MISM 6212 Data Mining Project

Northeastern University
Prof. Xavier Babu
23rd April, 2022
Team 10: Sayantani Roy

Contents

1. Abstract:

In the United States, small businesses have long been a primary driver of job generation. As the number of small businesses grows, so does the level of competition. As a result, many businesses will seek bank loans to solve their capital problems and boost their growth. When these businesses run into financial difficulties, applying for bank loans is generally less expensive in the long run than alternative means of generating capital like equity. Under normal circumstances, banks are more likely to lend to enterprises based on a particular profitability level and stature. Bank personnel will generally visit the business to conduct a survey that goes in consideration for the loan approval; however, this generally has subjectivity as each surveyor might have a different perspective of the risks associated with businesses. As most small businesses fail by the third year, banks take several metrics in consideration when assessing a business for their loan application. In fact, the bank will examine the enterprise's debt ratio. If the debt ratio is excessively high, it indicates that the company will be unable to repay the loan. Moreover, from a risk perspective, they will reduce the loan amount willing to offer for firms with a high debt ratio. There is data with several such metrics that go into consideration to help banks with loan decisions.

The U.S. Small Business Administration (SBA) was formed with the goal of fostering the establishment and expansion of small businesses and has a positive social impact by increasing job opportunities and lowering unemployment. As SBA guarantees a fraction of the principal loan amount, banks need to write off the debt if a business defaults on its partially SBA-guaranteed loan. Hence, banks are liable to make the difficult decision as to whether they should approve a loan even with the high risk of failure associated with small businesses

2. Introduction:

As more and more small businesses are applying for business loans with SBA guarantee program it is difficult for the bankers to take the decision whether to approve or reject the loan. One way to inform and improve their decision making is through analyzing relevant historical data such as the datasets provided here and creating a machine learning model to assist the decision making.

In this project, I shall use a dataset from small businesses in the United States, leveraging data mining techniques like logistic regression, decision tree, SGD classifier etc. to evaluate the risk of default in order to approve or deny loans to small firms and also to predict the best model for this business case.

3. Literature Review:

Information about the Bank Loan Industry and Infrastructure:

Bank loans are among the most popular types of financing for small and medium enterprises (SMEs). They are often offered over a set period and are a quick and easy approach to get the essential funding. Bank loans can be designed to meet the needs of the business and can be capital/principal repayment or interest-only. A series of activities leads to the approval or rejection of a bank loan application during the lending process. A bank's loan department employs a variety of credit specialists, each with their own set of functions and responsibilities that work together to complete the lending process.
Loan analysis is a method of determining if loans are given on feasible terms and whether potential lenders can and are willing to repay the loan. It examines the potential borrower's eligibility against the lending requirements.
The following are the most important traits that most prospective lenders will look for:

- Credit History - personal and business credit score can be obtained through credit bureaus: Experian, Equifax and Dun & Bradstreet

- The company's cash flow history and projections - Financial statements like balance sheet, cash flow statement, income statement.
- Available collateral to secure the loan - Accounts Receivable, Inventory, Real Estate and Equipment
- A plethora of loan documents, such as business and personal financial records, income tax returns, and a business plan, which essentially summarizes up and gives evidence for the first four items listed.
- Business Plan - Documentation of the business strategy and how the loan proceeds will be allocated to help the business.

Loan Application Process: Applying for business loans is still conventional, where the applicant needs to schedule an appointment with a bank officer to provide relevant documentation and interview for the loan. While much of the banking process, like applying for credit cards, has been automated, business loans still require the human component of collecting data due to the size of the loans requested, and the probability of fraud/default.
Loan Eligibility Prediction: Loan lending has always been an important component of both organizations' and people' daily lives. With ever-increasing financial competitiveness and a large quantity of financial restraints, the practice of getting a loan has become rather unavoidable. Companies all over the world rely on loan lending for a variety of reasons. Banks and other small to large businesses rely on loan lending for the essential focus on managing their operations and trying to maximize during times of financial restriction.
Predicting how likely a customer is to default on a loan is always a fascinating and difficult topic for banks when they only have a few pieces of information. In the modern era, data science teams in banks use machine learning to build prediction models. The datasets they employ are almost certainly proprietary and are typically gathered internally as part of their day-to-day operations.

Despite the fact that it is extremely profitable and advantageous to both lenders and borrowers. Nevertheless, it carries a significant risk in the loan lending industry. Credit scores are numerical values assigned to individuals by industry experts and researchers all over the world to assess risk and creditworthiness.

Machine learning techniques have been used for years to calculate and predict credit risk by analyzing an individual's historical data. Our study examines the current research about how the bank wants to estimate which clients would default on their debts based on their financial information for this assignment by using Machine Learning algorithms.

Real-world Scenario of implementing ML algorithms for Loan Prediction:

Case: Biz2Credit offers Biz2X, a fully managed digital lending platform which assists financial institutions in extending loans to small enterprises. It also lends to small enterprises, having funded $2.5 billion in loans to date. Biz2Credit is supported by Nexus Venture Partners and has closed a $52 million Series B capital funding transaction managed by WestBridge Capital.

