DEFAULT OR PAID IN FULL? A LOAN APPROVAL RECOMMENDATION

Muhammad Rijal Senjaya

FINAL PROJECT DIBIMBING BATCH 17

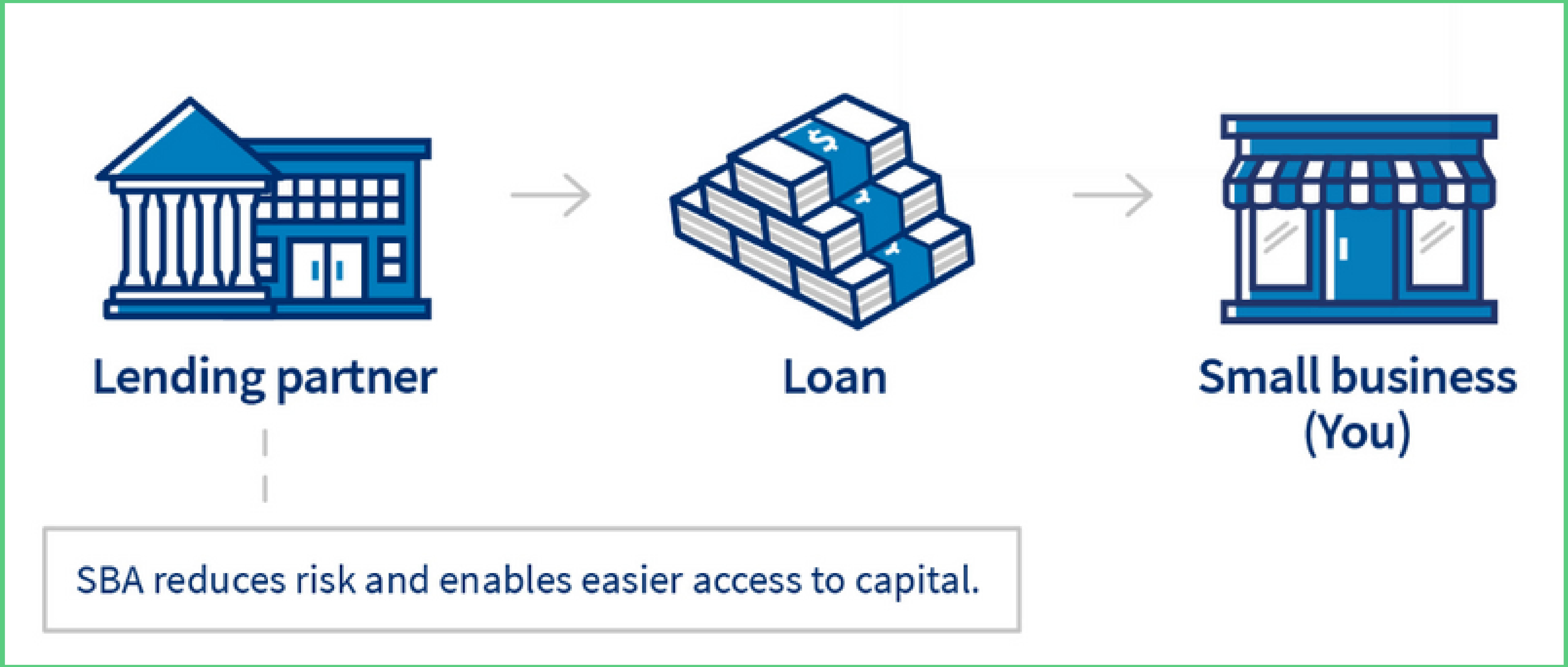
dibimbing

TABLE OF CONTENTS

PRESENTATION HIGHLIGHTS

ibimbing

BACKGROUND & OBJECTIVE

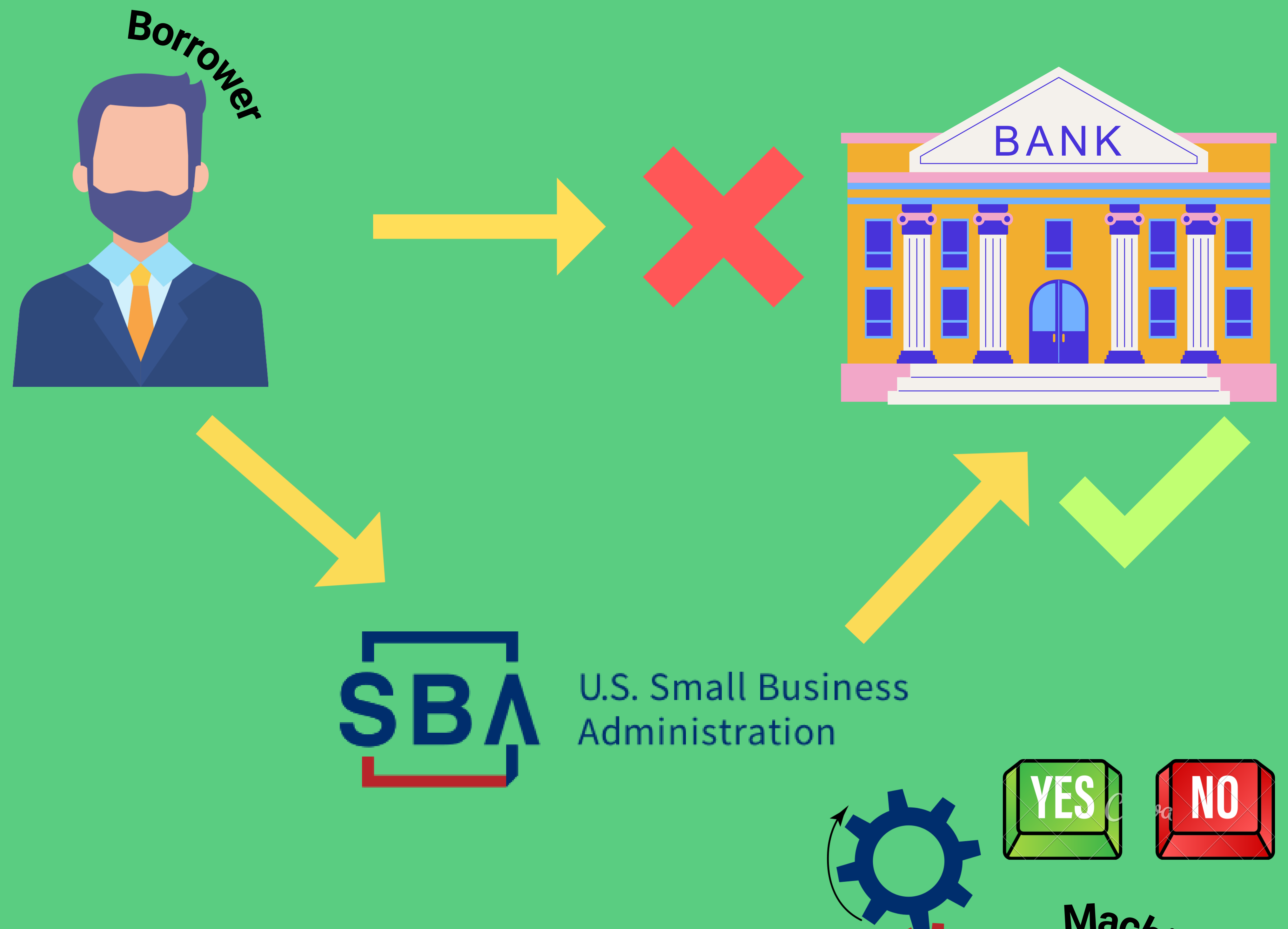


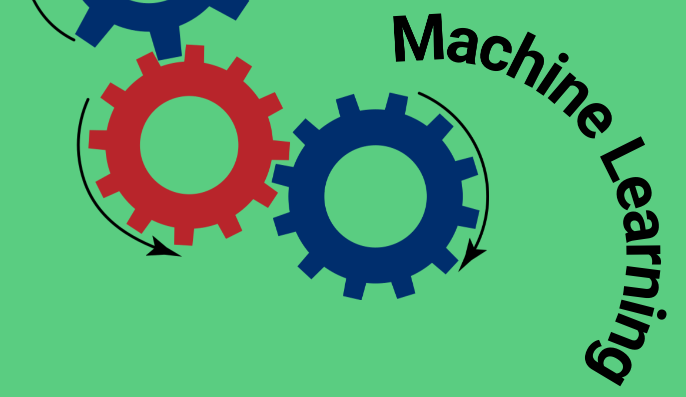
https://www.sba.gov/funding-programs/loans

The U.S. Small Business Administration helps small businesses get funding by setting guidelines for loans and reducing lender risk. These SBA-backed loans make it easier for small businesses to get the funding they need. This increases the risk to the SBA however, which can sometimes make it difficult to get accepted for one of their loan programs.

ibimbing

BACKGROUND & OBJECTIVE

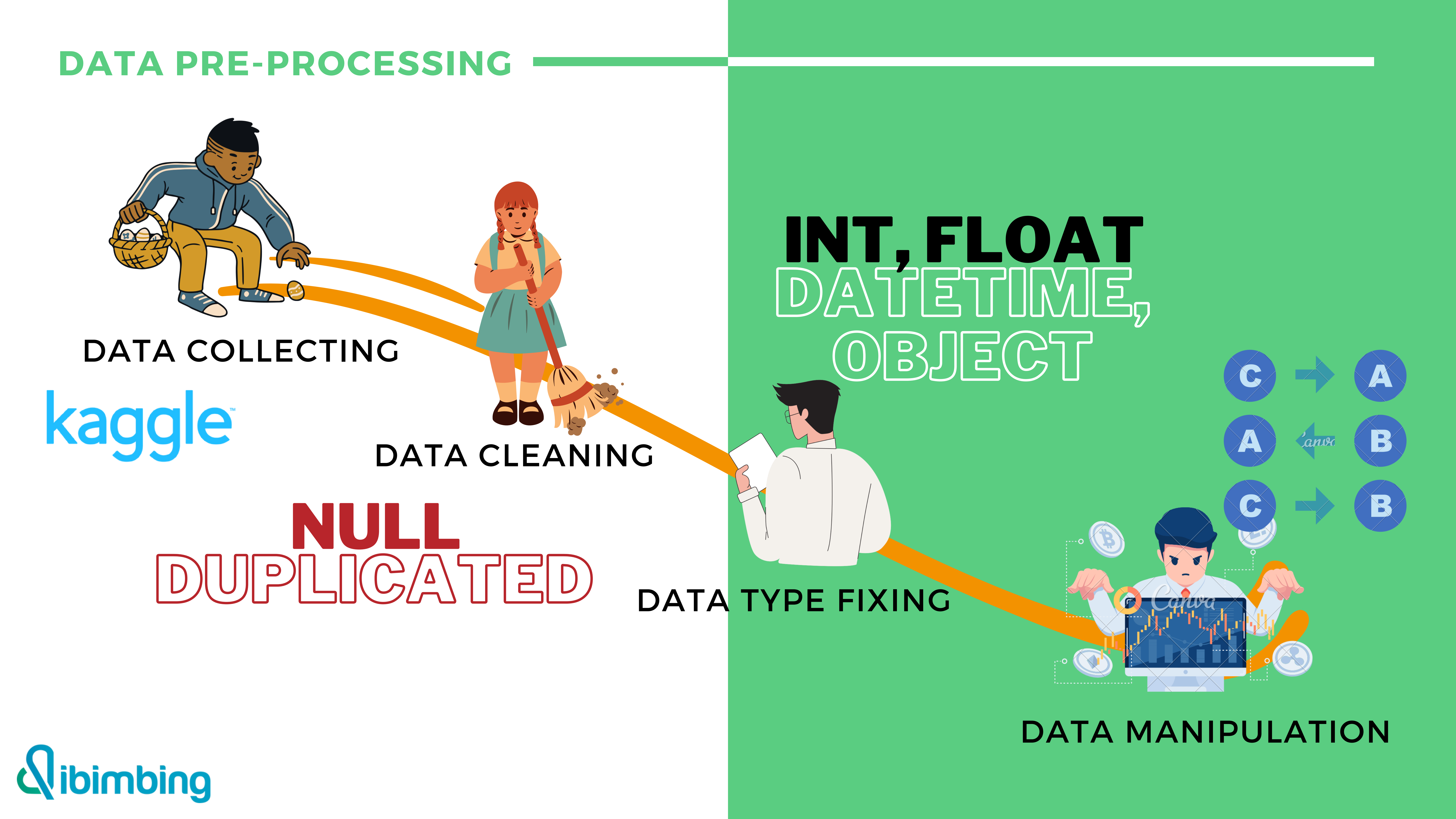




Making a machine learning predictor to help asses whether the borrower have a potential to default or paid in full successfully

A LOAN APPROVAL PREDICTOR: WILL DEFAULT OR PAID IN FULL?

ibimbing


DATA PRE-PROCESSING
DATA COLLECTING
kaggle
DATA CLEANING
NULL
DUPLICATED
INT, FLOAT
DATETIME,
OBJECT
DATA TYPE FIXING
C → A
A ← B
C → B
DATA MANIPULATION
ibimbing

DATA UNDERSTANDING

The dataset is a real dataset from US Small Business Administration SBA. US Small Business Administration is a US Goverment Agency, which has purpose to promote the economy by assisting country small businesses.

