S. Shaposhnik, & Y. (2023). Globally consistent rull based summary explainations for machine learning models: Applications to credit risk evaluation. Journal of machine learning research

17. Feng, Y. & Shen, Z. (2023). Machine learning based predictive models and drug prediction for Shizophernia in multiple programmed cell death patterns. Ntional library of medicine. USA

18. Kryzanowski. (1985). Small businesses debt finance: An emprical investigation of default risk

19. Biecek, P. Burzykowski, & T. (2021). Explanatory model analysis: Explore, explain and examine predictive models. 1st edition. CRC press

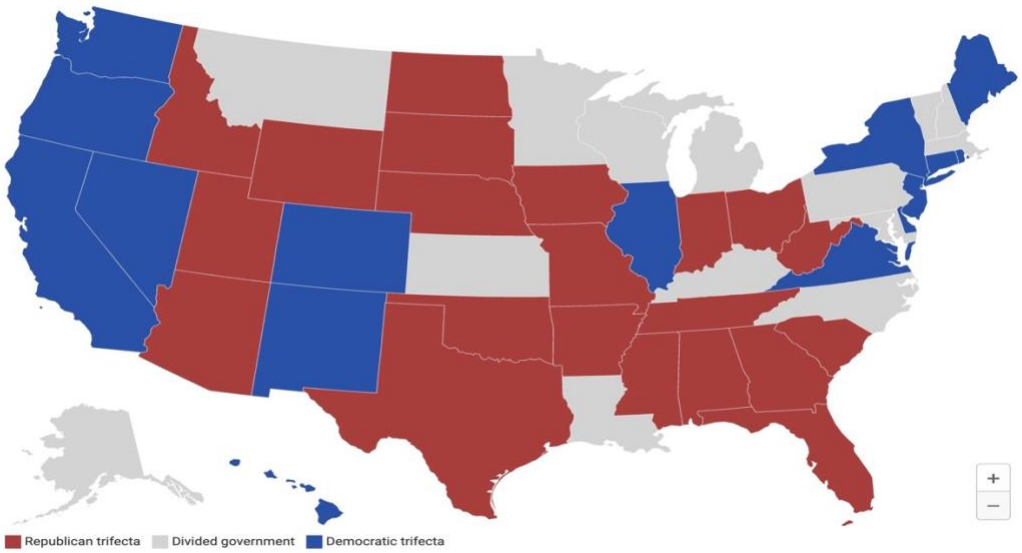
20. Jammalamadaka, K. (2023). Responsible AI in automated credit scoring system. AI and ethics

21. Modarres, C. & Louie, M. (2018). Towards expalinable deep learning for credit lending: A case study. Capital One

Appendix

Appendix A

Map of Political Orientation of USA in 2020 (Ballotpedia.pdf)



Appendix B

PPP Loans Dictionary (pandemicoversight.gov)

Field Name	Field Description
LoanNumber	Loan Number (unique identifier)
DateApproved	Loan Funded Date
SBAOfficeCode	SBA Origination Office Code
ProcessingMethod	Loan Delivery Method (PPP for first draw; PPS for second draw)
BorrowerName	Borrower Name
BorrowerAddress	Borrower Street Address
BorrowerCity	Borrower City
BorrowerState	Borrower State
BorrowerZip	Borrower Zip Code
LoanStatusDate	Loan Status Date - Loan Status Date is blank when the loan is disbursed but not Paid In Full or Charged Off
LoanStatus	Loan Status Description - Loan Status is replaced by 'Exemption 4' when the loan is disbursed but not Paid in Full or Charged Off
Term	Loan Maturity in Months
SBAGuarantyPercentage	SBA Guaranty Percentage
InitialApprovalAmount	Loan Approval Amount(at origination)
CurrentApprovalAmount	Loan Approval Amount (current)
UndisbursedAmount	Undisbursed Amount
FranchiseName	Franchise Name
ServicingLenderLocationID	Lender Location ID (unique identifier)
ServicingLenderName	Servicing Lender Name
ServicingLenderAddress	Servicing Lender Street Address
ServicingLenderCity	Servicing Lender City
ServicingLenderState	Servicing Lender State
ServicingLenderZip	Servicing Lender Zip Code
RuralUrbanIndicator	Rural or Urban Indicator (R/U)
HubzoneIndicator	Hubzone Indicator (Y/N)
LMIIndicator	LMI Indicator (Y/N)
BusinessAgeDescription	Business Age Description
ProjectCity	Project City
ProjectCountyName	Project County Name
ProjectState	Project State

