Institute of Technology of Cambodia
Department of Applied Mathematics and Statistics

Logistic Regression Project Idea for Loan Default Prediction

INSTRUCTORS

MR. PHUAK SOKKHEY (COURSE)

MR. NHIM MALAI (TD)

Date: June 16, 2023

Prepared by: SENG Lay, VANN Visal, HOK Kimleang, BUT Cheableng, MEACH Seaklav, LIM Sunheng

LET'S GET STARTED!

GROUP MEMBERS

NAME	ID
SENG LAY	e20200872
VANN VISAL	e20200537
HOK KIMLEANG	e20200637
BUT CHEABLENG	e20200861
MEACH SEAKLAV	e20200683
LIM SUNHENG	e20200807

TABLE OF CONTENTS

1. Introduction

Sanctioning a loan is an essential decision for any lending institution. A bank loses out on potential income by rejecting a loan to an individual or a company. At the same time, granting loans where lending risks exceed the returns could result in heavy losses. This is why banks stand much to gain from relying on good loan default prediction models based on actual statistics.

2. Background and Data Description

The Original Source

- The original data set is from the *U.S.SBA* loan database, which includes historical data from 1987 through 2014 (899,164 observations) with 27 variables.
- The original data set includes information on whether the loan was paid off in full or if the SMA has to charge off any amount and how much that amount was.

THE USED DATA SET IN OUR PROJECT

- The dataset of this project is "***SBAcase.11.13.17.csv***".
- The dataset used is a subset of the original set. It contains loans about Real Estate and Rental and Leasing industry in California.
- In this subset file has 2102 observations and 35 variables.

Data Description

Variable	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

Variable	Data Type	Description of variable
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	Franchised code, (00000 or 00001) = No franchise

Variable	Data Type	Description of variable
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

Variable	Data Type	Description of variable
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approved Loan
New	Number	=1 if NewExist=2 (New Business), =0 if NewExist=1 (Existing Business)
Portion	Number	Propotion of gross amount guaranteed by SBA
RealEstate	Number	=1 if loan is backed by real estate, =0 otherwise

Variable	Data Type	Description of variable
Recession	Number	=1 if loan is active during Great Recession, =0 if NewExist=1 (Existing Business)
Selected	Number	=1 if the data are selected as training data to build model for assignment =0 if the data are selected as testing data to validate model
Default	Number	=1 if MIS_Status=CHGOFF, =0 if MIS_Status=PIF
daysterm	Number	Extra variable generated when creating "Recession"
xx	Number	Extra variable generated when creating "Recession"

3. Project Scope
3.1. Learning Objective

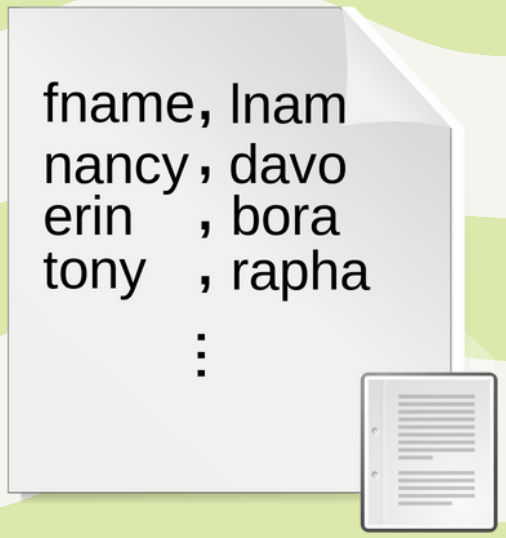
STEP 1: Identify the input and output variables.
STEP 2: Understanding the case study and dataset.
STEP 3: Building the model (Logistic Regression).
STEP 4: Make a decision base on the model.

3.2. Statistical Software

Programming Language

PYTHON

Imported File



Comma-Separated Values (CSV)

4. Exploratory Data Analysis (EDA)

4.1. Data Exploration

In [1]:

```python
# Import packages used for analysis
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.feature_selection import SelectKBest
from sklearn.linear_model import SGDRegressor
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
```

Import packages used for analysis

In [2]:

```python
# Load the SBA loan data and make a copy for exploration
df = pd.read_csv('SBAnational.csv')

df_copy = df.copy()
```

```
/opt/anaconda3/lib/python3.9/site-packages/IPython/core/interactiveshell.py:3444: DtypeWarning: Columns (9) have mixe
d types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

In [3]:

```python
df_copy.head()
```

Out[3]:

	LoanNr_ChkDgt	Name	City	State	Zip	Bank	BankState	NAICS	ApprovalDate	ApprovalFY	...	RevLineCr	LowDoc	ChgOffDate
0	1000014003	ABC HOBBYCRAFT	EVANSVILLE	IN	47711	FIFTH THIRD BANK	OH	451120	28-Feb-97	1997	...	N	Y	NaN
1	1000024006	LANDMARK BAR & GRILLE (THE)	NEW PARIS	IN	46526	1ST SOURCE BANK	IN	722410	28-Feb-97	1997	...	N	Y	NaN
2	1000034009	WHITLOCK DDS, TODD M.	BLOOMINGTON	IN	47401	GRANT COUNTY STATE BANK	IN	621210	28-Feb-97	1997	...	N	N	NaN
3	1000044001	BIG BUCKS PAWN & JEWELRY, LLC	BROKEN ARROW	OK	74012	1ST NATL BK & TR CO OF BROKEN	OK	0	28-Feb-97	1997	...	N	Y	NaN
4	1000054004	ANASTASIA CONFECTIONS, INC.	ORLANDO	FL	32801	FLORIDA BUS. DEVEL CORP	FL	0	28-Feb-97	1997	...	N	N	NaN

5 rows × 27 columns

Load and Copy data for Exploration

4.1. Data Exploration