SBA U.S. Small Business Administration

27 columns
899.164 rows

MIS_Status
Charge Off (1)
Paid in Full (0)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 899164 entries, 0 to 899163
Data columns (total 27 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   LoanNr_ChkDgt      899164 non-null  int64
 1   Name               899150 non-null  object
 2   City               899134 non-null  object
 3   State              899150 non-null  object
 4   Zip                899164 non-null  int64
 5   Bank               897605 non-null  object
 6   BankState          897598 non-null  object
 7   NAICS              899164 non-null  int64
 8   ApprovalDate       899164 non-null  object
 9   ApprovalFY         899164 non-null  object
 10  Term               899164 non-null  int64
 11  NoEmp              899164 non-null  int64
 12  NewExist           899028 non-null  float64
 13  CreateJob          899164 non-null  int64
 14  RetainedJob        899164 non-null  int64
 15  FranchiseCode      899164 non-null  int64
 16  UrbanRural         899164 non-null  int64
 17  RevLineCr          894636 non-null  object
 18  LowDoc             896582 non-null  object
 19  ChgOffDate         162699 non-null  object
 20  DisbursementDate   896796 non-null  object
 21  DisbursementGross  899164 non-null  object
 22  BalanceGross       899164 non-null  object
 23  MIS_Status         897167 non-null  object
 24  ChgOffPrinGr       899164 non-null  object
 25  GrAppv             899164 non-null  object
 26  SBA_Appv           899164 non-null  object
dtypes: float64(1), int64(9), object(17)
memory usage: 185.2+ MB
```

ibimbing

DATA CLEANING

```
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
```

<0%

```
#Checking for duplicate
df.duplicated().sum()
```

```
0
```

DATASET DOESN'T CONTAIN DUPLICATED VALUE

<1%

EACH COLUMN CONTAIN LESS THAN ONE PERCENT OF MISSING VALUES

82%

'CHGOFFDATE' COLUMN CONTAIN HUGE NULL VALUES, HARD TO IMPUTE, UNUSEFULL

DATA TYPE FIXING

OBJECT TO INTEGER

Change the data type of **'DisbursementGross'**, **'BalanceGross'**, **'ChgOffPrinGr'**, **'GrAppv'**, **'SBA_Appv'** which should be integer instead of object. This is because the data come with $.

OBJECT TO DATETIME

Convert **'ApprovalDate'** and **'DisbursementDate'** columns to datetime values.

OBJECT TO INTEGER

Change data type of **'ApprovalFY'** which should be an integer but is coming up as an object type because there's a mixture of integers and strings here, with one record including an 'A' as well

ibimbing

DATA MANIPULATION

CHANGE CODE TO TEXT

Convert the data of **'NAICS'** from code number to the name of industry