Appendix C

PPP Loan Application Form, 1st page (pandemicoversight.gov)

Paycheck Protection Program
Borrower Application Form Revised March 18, 2021

OMB Control No.: 3245-0407
Expiration Date: 9/30/2021

Check One: ☐ Sole proprietor ☐ Partnership ☐ C-Corp ☐ S-Corp ☐ LLC ☐ Independent contractor ☐ Self-employed individual ☐ 501(c)(3) nonprofit ☐ 501(c)(6) organization ☐ 501(c)(19) veterans organization ☐ Other 501(c) organization ☐ Housing cooperative ☐ Tribal business ☐ Other _____	**DBA or Tradename (if applicable)**	**Year of Establishment (if applicable)**
Business Legal Name	**NAICS Code**	**Applicant (including affiliates, if applicable) Meets Size Standard** (check one):
		☐ No more than 500 employees (or 300 employees, if applicable) unless "per location" exception applies ☐ SBA industry size standards ☐ SBA alternative size standard
Business Address (Street, City, State, Zip Code - No P.O. Box addresses allowed)	**Business TIN (EIN, SSN, ITIN)**	**Business Phone**
	Primary Contact	**Email Address**

Average Monthly Payroll:	$	**x 2.5 + EIDL (Do Not Include Any EIDL Advance) equals Loan Request Amount:**	$	**Number of Employees:**	
Purpose of the loan (select all that apply):	☐ Payroll Costs	☐ Rent / Mortgage Interest	☐ Utilities	☐ Covered Operations Expenditures	
	☐ Covered Property Damage	☐ Covered Supplier Costs	☐ Covered Worker Protection Expenditures	☐ Other (explain): _____	

Applicant Ownership

List all owners of 20% or more of the equity of the Applicant. Attach a separate sheet if necessary.

Owner Name	**Title**	**Ownership %**	**TIN (EIN, SSN, ITIN)**	**Address**

PPP Applicant Demographic Information (Optional)

Veteran/gender/race/ethnicity data is collected for program reporting purposes only. Disclosure is voluntary and will have no bearing on the loan application decision.

Principal Name	Principal Position

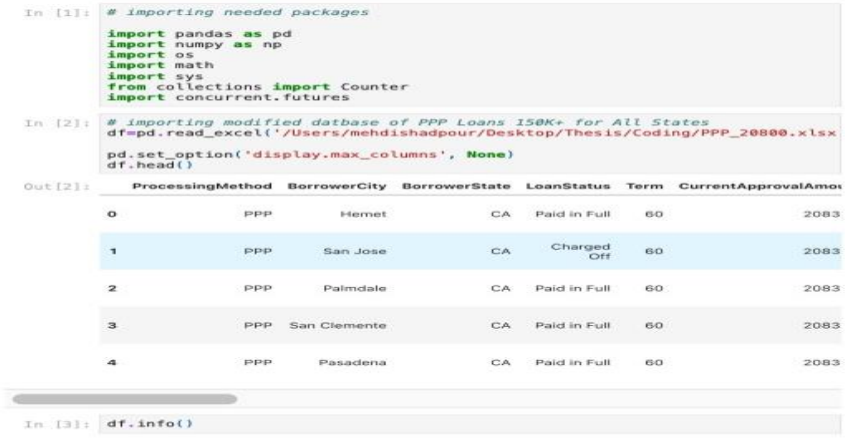
	Select Response Below:
Veteran	☐ Non-Veteran; ☐ Veteran; ☐ Service-Disabled Veteran; ☐ Spouse of Veteran; ☐ Not Disclosed
Gender	☐ Male; ☐ Female; ☐ Not Disclosed
Race (more than 1 may be selected)	☐ American Indian or Alaska Native; ☐ Asian; ☐ Black or African-American; ☐ Native Hawaiian or Pacific Islander; ☐ White; ☐ Not Disclosed
Ethnicity	☐ Hispanic or Latino; ☐ Not Hispanic or Latino; ☐ Not Disclosed