```
In [4]:  1  df_copy.shape
Out[4]: (899164, 27)
```

Shape of dataset

```
In [5]:  1  df_copy.isnull().sum()
Out[5]: LoanNr_ChkDgt              0
        Name                      14
        City                      30
        State                     14
        Zip                        0
        Bank                    1559
        BankState               1566
        NAICS                      0
        ApprovalDate               0
        ApprovalFY                 0
        Term                       0
        NoEmp                      0
        NewExist                 136
        CreateJob                  0
        RetainedJob                0
        FranchiseCode              0
        UrbanRural                 0
        RevLineCr               4528
        LowDoc                  2582
        ChgOffDate            736465
        DisbursementDate        2368
        DisbursementGross          0
        BalanceGross               0
        MIS_Status              1997
        ChgOffPrinGr               0
        GrAppv                     0
        SBA_Appv                   0
        dtype: int64
```

Checking missing value in the dataset.

4.1. Data Exploration

```
In [6]:
# Drop null values from specified columns
df_copy.dropna(subset=['Name', 'City', 'State', 'BankState', \
                       'NewExist','RevLineCr', 'LowDoc', \
                       'DisbursementDate', 'MIS_Status'],
               inplace=True)
df_copy.isnull().sum()

Out[6]:
LoanNr_ChkDgt             0
Name                      0
City                      0
State                     0
Zip                       0
Bank                      0
BankState                 0
NAICS                     0
ApprovalDate              0
ApprovalFY                0
Term                      0
NoEmp                     0
NewExist                  0
CreateJob                 0
RetainedJob               0
FranchiseCode             0
UrbanRural                0
RevLineCr                 0
LowDoc                    0
ChgOffDate           725369
DisbursementDate          0
DisbursementGross         0
BalanceGross              0
MIS_Status                0
ChgOffPrinGr              0
GrAppv                    0
SBA_Appv                  0
dtype: int64
```

Drop null values from specific columns