One of the issues that Biz2Credit has attempted to address with its Biz2X platform is the length of time it takes to process a loan application. In a traditional bank, numerous departments, including credit, enrollment, and operational, are engaged in processing a loan application, which can result in a significant turnaround time of 7-10 days – especially if there are different systems and silos among departments. Biz2X offers a unified platform for increasing the efficiency of the loan application procedure, delivering a 3-5X reduction in response time.

The use of Machine Learning to transform and evaluate client information is a critical component of Biz2Credit's capacity to increase loan application process efficiency. This begins when the firm files its application, which contains, among other things, Know Your Customer (KYC) information to identify and authenticate the company's license, executives, authorized signatories, ownership group, and current bank statements. The Biz2X Platform, using Amazon Rekognition, can extract relevant fields from a scanned image of a customer's identification document, such as the tax payer ID, which it can subsequently validate using an online API. In contrast, in a normal bank, an onboarding team must manually record and verify a customer's specific information into a CRM system.

Biz2Credit also employs a proprietary AI/ML algorithm to assess applicants' creditworthiness based on financial records or transaction information. The model calculates the enterprise's credit risk by estimating the applicant's revenue, expenses, and seasonality. According to Biz2Credit, their bad-loan percentage is one-third of the industry average, allowing banks who utilize its Biz2X platform to provide credit with less risk.

Global Recession : The expansion of recession began in the 1990s and continued unabated through the 2001 recession, accelerating in the mid-2000s. Average home prices in the United States more than doubled between 1998 and 2006, the sharpest increase recorded in US history, and even larger gains were recorded in some regions. Mortgage debt of US households rose from 61 percent of GDP in 1998 to 97 percent in 2006 and businesses were equally impacted

Global Environment: Implication of the COVID-19 pandemic on small business loans
Due to the difficulties small businesses suffered during the pandemic, SBA launched the Paycheck Protection Program (PPP), which enabled businesses to cover overheads, maintain payroll and keep business afloat during the turmoil through loans of up to $10 million. According to SBA, more than 11.5 million loans were disbursed equating to almost $800 billion US dollars. Given the right circumstances, the loans would be forgiven or if not fully forgiven, only 1 percent interest would be charged. As of March 27 2022, 89% of the total PPP loans disbursed had been forgiven. This program however, also attracted fraudulent practices as there were more than 100 cases of suspiciously acquired PPP loans equating to over $75 million USD.

Rising inflation rates: While the PPP loan program is over and much of the effects of the pandemic has subsided, issues arise from growing inflation as the global supply chain issues and political conflict has led to sharp rise in costs of raw materials, leading to businesses looking towards debt solutions to aid economic hardship. This will also more likely lead to greater default rates, as was the case during the 2008 recession, as shown in the graph below which is extracted from one of the exploratory analysis in the visualizations section. Please review reference section for details.

4. Data Source and Processing

Dataset: The dataset is collected from SBA website. Please refer appendix column for the datalink.13
This data set includes several columns with useful variables including Location (State), Industry, Gross Disbursement, New versus Established Business, Loans Backed by Real Estate, SBA's Guaranteed Portion of Approved Loan. I intend to learn which statistics are critical for small firms seeking loans from the pool of these variables, work on exploratory analysis and propose a ML model for future prediction. There are 27 columns and 899,164 rows in the dataset. Few of the notable variables are as below:

Variable name	Data type	Description of variable
LoanNr_ChkDgt	Text	Identifier – Primary key
Name	Text	Borrower name
City	Text	Borrower city
State	Text	Borrower state
Zip	Text	Borrower zip code
Bank	Text	Bank name
BankState	Text	Bank state
NAICS	Text	North American industry classification system code
ApprovalDate	Date/Time	Date SBA commitment issued
ApprovalFY	Text	Fiscal year of commitment
Term	Number	Loan term in months
NoEmp	Number	Number of business employees
NewExist	Text	1 = Existing business, 2 = New business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise code, (00000 or 00001) = No franchise
UrbanRural	Text	1 = Urban, 2 = rural, 0 = undefined
RevLineCr	Text	Revolving line of credit: Y = Yes, N = No
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Currency	Amount disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan status charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approved loan

Sector	Description
11	Agriculture, forestry, fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31–33	Manufacturing
42	Wholesale trade
44–45	Retail trade
48–49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services (except public administration)
92	Public administration

5. **Data Cleaning:**

There were no strong correlation between variables initially as we can see the below plot

There were some missing values in our dataset. Since I have almost 900,000 rows, I can remove rows with missing values rather than imputing them. Also, it is difficult to impute whether the business was new or existing without proper information and this could potentially lead to model error. As an example, it is difficult to infer whether the business was new or existing without knowing and which could potentially give model error in future.

```
[ ] # Drop null values from specified columns
    data.dropna(subset=['Name', 'City', 'State', 'BankState', 'NewExist','RevLineCr', 'LowDoc', 'DisbursementDate', 'MIS_Status'], inplace=True)
    data.isnull().sum()
```

Furthermore, I decided to create the amount of the loan that the SBA guaranteed. This is a unique feature of SBA loans in which the SBA will 'guarantee' a % of the loan if it is lost. For example, if a company takes out a 1,000,000 loan and the SBA guarantee 50% of the loan, if the company is unable to repay $400,000 of the loan, the SBA will cover $200,000 of the loss. This makes these loans particularly appealing to small firms because it reduces their risk, but that also raises the SBA's risk.