```
'11': 'Ag/For/Fish/Hunt',
'21': 'Min/Quar/Oil_Gas_ext',
'22': 'Utilities',
'23': 'Construction',
'31': 'Manufacturing',
```

ENCODE WITH BOOLEAN

We use boolean to encode whether the business is franchise or not, New Business or not, joining another loan program or not, location similarity between bank and domicile, and the target column of course (**'MIS_Status'**),

MAKE A NEW COLUMN BY FORMULA

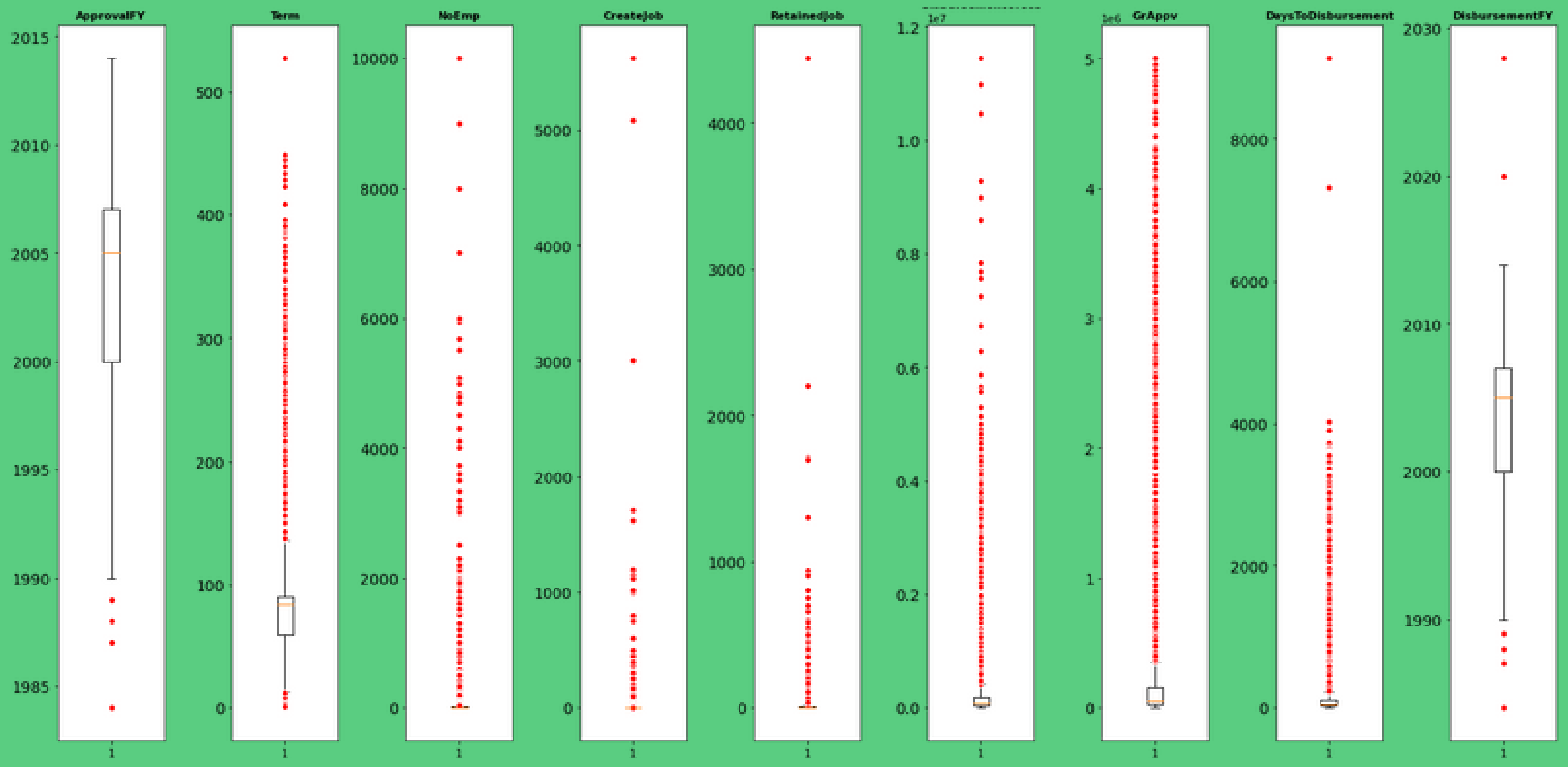
Create **'DaysToDisbursement'** column which calculates the number of days passed between **'DisbursementDate'** and **'ApprovalDate'**.

STATISTICAL ANALYSIS

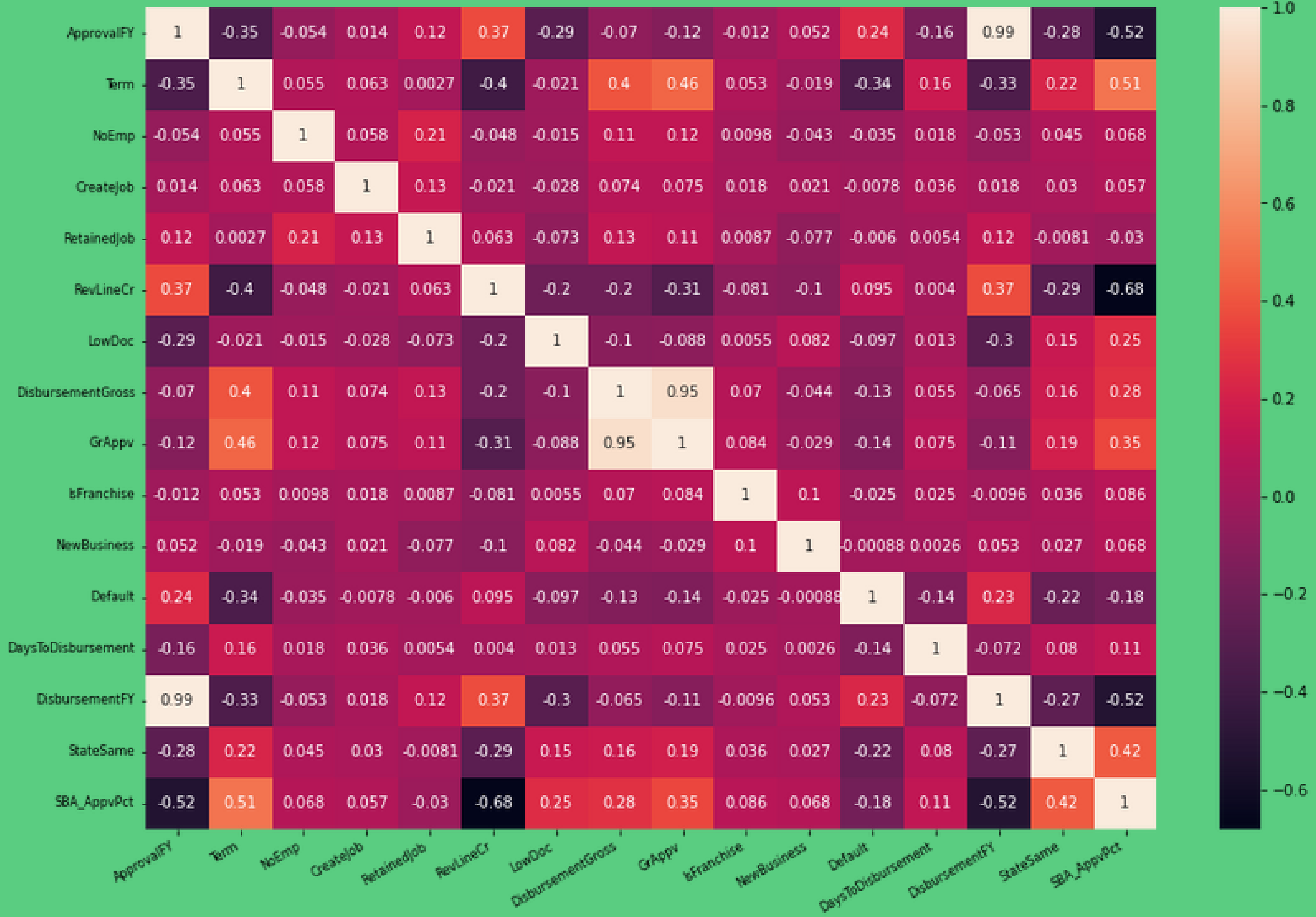
- The average loan term is 94 months with a standard deviation of 69 months, suggesting the loan terms are pretty spread out; Max loan term of 527 months could suggest some outliers in the data