```
In [7]:
# Check data types of each feature
df_copy.dtypes

Out[7]:
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
```

Checking data types of each features

4.2. Data Analysis

In [36]:
```
df_copy.describe(include=['object', 'float', 'int'])
```

Out[36]:

	State	BankState	ApprovalFY	Term	NoEmp	CreateJob	RetainedJob	UrbanRural	RevLineCr
count	438504	438504	438504.000000	438504.000000	438504.000000	438504.000000	438504.000000	438504	438504.000000
unique	51	53	NaN	NaN	NaN	NaN	NaN	3	NaN
top	CA	NC	NaN	NaN	NaN	NaN	NaN	1	NaN
freq	59171	55644	NaN	NaN	NaN	NaN	NaN	270482	NaN
mean	NaN	NaN	2002.665604	94.119445	9.794887	1.843611	4.568973	NaN	0.418959
std	NaN	NaN	5.492623	68.548785	57.674184	16.496650	15.330176	NaN	0.493389
min	NaN	NaN	1984.000000	0.000000	0.000000	0.000000	0.000000	NaN	0.000000
25%	NaN	NaN	1999.000000	58.000000	2.000000	0.000000	0.000000	NaN	0.000000
50%	NaN	NaN	2005.000000	84.000000	4.000000	0.000000	1.000000	NaN	0.000000
75%	NaN	NaN	2007.000000	90.000000	9.000000	1.000000	4.000000	NaN	1.000000
max	NaN	NaN	2011.000000	527.000000	9999.000000	5621.000000	4441.000000	NaN	1.000000