Currency has a '$' symbol and commas included in some columns, it appears that these are represented as strings. Thus, I altered the datatype to float without deleting those and also cleaned the data using lambda function.

```
[12] # Remove '$', commas, and extra spaces from records in columns with dollar values that should be floats
     data[['DisbursementGross', 'BalanceGross', 'ChgOffPrinGr', 'GrAppv', 'SBA_Appv']] = \
     data[['DisbursementGross', 'BalanceGross', 'ChgOffPrinGr', 'GrAppv', 'SBA_Appv']].applymap(lambda x: x.strip().replace('$', '').replace(',', ''))
```

Column 'ApprovalFY' has a combination of integers and strings, and one even terminates with an A. Following that, I would clear this section using if function

NAICS codes designate which industry each business requesting the loan application. The dataset has first two digits but I am more concerned with the general industry. So, I created the 'Sector' variable mapped to specific industries in my dataset.

I also created a flat column (0,1) for 'FranchiseCode' where franchisecode <=1 is tagged as 0, and >1 is tagged as 1

NewExist is about whether the business existed before. The values were initially 1 or 2 but I proceeded to change them to 0, for the business did not exist, and 1 for it existed before.

RevLineCr and LowDoc columns had multiple unique values but I really only need 'Y' for Yes and 'N' for No. In this case, I removed all these rows which had values other than Y or N and subsequently changed the values of the remaining rows from Y and N to 1 and 0 respectively.

For ApprovalDate and DisbursementDate I converted these variables to datetime values to create a new variable that can help find days between approval and disbursement.

I decided to create my target variable from the 'MIS_Status' column, as the two values in this column designated whether the loan was fully repaid('PIF') or defaulted ('ChgOff'). This binary variable helps create our logistic regression model(s).

```
[ ] data['Default'] = np.where(data['MIS_Status'] == 'P I F', 0, 1)
    data['Default'].value_counts()

    0    358558
    1     98382
    Name: Default, dtype: int64
```

I incorporated the number of days it took to disbursement once approved. My hypothesis is that the timing at which funds were received could have a negative relationship with business's ability to repay the loan, which means the longer it took to receive funds, the more difficult it would be to pay off the loan.

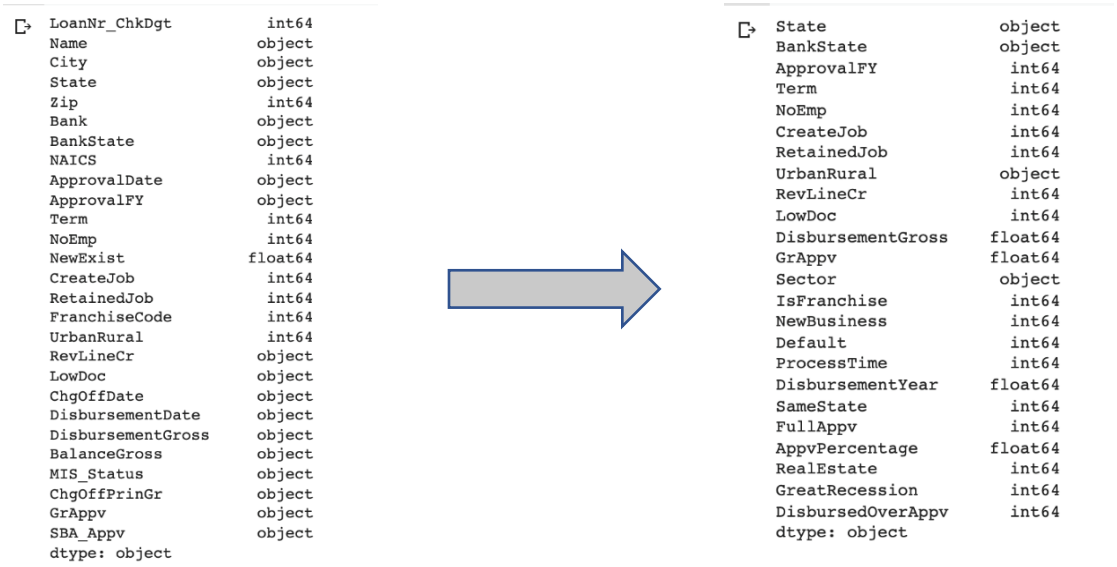
```
[26] # Create ProcessTime column which calculates the number of days passed between DisbursementDate and ApprovalDate
     data['ProcessTime'] = data['DisbursementDate'] - data['ApprovalDate']

     # Change ProcessTime from a timedelta64 dtype to an int64 dtype
     # Converts series to str, removes all characters after the space before 'd' in days for each record, then changes the dtype to int
     data['ProcessTime'] = data['ProcessTime'].astype('str').apply(lambda x: x[:x.index('d') - 1]).astype('int64')
```

I also incorporated the SameState flag field to check if the business state is same as the BankState. My assumption is that it would be difficult to service a loan for a business which is not in the same state and this could also negatively impact the business's ability to repay the loan.