- The average number of employees is about 9.8 with 75% of of businesses having 9 or less employees, suggesting '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 again
- 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%

UNIVARIATE ANALYSIS

There are so many outliers in all numerical features except **'ApprovalFY'** dan **'DisbursementFY'** that just a little bit.



MULTIVARIATE ANALYSIS

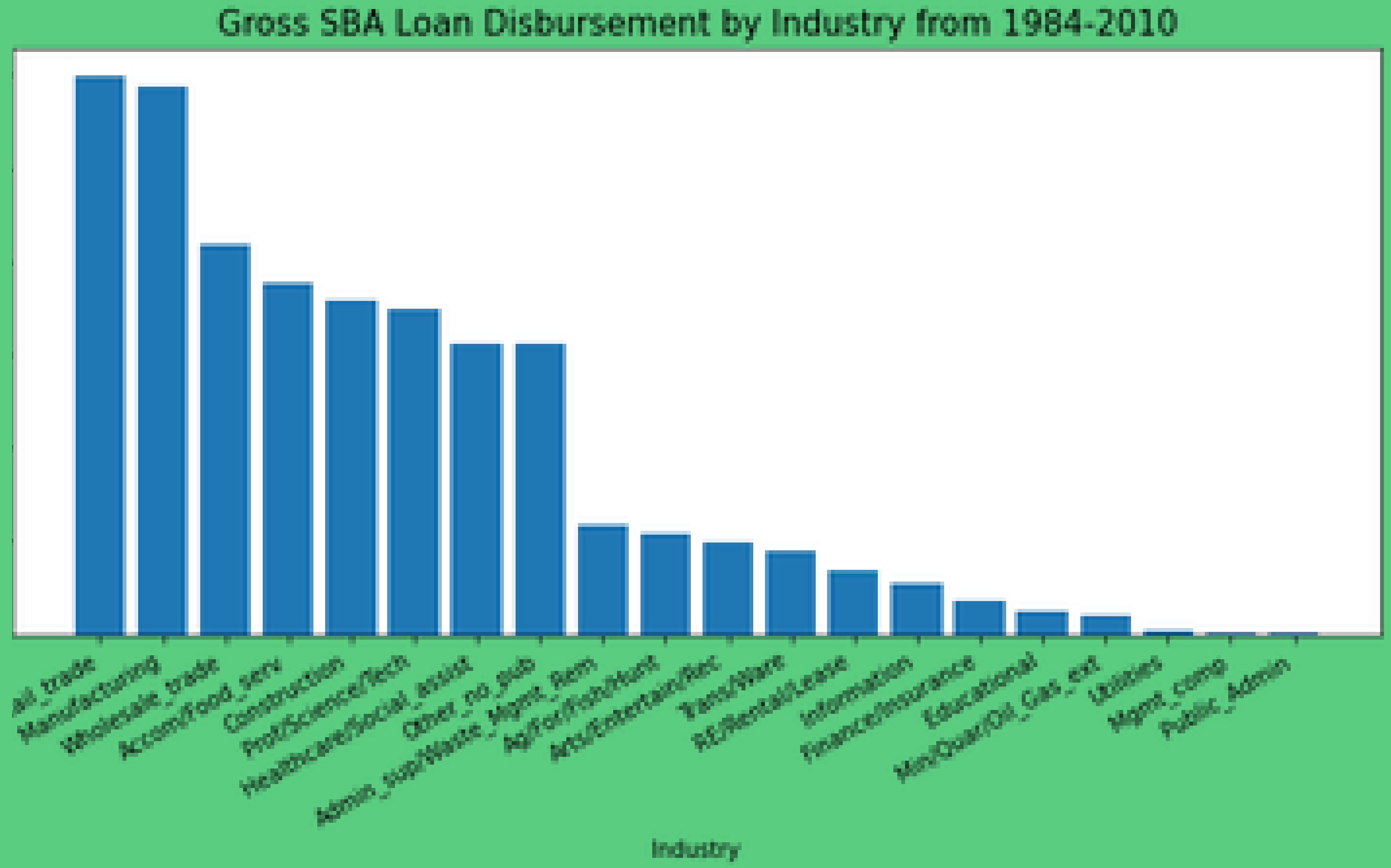


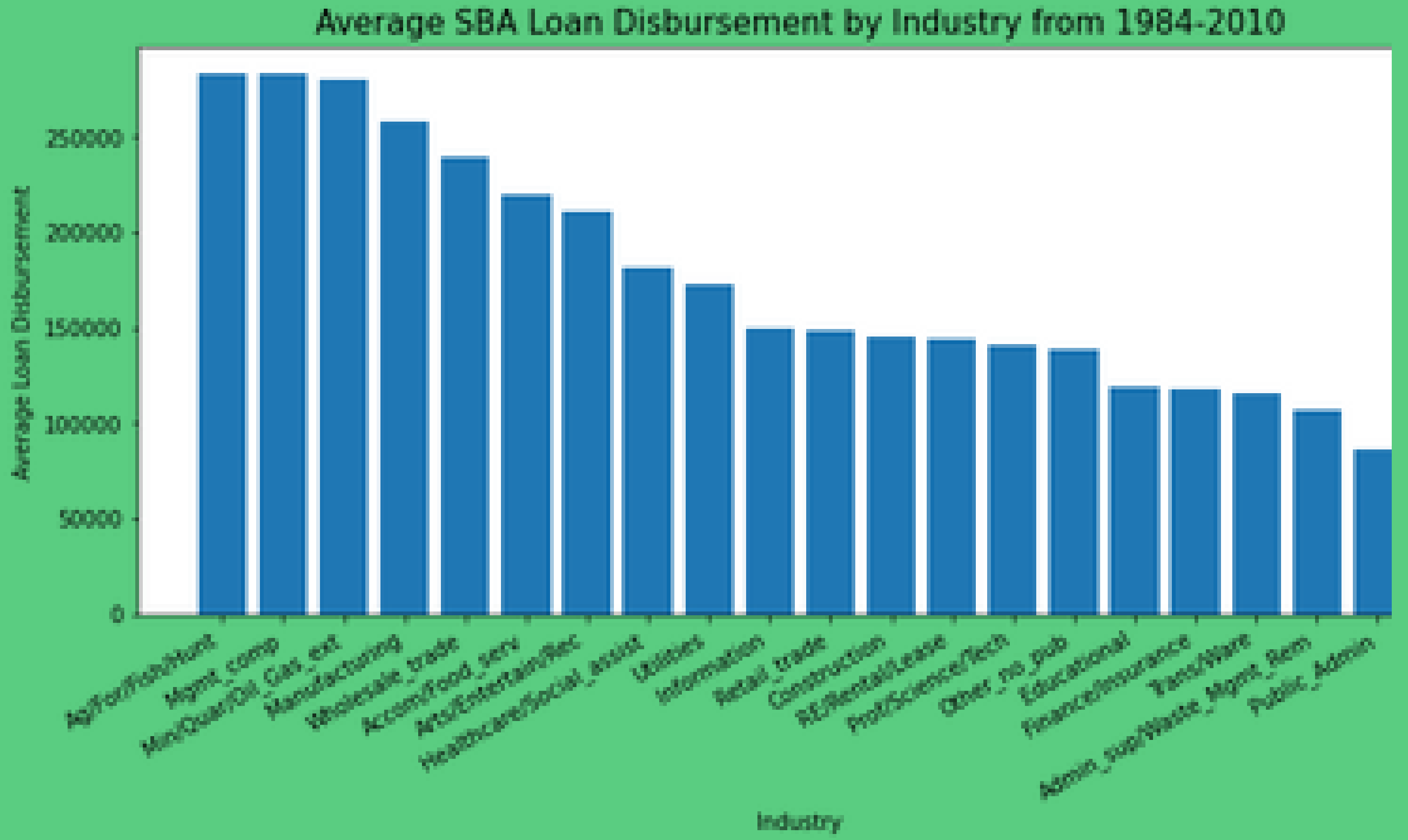
There are columns that have strong correlation such as **'GrAppv' - 'DisbursementGross'** and **'ApprovalFY' - 'DisbursementFY'**. It's mean that the year of approval and disbursement is mostly same, as well as the gross of approval and disbursement.