11 rows × 23 columns

After cleaning data, now take a look at the summary statistics table

In [36]:
```
df_copy.describe(include=['object', 'float', 'int'])
```


Out[36]:

LowDoc	...	IsFranchise	NewBusiness	Default	DaysToDisbursement	DisbursementFY	StateSame	SBA_AppvPct
438504.000000	...	438504.000000	438504.000000	438504.000000	438504.000000	438504.000000	438504.000000	438504.000000
NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.057247	...	0.030597	0.263840	0.221918	109.090631	2002.705704	0.454094	0.654071
0.232314	...	0.172224	0.440714	0.415537	182.221498	5.403909	0.497889	0.179932
0.000000	...	0.000000	0.000000	0.000000	-3614.000000	1984.000000	0.000000	0.050000
0.000000	...	0.000000	0.000000	0.000000	27.000000	2000.000000	0.000000	0.500000
0.000000	...	0.000000	0.000000	0.000000	51.000000	2005.000000	0.000000	0.500000
0.000000	...	0.000000	1.000000	0.000000	109.000000	2007.000000	1.000000	0.829994
1.000000	...	1.000000	1.000000	1.000000	4029.000000	2010.000000	1.000000	1.000000

Other columns

4.2. Data Analysis

Out[44]:

Industry	0	1	Def_Percent
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