I wanted to look at are whether a loan was backed by Real Estate, and whether a loan was active during the Great Recession (2007-2009). These were both mentioned in the document which describes the dataset and how it was used for educational purposes, and I think they will be very important features to consider. To determine whether a loan was backed by Real Estate, I made a flag that signifies if the loan term is >= 20 years, as real estate-backed loans are typically at least this long since the loan term is usually tied to the useful life of the assets used for collateral. Unfortunately, there's no way to know this for sure since it is not included explicitly in the data.

For loans active during the Great Recession, I created a flag for loans where the Great Recession (2007-2009) between DisbursementFY and DisbursementFY plus the loan term (in years).

Below is the initial data type and final data type after all the data cleaning, deleting rows and conversion:



```
LoanNr_ChkDgt          int64
Name                  object
City                  object
State                 object
Zip                    int64
Bank                  object
BankState             object
NAICS                  int64
ApprovalDate          object
ApprovalFY            object
Term                   int64
NoEmp                  int64
NewExist             float64
CreateJob              int64
RetainedJob            int64
FranchiseCode          int64
UrbanRural             int64
RevLineCr             object
LowDoc                object
ChgOffDate            object
DisbursementDate      object
DisbursementGross     object
BalanceGross          object
MIS_Status            object
ChgOffPrinGr          object
GrAppv                object
SBA_Appv              object
dtype: object
```

```
State                 object
BankState             object
ApprovalFY             int64
Term                   int64
NoEmp                  int64
CreateJob              int64
RetainedJob            int64
UrbanRural            object
RevLineCr              int64
LowDoc                 int64
DisbursementGross    float64
GrAppv               float64
Sector                object
IsFranchise            int64
NewBusiness            int64
Default                int64
ProcessTime            int64
DisbursementYear     float64
SameState              int64
FullAppv               int64
AppvPercentage       float64
RealEstate             int64
GreatRecession         int64
DisbursedOverAppv      int64
dtype: object
```

Finally we have a total of 438,504 entries or rows and 23 columns. Though we have deleted rows it is still quite rich dataset to work on.

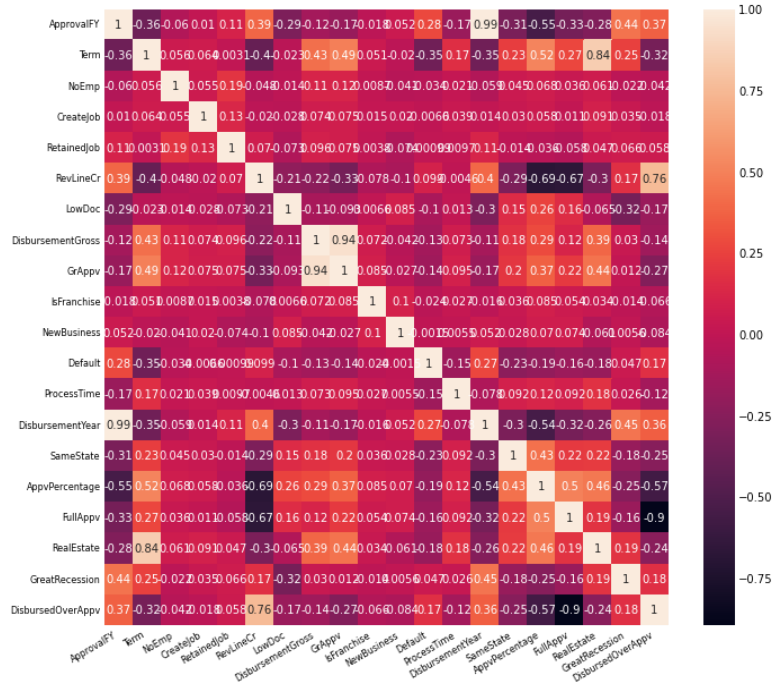
6. Exploratory Analysis:

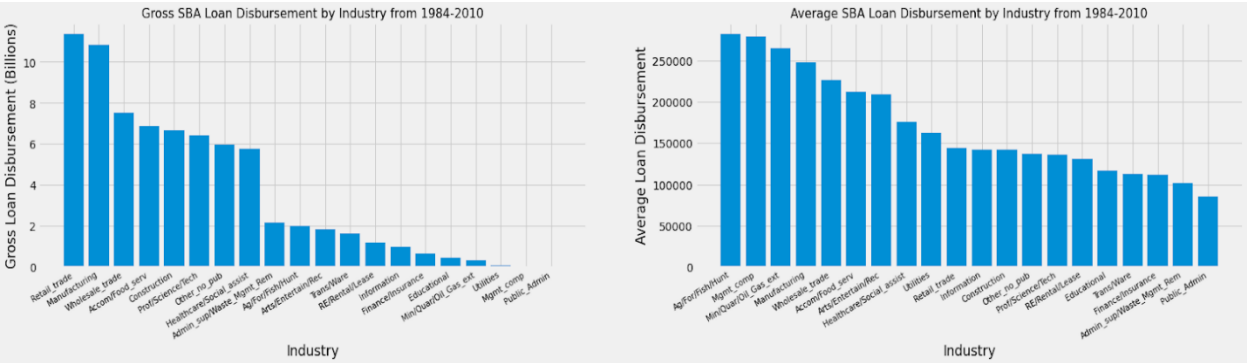
Some interesting information I found which could be useful for business decisions:

- The average loan term is ~94 months with a standard deviation of ~69 months, suggesting the loan terms are pretty spread out;
- Max loan term of 527 months could suggest some outliers in the data. The average number of employees is about 9.8 with 75% of businesses having 9 or less employees, suggesting NoEmp is very left skewed; Similar situations for created and retained jobs
- The mean for flag fields essentially shows a percentage, so roughly 42% of loans in the sample are revolving lines of credit and about 6% of loans were a part of the Low Doc program
- Average gross loan disbursement was ~166,000 with 75% of loans being less than 188,000, suggesting left skewness. About 77.8% of loans in the sample were paid in full Only 3% of businesses were franchised.
- About 26% of loan applicants were considered new businesses. The average days to loan disbursement was 109. The min was -3,614, suggesting at least one error in the data (since that's ~301 years)
- Approximately 45.4% of loans were serviced by banks in the same state as the applying business
- The average percentage of SBA loan guaranteed amount was 65.4%
- About 11.2% of the loans backed by real estate per my assumptions
- About 73.4% of the loans in the sample were active at some point during the Great Recession

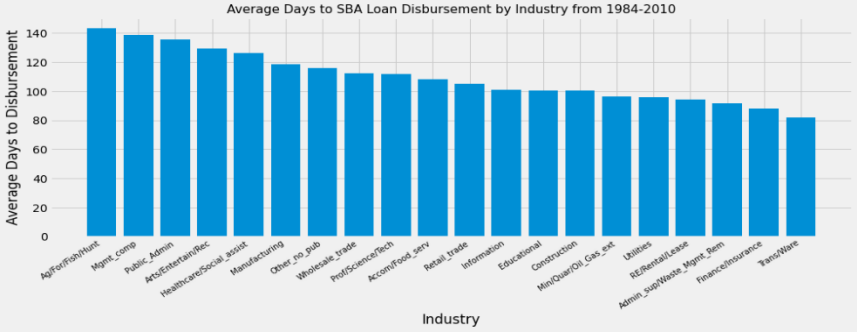
7. Data Visualizations:

Below is the correlation matrix with the cleaned data. We can see the change from our initial matrix with stronger correlation between variables as some variables have been removed, while others have been formatted or added. GrAppv & DisbursementGross has positive correlation, DisbursedOverAppv & FullAppv has negative correlations. DisbursedOverAppv & FullAppv, Negative -RevLineCr & DisbursedOverAppv are positive while DisbursementYear & ApprovalFY, Positive -FullAppv & RevLineCr, Negative -AppvPercentage & RevLineCr are negative.





From the above graphs we can notice that during 1984-2010, retail trade, manufacturing industries have highest gross loan disbursements followed by wholesale, food, construction industries. On the contrary Agri or fish has the highest average loan disbursement amount followed by management of companies or services, Oil & Gas, manufacturing



From this graph we can notice that processing time can defer to almost 2 months depending on the industry the business is based on. Some of the industries with the highest average loan amount also had the highest number of days to disbursement of funds, including the Agriculture, forestry, fishing and hunting, and Management of companies and enterprises industries.

Based on both the graph I believe it is safe to suggest banks to consider the loan disbursement days for the Agriculture and Fish industry as it takes the highest loan from the bank. The delay can create dissatisfaction among the customers. Another alarming thing I can notice is that though public admin has lowest loan amount disbursement still it takes a lot of time to disburse. Banks could look into these aspects.

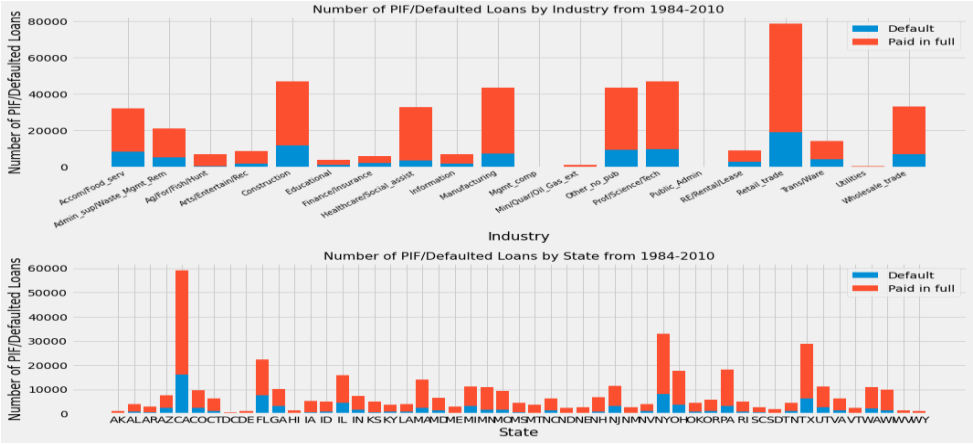
I also Checked default percentage by sector as below :