ibimbing

EXPLORATORY DATA ANALYSIS (1/4)

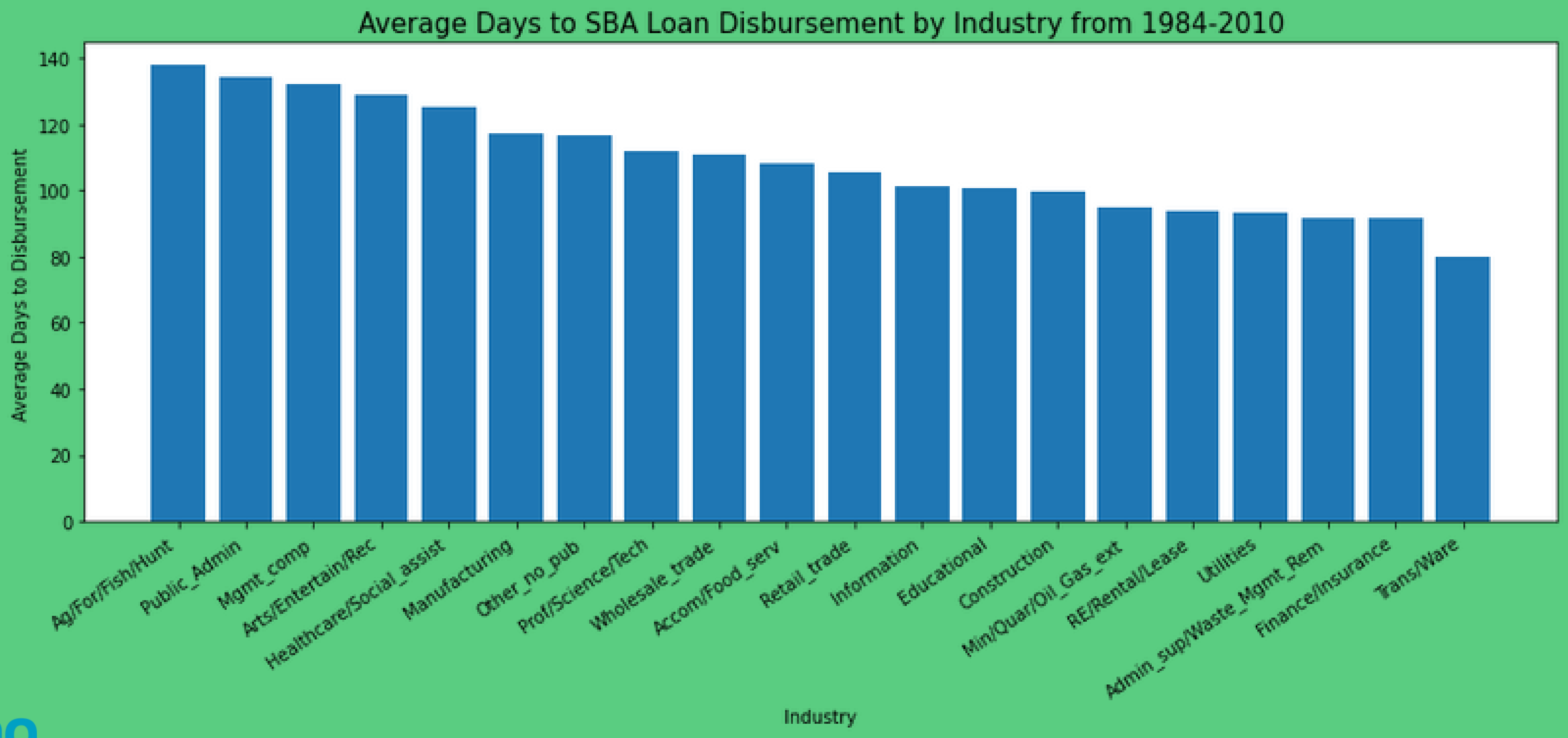
1. Retail trade and Manufacturing industries had significantly more loan funds. distributed to them during the sample period compared to other industries.
2. Although the Agriculture, forestry, fishing and hunting, Mining, quarrying, and oil and gas extraction, and Management of companies and enterprises industries had a small amount of total loan funds distributed to them during this time relative to most other industries, they had the highest average loan amount compared to other industries; This suggests they had a small number of large loans.



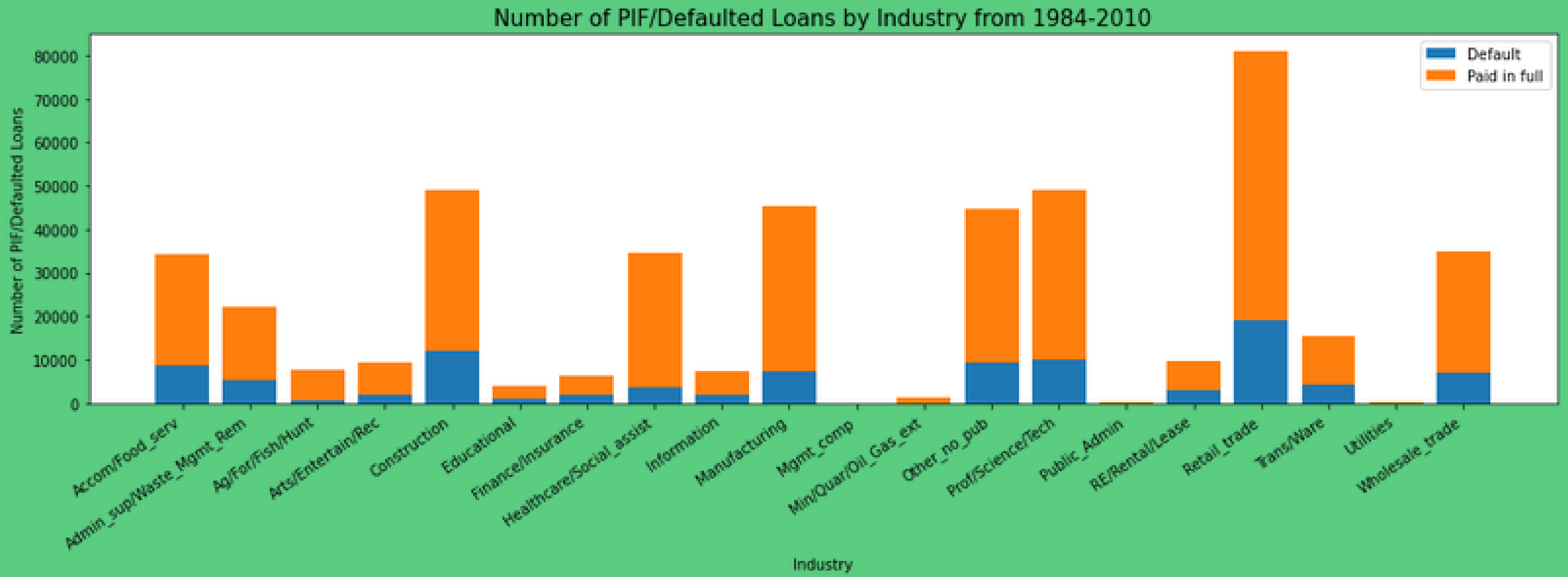


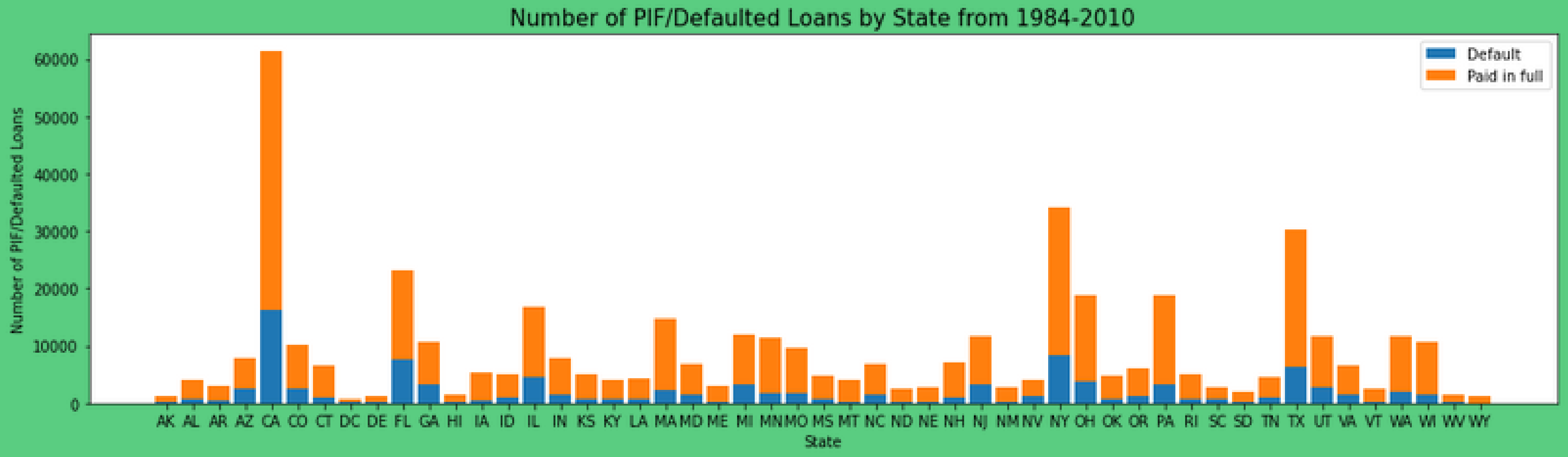
EXPLORATORY DATA ANALYSIS (2/4)

3. Interestingly, 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.



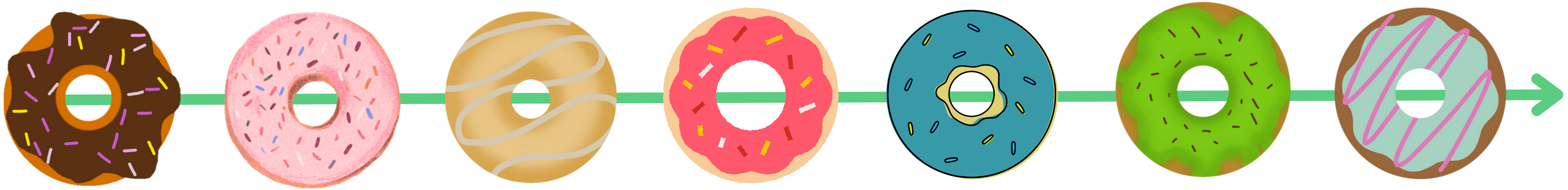
EXPLORATORY DATA ANALYSIS (3/4)



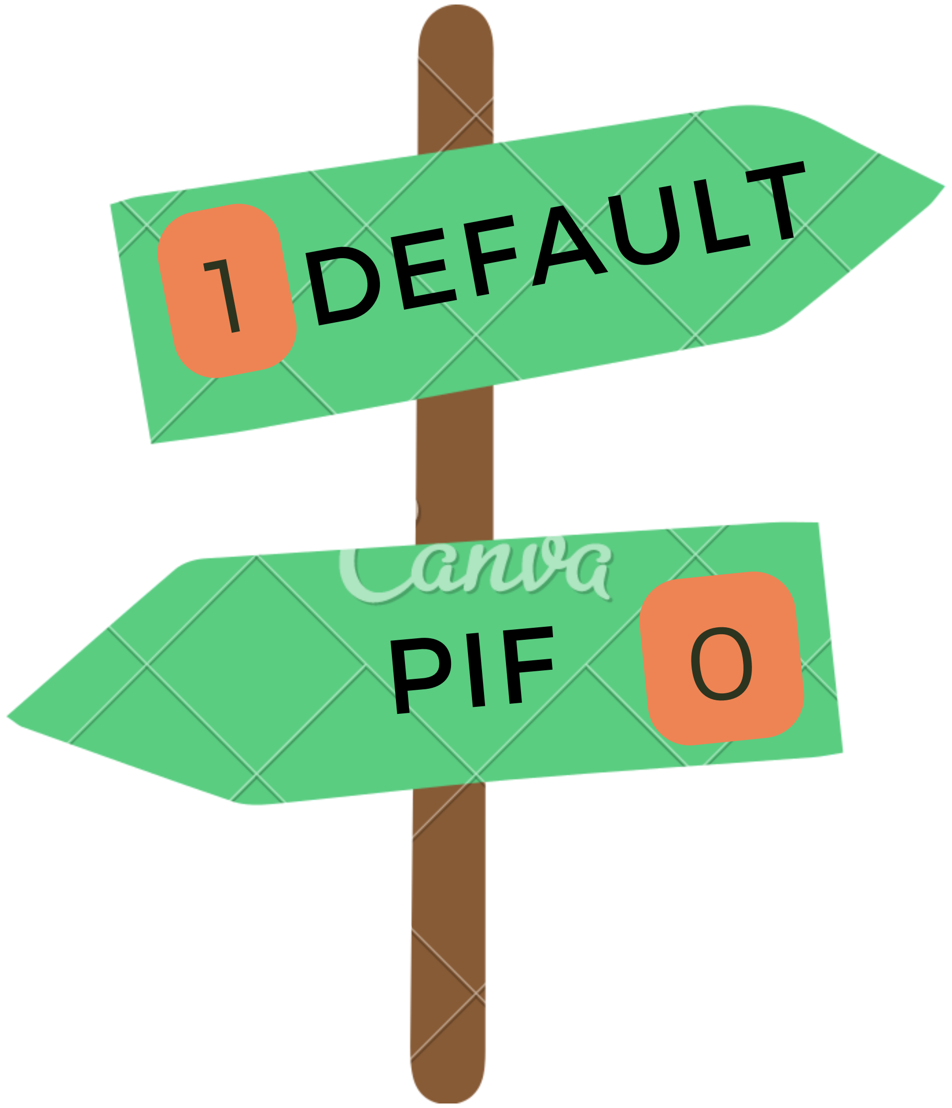


1. Industries with the highest number of loans during sample period: Retail trade (78,554), Professional, scientific and technical services (47,081) and Construction (47,047).
2. Industries with the highest Default percentage: Finance and Insurance (34.4%), Real Estate and rental leasing (33.8%) and Transportation and warehousing (30.7%).
3. States with the highest number of loans during sample period: California (59,121), New York (33,059) and Texas (28,941) State with the highest Default percentage: Florida (33.8%), Arizona (32.6%) and Nevada (31.6%).

BUILD THE MODEL



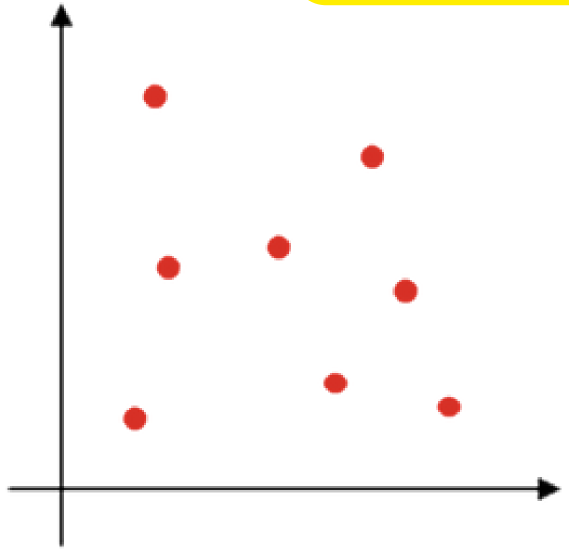
TARGET



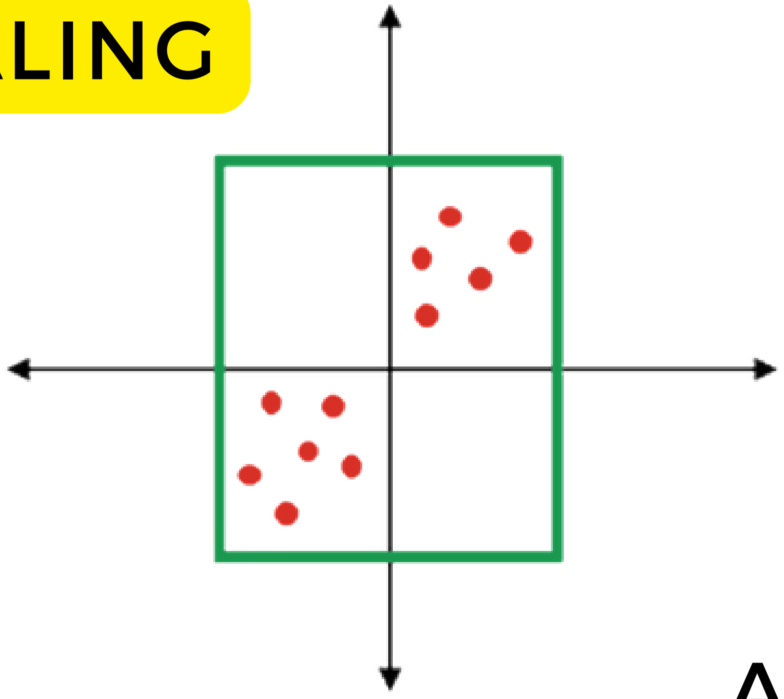
SPLIT DATA

- 75% Test Data
- 25% Train Data

DATA SCALING



ACTUAL DATA



AFTER STANDARIZATION

StandardScaler to normalize the data so that the data used does not have large deviations

ibimbing

FITTING MODEL AND EVALUATION

BASELINE MODEL

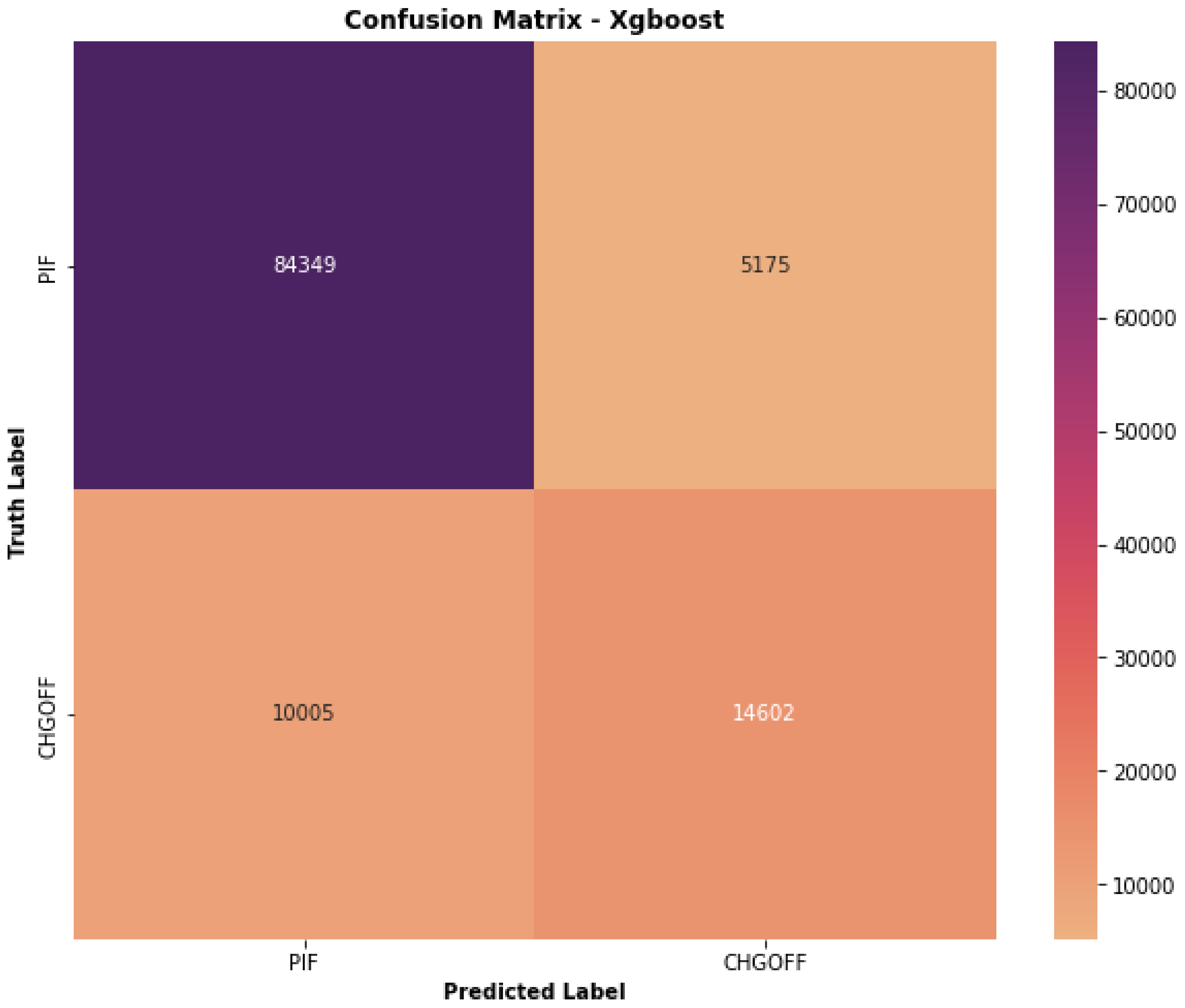
Model	Accuracy_Training_Set	Accuracy_Test_Set	Precision	Recall	f1_score
LogisticRegression	0.860593	0.859311	0.743701	0.530174	0.619042
XGBClassifier	0.961722	0.955621	0.906888	0.885033	0.895827
DecisionTreeClassifier	1.000000	0.933778	0.844856	0.848702	0.846775
RandomForestClassifier	0.999988	0.946264	0.922205	0.819929	0.868065

```
0    358180
1     98343
Name: Default, dtype: int64
```

Accuracy is percentage of prediction were correct.
Recall is actual positive rate.
Precision is predicted positive rate
F1-Score combines Recall and Precision to one performance matrix.

1. The ratio of target column show 78 : 22, we can't do balancing process to remodify the dataset because the difference is very high.
2. From model evaluation above, we know that the target data is not balance, so we can take out or ignore the accuracy score.
3. XGBoost and Random Forest are the model with high score in precision, recall, and f1 score. But we will choose XGboost (the highest) for the next tuning parameter process.

CONFUSION MATRIX



ibimbing

FEATURE SELECTION

BUILD PIPELING FOR FEATURE SELECTION AND MODELING;

SELECTKBEST DEFAULTS TO TOP 10 FEATURES

	precision	recall	f1-score	support
0	0.966	0.969	0.967	89524
1	0.884	0.875	0.880	24607
accuracy			0.948	114131
macro avg	0.925	0.922	0.923	114131
weighted avg	0.948	0.948	0.948	114131

It looks like reducing the number of features, and thereby dimensionality of the data, didn't affect the results too drastically. In fact, this model would likely perform better in a real world test because it is far more generalized.

1. ApprovalFY = 0.13102008
2. CreateJob = 0.026480757
3. DisbursementGross = 0.047658574
4. GrAppv = 0.09509127
5. IsFranchise = 0.07559523
6. LowDoc = 0.22293459
7. NoEmp = 0.017324772
8. RetainedJob = 0.01678543
9. RevLineCr = 0.08028162
10. Term = 0.2868277

CONCLUSION

Term is the most important feature. Term of a loan is highly related to real estate ownership. Loans with longer term (>= 240 months) are loans backed by real estate, whereas loans with shorter term (<240) are not loans backed by real estate. The ownership of land or real estate is often large enough to cover the amount of any principal outstanding. So this can lead to reduce of the probability of default.

RECOMMENDATION

There is something else that isn't captured in this data that is arguably the most important and relevant factor in determining the ability of a business to repay the loan: the business owner(s) and the business operations themselves! Although the industry does have some weight in this aspect, the data doesn't include the cash flow of each business, working capital, existing debt they had prior to applying for the SBA loan, etc.