Check default percentage by Industry

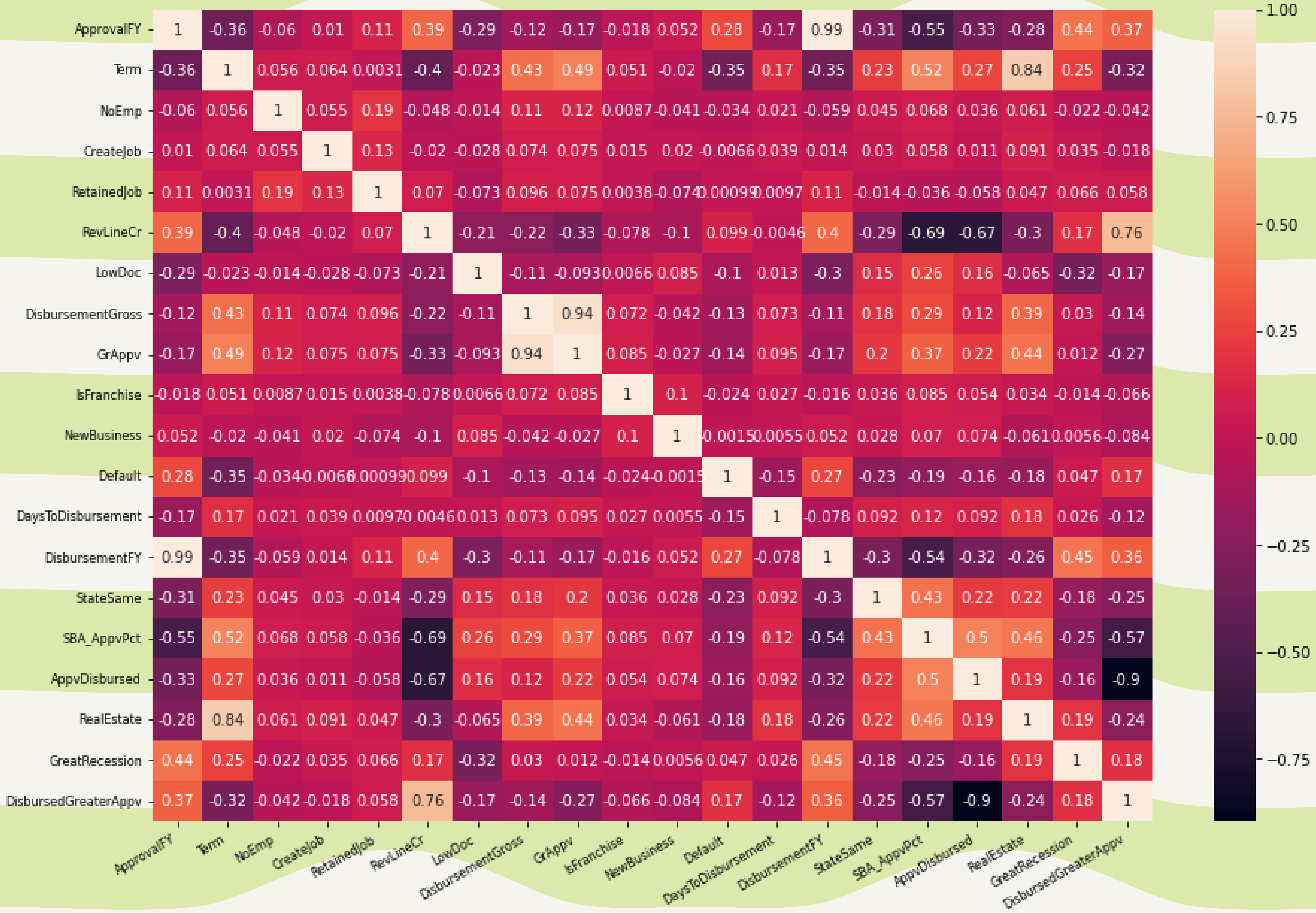
Out[45]:

State	0	1	Def_Percent
AK	979	94	0.087605
AL	3192	805	0.201401
AR	2414	528	0.179470
AZ	5119	2473	0.325738
CA	42983	16138	0.272966
CO	7439	2349	0.239988
CT	5328	1064	0.166458
DC	567	157	0.216851
DE	841	246	0.226311
FL	14820	7587	0.338600
GA	7080	3141	0.307308
HI	1164	263	0.184303
IA	4596	568	0.109992
ID	4046	886	0.179643
IL	11500	4505	0.281475
IN	5904	1524	0.205170
KS	4269	631	0.128776
KY	2959	789	0.210512
LA	3228	775	0.193605
MA	11812	2289	0.162329
MD	5084	1431	0.219647
ME	2502	317	0.112451
MI	7976	3287	0.291841
MN	9186	1688	0.155233
MO	7679	1636	0.175631
MT	3533	264	0.069529
NC	4922	1423	0.224271
ND	2319	174	0.069795
NE	2329	285	0.109028
NH	5966	922	0.133856
NJ	8217	3241	0.282859
NM	2408	321	0.117626
NV	2688	1239	0.315508
NY	24822	8237	0.249161
OH	14150	3592	0.202457
OK	3839	765	0.166160
OR	4519	1137	0.201025
PA	14959	3146	0.173764
RI	4227	730	0.147266
SC	1925	616	0.242424
SD	1759	132	0.069804
TN	3477	1003	0.223884
TX	22738	6203	0.214333
UT	8565	2607	0.233351
VA	4862	1371	0.219958