Sector			
Accom/Food_serv	23936	8381	0.259337
Admin_sup/Waste_Mgmt_Rem	15774	5427	0.255978
Ag/For/Fish/Hunt	6536	657	0.091339
Arts/Entertain/Rec	6976	1917	0.215563
Construction	34999	12048	0.256084
Educational	2750	1070	0.280105
Finance/Insurance	3984	2093	0.344413
Healthcare/Social_assist	29192	3571	0.108995
Information	5222	1830	0.259501
Manufacturing	36448	7281	0.166503
Mgmt_comp	90	23	0.203540
Min/Quar/Oil_Gas_ext	1133	117	0.093600
Other_no_pub	34192	9351	0.214753
Prof/Science/Tech	37278	9803	0.208216
Public_Admin	151	29	0.161111
RE/Rental/Lease	6079	3097	0.337511
Retail_trade	59503	19051	0.242521
Trans/Ware	10016	4430	0.306659
Utilities	334	79	0.191283
Wholesale_trade	26224	7018	0.211118

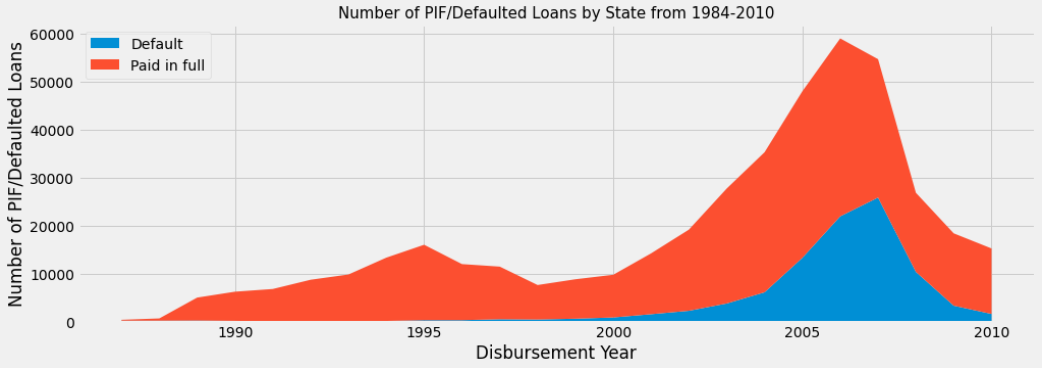
As per my analysis, industries with the highest number of loans during the sample period are Retail trade (78,554), Professional, Scientific and technical services (47,081) and Construction (47,047). Industries with the highest default percentage are Finance and Insurance (34.4%), Real Estate and rental leasing (33.8%) and Transportation and warehousing (30.7%).

States with the highest number of loans during the sample period are California (59,121), New York (33,059) and Texas (28,941). States with the highest Default percentage are Florida (33.8%), Arizona (32.6%) and Nevada (31.6%).

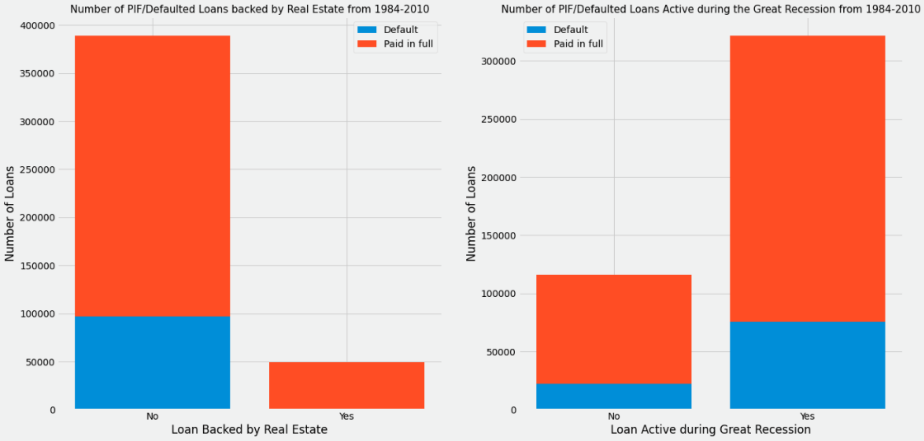
I am assuming because these metropolitan cities would have more business loans and hence more default rate (as pic below)



Based on the visualization below, there is a clear increase in loan volume leading up to the peak of the **Great Recession**, with a subsequent drop in loan volume immediately following that time. Looking at the graph, it appears the default rate of loans increased during that time as well.



The volume of loans backed by real estate was significantly less than those not backed by real estate, however the default rate is also lesser for loans backed by real estate as applicants might not want their collateral assets to be foreclosed



8. Methodology or Machine Learning Model:

As I have a large dataset (almost 500,000 entries after cleaning), I took a sample of 100,000 data in my train dataset . I have used 80-20 split for the dataset. I have also run the data with the entire sample and I did not see much difference in the model accuracy. I also created dummy variables for the categorical variable. Finally. I separated the

Model 1: Logistic Regression

I am working on a classification model and hence the first model came to my mind is to use the logistic regression. It's a type of regression analysis and is a commonly used algorithm for solving binary classification problems. Logistic regression algorithm can help predict the likelihood of events by looking at historical data points.