SBA Form 2483 (3/21)

1

Appendix D

A Sample of Python Codes



```
In [1]: # importing needed packages
        import pandas as pd
        import numpy as np
        import os
        import math
        import sys
        from collections import Counter
        import concurrent.futures

In [2]: # importing modified datbase of PPP Loans 150K+ for All States
        df=pd.read_excel('/Users/mehdishadpour/Desktop/Thesis/Coding/PPP_20800.xlsx

        pd.set_option('display.max_columns', None)
        df.head()
```

Out[2]:

	ProcessingMethod	BorrowerCity	BorrowerState	LoanStatus	Term	CurrentApprovalAmo
0	PPP	Hemet	CA	Paid in Full	60	2083
1	PPP	San Jose	CA	Charged Off	60	2083
2	PPP	Palmdale	CA	Paid in Full	60	2083
3	PPP	San Clemente	CA	Paid in Full	60	2083
4	PPP	Pasadena	CA	Paid in Full	60	2083

```
In [3]: df.info()
```

Define the Loan Status as the dependent varibale and others as independent variables

y=df['Loan_Status']
X=df.drop(['Loan_Status'],axis=1)

#importing scientific kit packages

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=123)

#importing xgboost

import xgboost as xgb

xgb = xgb.XGBClassifier()

xgb.fit(X_train,y_train)

predictions = xgb.predict(X_test)

#Importing accuracy package to see the accuracy of the prediction

from sklearn.metrics import accuracy_score

xgb_score=accuracy_score(y_test, predictions)
xgb_score

Appendix E

NAICS (NORTH AMERICAN INDUSTRY CLASSIFICATION SYSTEM) Codes List

Sector 11. Agriculture, Forestry, Fishing and Hunting
Subsector 111. Crop Production
Subsector 112. Animal Production and Aquaculture
Subsector 113. Forestry and Logging
Subsector 114. Fishing, Hunting and Trapping
Subsector 115. Support Activities for Agriculture and Forestry
Sector 21. Mining, Quarrying, and Oil and Gas Extraction
Subsector 211. Oil and Gas Extraction
Subsector 212. Mining (except Oil and Gas)
Subsector 213. Support Activities for Mining
Sector 22. Utilities
Subsector 221. Utilities
Sector 23. Construction
Subsector 236. Construction of Buildings
Subsector 237. Heavy and Civil Engineering Construction
Subsector 238. Specialty Trade Contractors
Sector 31-33. Manufacturing
Subsector 311. Food Manufacturing
Subsector 312. Beverage and Tobacco Product Manufacturing
Subsector 313. Textile Mills
Subsector 314. Textile Product Mills
Subsector 315. Apparel Manufacturing
Subsector 316. Leather and Allied Product Manufacturing
Subsector 321. Wood Product Manufacturing
Subsector 322. Paper Manufacturing
Subsector 323. Printing and Related Support Activities
Subsector 324. Petroleum and Coal Products Manufacturing
Subsector 325. Chemical Manufacturing
Subsector 326. Plastics and Rubber Products Manufacturing
Subsector 327. Nonmetallic Mineral Product Manufacturing
Subsector 331. Primary Metal Manufacturing
Subsector 332. Fabricated Metal Product Manufacturing
Subsector 333. Machinery Manufacturing
Subsector 334. Computer and Electronic Product Manufacturing
Subsector 335. Electrical Equipment, Appliance, and Component Manufacturing
Subsector 336. Transportation Equipment Manufacturing
Subsector 337. Furniture and Related Product Manufacturing
Subsector 339. Miscellaneous Manufacturing
Sector 42. Wholesale Trade
Subsector 423. Merchant Wholesalers, Durable Goods
Subsector 424. Merchant Wholesalers, Nondurable Goods
Subsector 425. Wholesale Trade Agents and Brokers
Sector 44-45. Retail Trade
Subsector 441. Motor Vehicle and Parts Dealers
Subsector 444. Building Material and Garden Equipment and Supplies Dealers
Subsector 445. Food and Beverage Retailers

Subsector 449. Furniture, Home Furnishings, Electronics, and Appliance Retailers
Subsector 455. General Merchandise Retailers
Subsector 456. Health and Personal Care Retailers
Subsector 457. Gasoline Stations and Fuel Dealers
Subsector 458. Clothing, Clothing Accessories, Shoe, and Jewelry Retailers
Subsector 459. Sporting Goods, Hobby, Musical Instrument, Book, and Miscellaneous Retailers
Sector 48-49. Transportation and Warehousing
Subsector 481. Air Transportation
Subsector 482. Rail Transportation
Subsector 483. Water Transportation
Subsector 484. Truck Transportation
Subsector 485. Transit and Ground Passenger Transportation
Subsector 486. Pipeline Transportation
Subsector 487. Scenic and Sightseeing Transportation
Subsector 488. Support Activities for Transportation
Subsector 491. Postal Service
Subsector 492. Couriers and Messengers
Subsector 493. Warehousing and Storage
Sector 51. Information
Subsector 512. Motion Picture and Sound Recording Industries
Subsector 513. Publishing Industries
Subsector 516. Broadcasting and Content Providers
Subsector 517. Telecommunications
Subsector 518. Computing Infrastructure Providers, Data Processing,
Web Hosting, and Related Services
Subsector 519. Web Search Portals, Libraries, Archives, and OtherInformation Services
Sector 52. Finance and Insurance
Subsector 521. Monetary Authorities-Central Bank
Subsector 522. Credit Intermediation and Related Activities
Subsector 523. Securities, Commodity Contracts, and Other Financial
Investments and Related Activities
Subsector 524. Insurance Carriers and Related Activities
Subsector 525. Funds, Trusts, and Other Financial Vehicles
Sector 53. Real Estate and Rental and Leasing
Subsector 531. Real Estate
Subsector 532. Rental and Leasing Services
Subsector 533. Lessors of Nonfinancial Intangible Assets (exceptCopyrighted Works)
Sector 54. Professional, Scientific, and Technical Services
Subsector 541. Professional, Scientific, and Technical Services
Sector 55. Management of Companies and Enterprises
Subsector 551. Management of Companies and Enterprises
Sector 56. Administrative and Support and Waste Management and Remediation Services
Subsector 561. Administrative and Support Services
Subsector 562. Waste Management and Remediation Services
Sector 61. Educational Services
Subsector 611. Educational Services
Sector 62. Health Care and Social Assistance
Subsector 621. Ambulatory Health Care Services

Subsector 622. Hospitals
Subsector 623. Nursing and Residential Care Facilities
Subsector 624. Social Assistance
Sector 71. Arts, Entertainment, and Recreation
Subsector 711. Performing Arts, Spectator Sports, and Related Industries
Subsector 712. Museums, Historical Sites, and Similar Institutions
Subsector 713. Amusement, Gambling, and Recreation Industries
Sector 72. Accommodation and Food Services
Subsector 721. Accommodation
Subsector 722. Food Services and Drinking Places
Sector 81. Other Services (except Public Administration)
Subsector 811. Repair and Maintenance
Subsector 812. Personal and Laundry Services
Subsector 813. Religious, Grantmaking, Civic, Professional, and Similar Organizations
Subsector 814. Private Households
Sector 92. Public Administration
Subsector 921. Executive, Legislative, and Other General Government Support
Subsector 922. Justice, Public Order, and Safety Activities
Subsector 923. Administration of Human Resource Programs
Subsector 924. Administration of Environmental Quality Programs
Subsector 925. Administration of Housing Programs, Urban Planning,
and Community Development
Subsector 926. Administration of Economic Programs
Subsector 927. Space Research and Technology
Subsector 928. National Security and International