VT	2222	199	0.082197
WA	9015	2074	0.187032
WI	8591	1463	0.145514
WV	1188	212	0.151429
WY	1156	69	0.056327

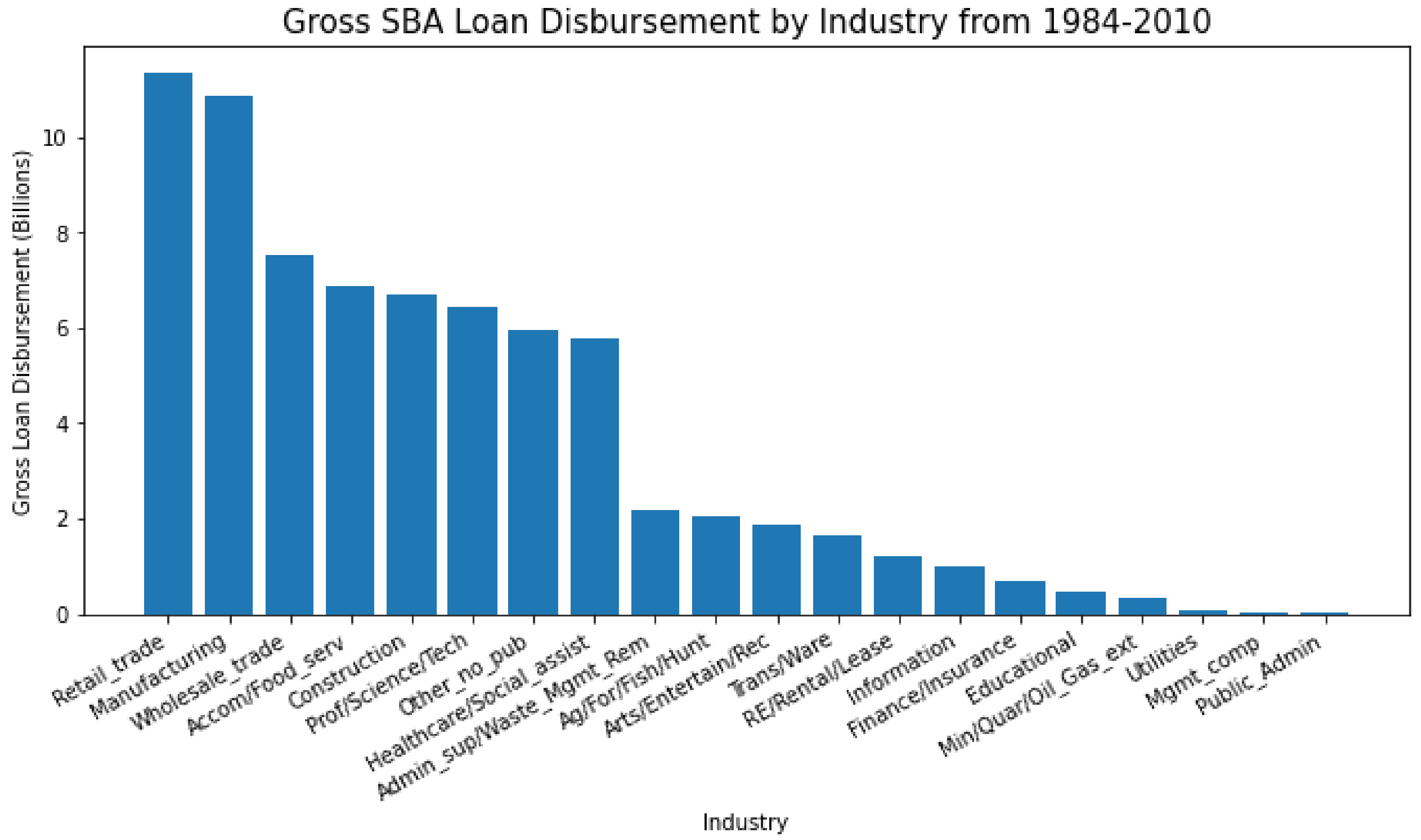
Check default percentage by States

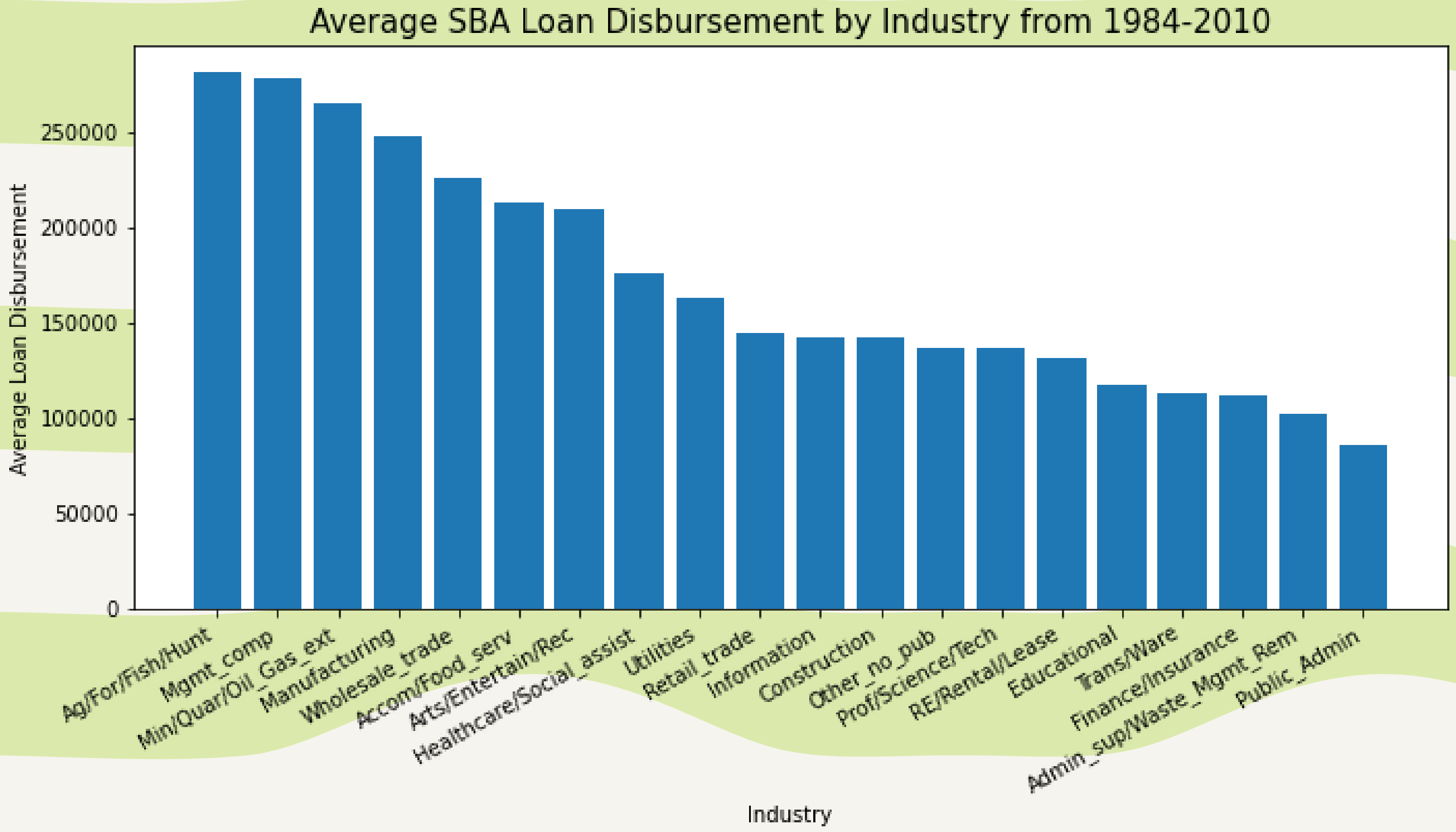
4.3. Data Visualization

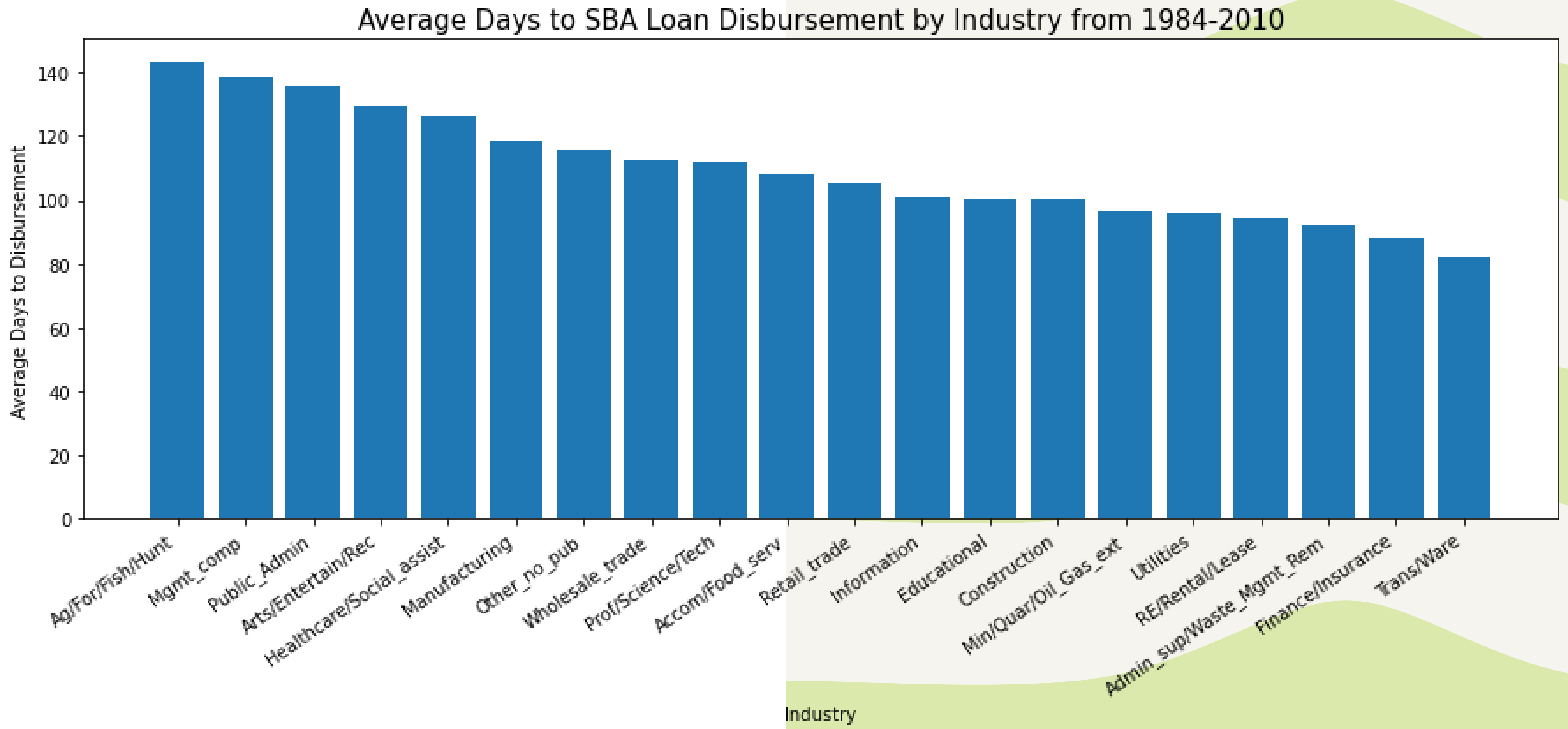
- GrAppv & DisbursementGross, Positive - Makes sense that in most situations, the amount disbursed is close to what was approved.
- DisbursedGreaterAppv & AppvDisbursed, Negative - Also makes sense since when the disbursed amount is greater than approved, the disbursed amount is then not equal to the approved amount.
- RevLineCr & DisbursedGreaterAppv, Positive - Due to the nature of revolving lines of credit (think of it like a credit card for businesses where the business can draw funds with a limit, pays it off when able, and then draw more funds again), this makes sense that over time more funds are used then the limit set for the loan.
- DisbursementFY & ApprovalFY, Positive - More often than not, the funds will be disbursed in the same year they are approved.