Logistic Regression Confusion Matrix :

```
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

```
[[15982  1059]
 [ 3176  1688]]
              precision    recall  f1-score   support

           0       0.83      0.94      0.88     17041
           1       0.61      0.35      0.44      4864

    accuracy                           0.81     21905
   macro avg       0.72      0.64      0.66     21905
weighted avg       0.79      0.81      0.79     21905
```

Accuracy score for **logistic regression is 81%.** I have not tuned or used cross validation on this model as the accuracy score is already quite high. Also, I wish to explore other models and compare the results. 0 denotes the not default rate and 1 is default rate

Model 2: Decision Tree

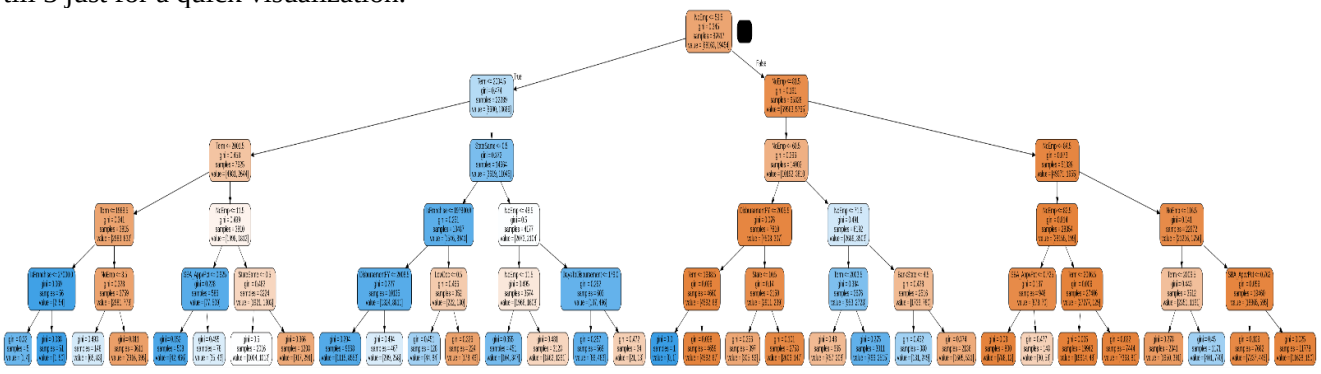
For decision tree I first fit the tree with default parameter and check till depth 5.

Confusion Matrix for decision tree :

```
# Let's check the evaluation metrics of our default model

# Importing classification report and confusion matrix from sklearn metrics
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score

# making predictions
y_pred_default = dt_default.predict(x_test)

# Printing classifier report after prediction
print(classification_report(y_test,y_pred_default))
```

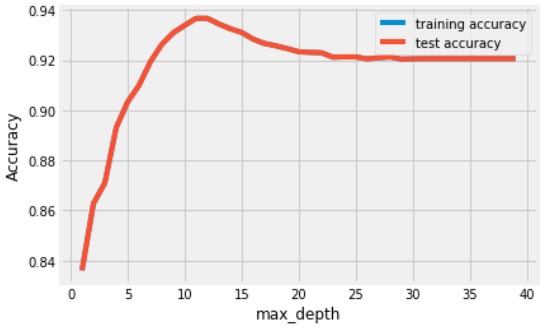
```
              precision    recall  f1-score   support

           0       0.93      0.94      0.94     17041
           1       0.79      0.76      0.78      4864

    accuracy                           0.90     21905
   macro avg       0.86      0.85      0.86     21905
weighted avg       0.90      0.90      0.90     21905
```

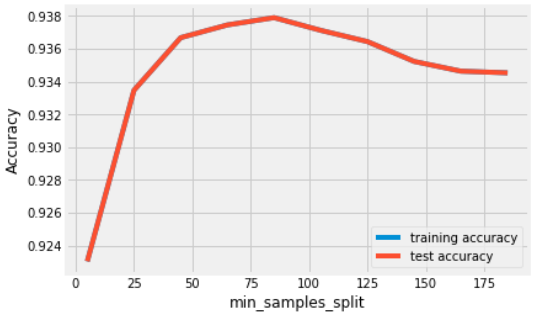
My initial Decision tree score is 90 %. This is an improvement on the logistic regression model. The precision and recall score also improved as we can see above. Below is the decision tree with max depth till 3 just for a quick visualization.



Please refer to my code (attached in the appendix) for better visualization and clarity.

I used 5-fold grid search cv to improve the model. I also checked the percentage score for the best split. Next, I focused to tune the model for max depth and length. For an example we can see in the below plot after 10-12 depth it starts to overfit. I can also checked the min sample leaf (50 as below).





Using the grid search with the optimal hyperparameter my **decision tree score is now 93.06%**
The maximum depth is 10, min sample leaf is 50, min sample split is 50 for the optimal hyperparameter.

```
[ ]  # printing the optimal accuracy score and hyperparameters
     print("best accuracy", grid_search.best_score_)
     print(grid_search.best_estimator_)

     best accuracy 0.930607094106063
     DecisionTreeClassifier(criterion='entropy', max_depth=10, min_samples_leaf=50,
                            min_samples_split=50)
```

Please refer to my code (attached in the appendix) for better visualization and clarity.

Model 3 : Stochastic Gradient Descent (SGD) Classifier

Stochastic Gradient Descent (SGD) is a simple yet efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost function. In other words, it is used for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic regression. It has been successfully applied to large-scale datasets because the update to the coefficients is performed for each training instance, rather than at the end of instances. Stochastic Gradient Descent (SGD) classifier basically implements a plain SGD learning routine supporting various loss functions and penalties for classification. Scikit-learn provides SGD Classifier module to implement SGD classification. I have used by default penalty or regularization which is L2 and also L1. Default loss function is 'hinge' but I have also used log. We can also use :

- modified_huber − a smooth loss that brings tolerance to outliers along with probability estimates. In my dataset we do not have much outliers and hence might not be relevant
- squared_hinge − similar to 'hinge' loss but it is quadratically penalized.
- perceptron − as the name suggests, it is a linear loss which is used by the perceptron algorithm.

Below is my best parameters :

```
grid_search.best_params_

{'alpha': 0.0001, 'loss': 'log', 'penalty': 'l1'}
```

Next, I predict using the best parameters and checked the model accuracy.

```
sgd = SGDClassifier(alpha = 0.0001, loss = 'log', penalty = 'l1')

sgd.fit(x_train, y_train)

# making predictions
y_pred = sgd.predict(x_test)

# Printing classifier report after prediction
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.85      0.93      0.89     17041
           1       0.62      0.41      0.49      4864

    accuracy                           0.81     21905
   macro avg       0.73      0.67      0.69     21905
weighted avg       0.80      0.81      0.80     21905
```

The **SGD classifier gives the accuracy score of 81%.**

9. Conclusion or Result:

Based on all the 3 models Decision Tree is giving the highest accuracy score with overall performance and hence I would recommend the Banks to use the Decision Tree model for best result and predict the default rate with 93.06% accuracy.

10. Future Recommendation:

I have worked on the 3 models as a part of my project however we can add random forest or any other model to check the results. I would also recommend including few external data like interest rates or inflation rates etc which might even improve our model.

11. Limitations:

- My dataset is dated till 2015 and I do not have updated datasets.
- There could be external factors like global pandemic which could impact the model and accuracy scores which is not in our hands sometimes.
- I also do not have individual credit scores or any other loan active datasets which might also impact the overall loan default rate

12. References:

[1] Article title: "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines , Website title: Taylor & Francis
URL:https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342

[2] Grimshaw, A. and Edmister, R., 1982. SBA default rates. Journal of Economics and Business, 34(4), pp.343-348.

[3] Weinberg, by J. (no date) The great recession and its aftermath, Federalreservehistory.org. Available at: https://www.federalreservehistory.org/essays/great-recession-and-its-aftermath

[4] (No date) Researchgate.net. Available at: https://www.researchgate.net/publication/5168423_Measuring_the_Default_Risk_of_Small_Business_Loans_A_Survival_Analysis_Approach

[5] Stangler, D. (2022) Questions about the post-pandemic small business financing landscape, Forbes. Available at: https://www.forbes.com/sites/danestangler/2022/03/17/questions-about-the-post-pandemic-small-business-financing-landscape/?sh=1a889742514d

[6] Constantino, A. K. (2022) These small businesses have survived the pandemic despite being rejected for PPP loans. Here's how they did it, CNBC. Available at: https://www.cnbc.com/2022/01/29/ppp-loans-how-some-small-businesses-have-survived-the-covid-pandemic-without-them.html

[7] How Feedzai, AWS reduced fraud for a U.K. bank (2021) Feedzai. Available at: https://feedzai.com/resource/major-uk-based-bank-leverages-feedzai-and-aws-to-reduce-fraud/

[8] Machine learning in finance: 10 companies to know (no date) Built In. Available at: https://builtin.com/artificial-intelligence/machine-learning-finance-examples

[9] Vardhan, V. (2018) Case study — loan prediction - Vishnu vardhan, Medium. Available at: https://medium.com/@vishnumbaprof/case-study-loan-prediction-ac035f3ec9e4

[10] Coiman, A. (2021) "Binary classification Machine Learning. Case study loan prediction," Abraham Coiman, 6 February. Available at: https://acoiman.github.io/post/loan_prediction/

[11] Lemus, G. (2018) AI in finance: Cutting through the hype (with case studies), DataDrivenInvestor. Available at: https://medium.datadriveninvestor.com/ai-in-finance-cutting-through-the-hype-with-case-studies-f361518b00d4

13. Appendix:

Data source link

https://amstat.tandfonline.com/doi/figure/10.1080/10691898.2018.1434342?scroll=top&needAccess=true
Final Code\SBAnational.csv

Python code :

Final Code\Data_Mining_Project_Sayantani.ipynb