- AppvDisbursed & RevLineCr, Negative - Typically, based on my experience underwriting loans as a Credit Analyst, the limit for a line of credit is lower than a term loan on average since the business can continually draw funds from the line of credit when needed after paying off the balance, which would explain the negative relationship.
- SBA_AppvPct & RevLineCr, Negative - SBA lines of credit can still be eligible for guarantees, however the guarantee percentage is dependant on the size of the loan. Although this doesn't quite explain the negative relationship between SBA guarantee percentage and a loan being RevLineCr, what could is the type of SBA loan program used for the loan application.



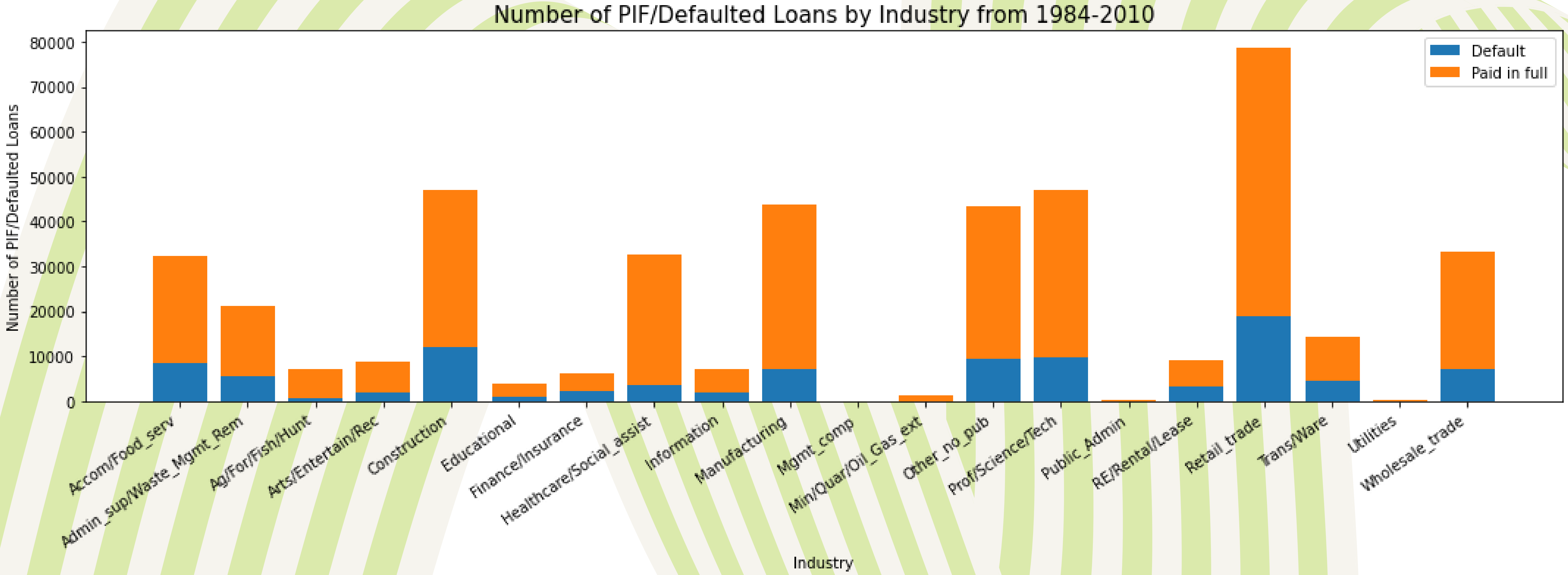
4.3. Data Visualization

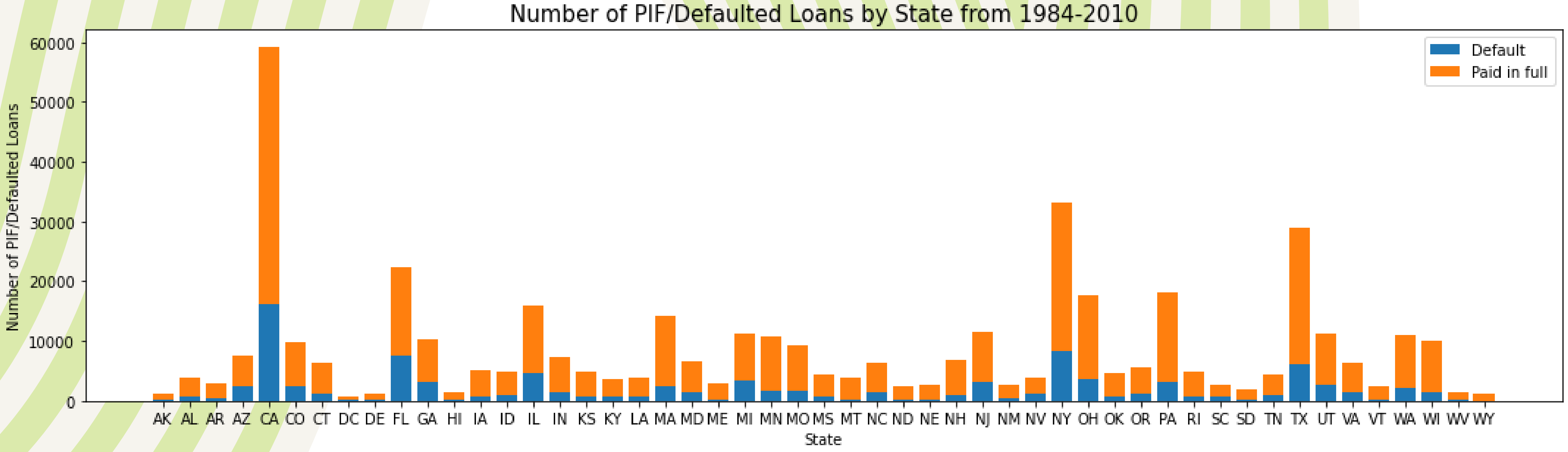




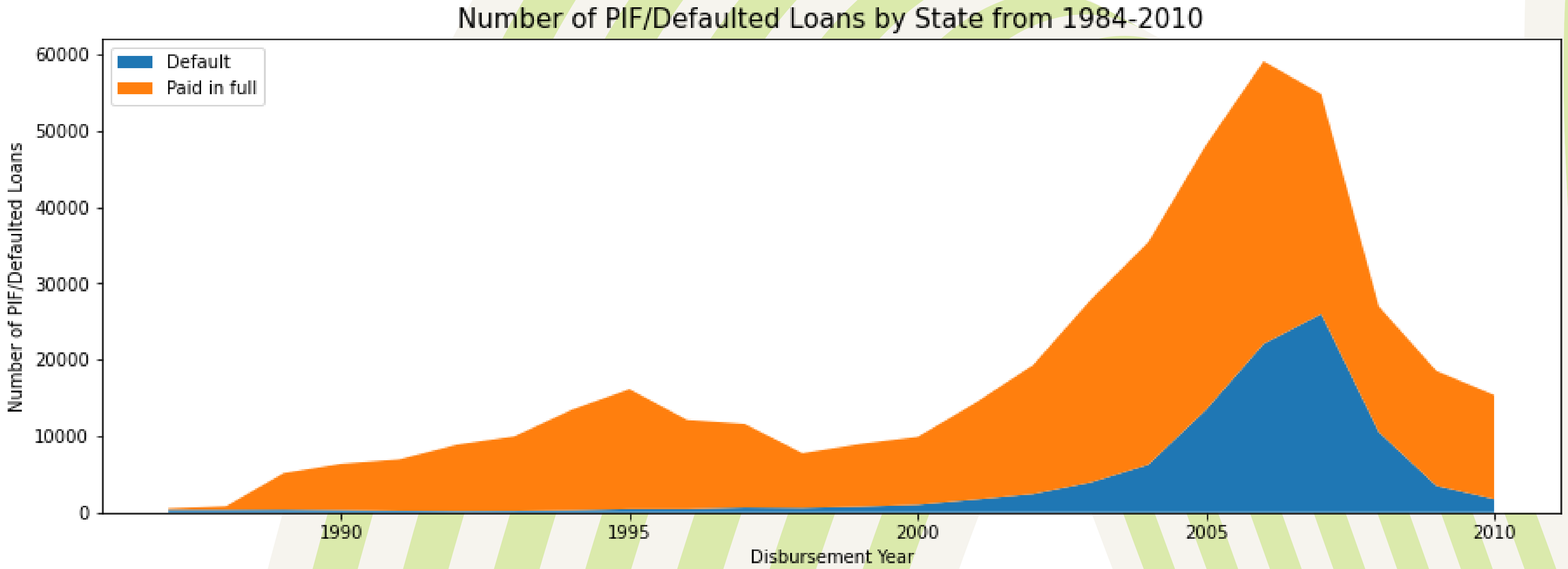


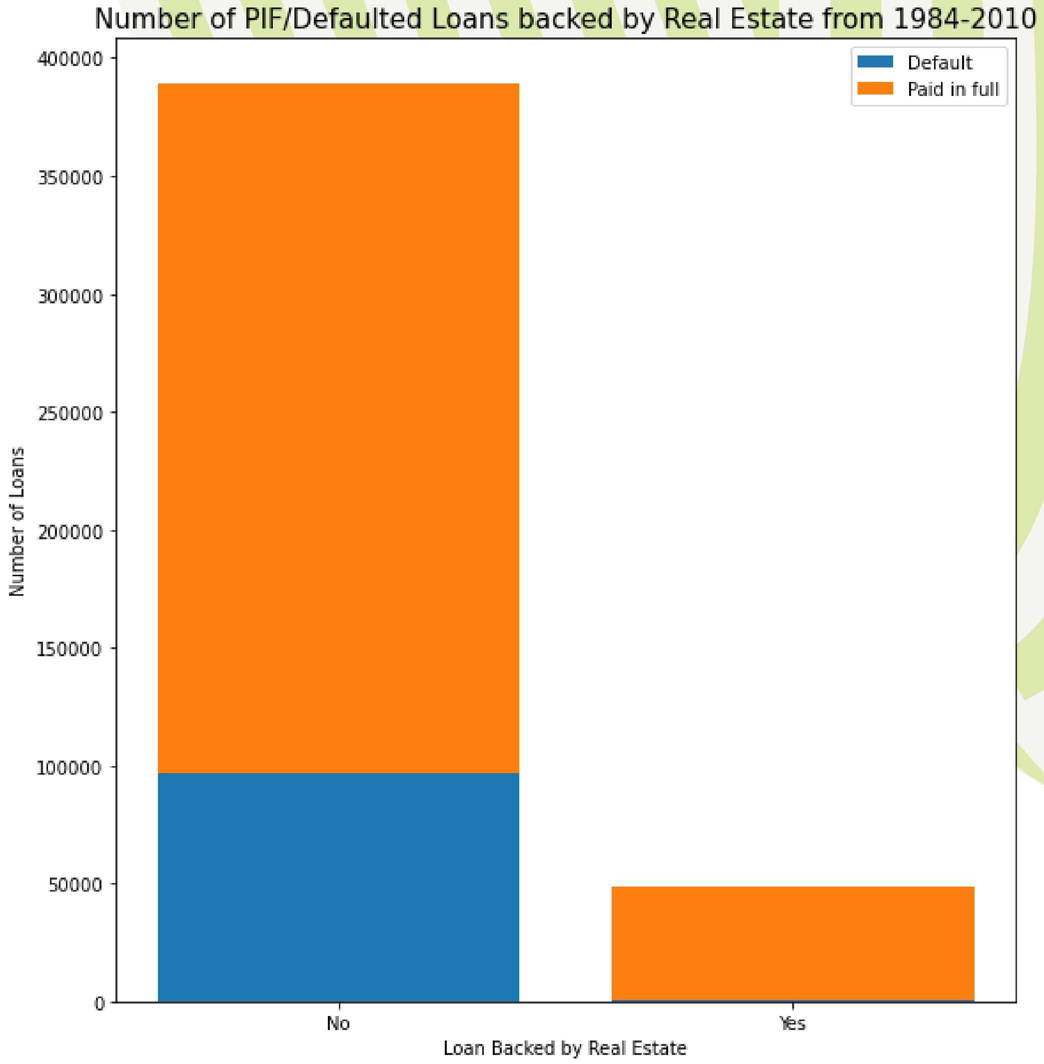
4.3. Data Visualization

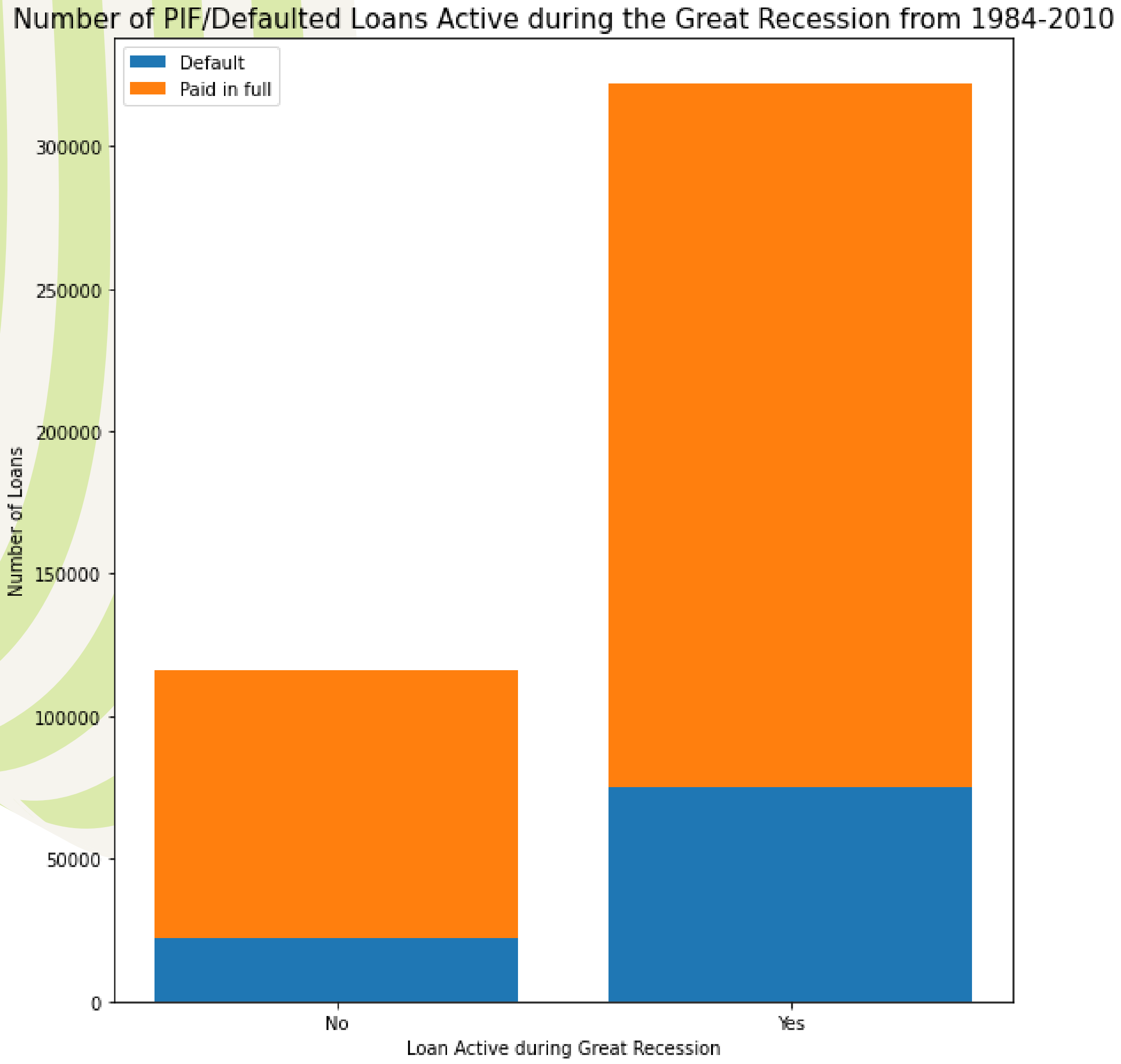




4.3. Data Visualization







WE'RE ABOUT TO COME TO AN END!

5. Statistical Model

IN OUR PROJECT, LOGISTIC REGRESSION WILL BE USED!

5.1. Logistic Regression (Code and Result)

In [54]:

```
# Initialize model
log_reg = LogisticRegression(random_state=2)

# Train the model and make predictions
log_reg.fit(X_train, y_train)
y_logpred = log_reg.predict(X_val)

# Print the results
print(classification_report(y_val, y_logpred, digits=3))
```

```
              precision    recall  f1-score   support

           0      0.895     0.952     0.923     85147
           1      0.785     0.610     0.686     24376

    accuracy                          0.876    109523
   macro avg      0.840     0.781     0.804    109523
weighted avg      0.870     0.876     0.870    109523
```

5.2. Interpret the result of Logistic Regression

We can see here that with the Logistic Regression model, we have a decent accuracy at 87.6%, however the F1-score of 68.6% for defaulted loans does not seem very promising. The precision suggests that the model is correct 78.5% of the time when the loan defaults, and the recall suggests that the model identifies 61% of defaulted loans correctly. That means that 39% of loans that defaulted were incorrectly classified as loans that would be paid in full, which is NOT very good.

6. Conclusion

This analysis found that the length of the loan term is the most important factor in determining whether or not a loan goes into default. Further analysis could be done to consider other factors, such as which SBA loan programs each loan fell under. Additionally, the data does not capture the cash flow of each business, working capital, the existing debt they had prior to applying for the SBA loan, and the personality, attitude, and drive of a business owner. It is important to note that if the owner doesn't want to pay the loan, they won't.

7. References

- ***SBA Loan Kaggle:*** https://www.kaggle.com/datasets/larsen0966/sba-loans-case-data-set
- ***SBA Loan Kaggle:*** https://www.kaggle.com/code/kevinm6720/sba-loan-approval-analysis
- ***Full Article Guideline of "Should this loan be Approved or Denied?":*** https://amstat.tandfonline.com/doi/full/10.1080/10691898.2018.1434342#.ZBWIk-xBxQI

Thank you for your attention! Do you have any questions?

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS