**Loan Prediction**

**Should the loan be approved?**

|  |  |  |  |
| --- | --- | --- | --- |
| Glenda Pinho  SEIS 763 Machine Learning  University of St Thomas  St Paul, MN, USA  sant3463@stthomas.edu | Mollika Tahsin  SEIS 763 Machine Learning  University of St Thomas  St Paul, MN, USA  tahs9401@stthomas.edu | Somdath Rampersad  SEIS 763 Machine Learning  University of St Thomas  St Paul, MN, USA  ramp6791@stthomas.edu | Subhransu Nanda  SEIS 763 Machine Learning  University of St Thomas  St Paul, MN, USA  nand7313@stthomas.edu |

**ABSTRACT**

This paper looks at a vast dataset (almost 900,000 instances) of historical loan approvals and their default rates, in order to predict whether an SBA loan should be approved or denied based on the loan’s default risk profile assessed from SBA loan data. The dataset was acquired from the U.S. Small Business Administration (SBA), which acts as an insurer for bank loans to small businesses. After cleaning the data, different methodologies were used to determine which variables play significant roles in determining the credit worthiness of a loan. Different machine learning algorithms like logistic regression, k-nearest neighbors (k-NN), and support vector machines were used to design and evaluate a predictive model for loan approval. The best prediction accuracy is 0.9957 with Kernel SVM with a linear kernel. Compared with logistic regression and k-nearest neighbors, kernel SVM is a better prediction model for this dataset.

**CCS CONCEPTS**

• Data Cleaning • Cross Validation • Feature Selection • Feature Extraction • Dimensionality Reduction • Classification • Kernel Methods • Supervised Learning • Modeling • OneHotEncoding • Model Selection

**KEYWORDS**

Loan Default, Logistic Regression, Kernel SVM, k-NN, PCA, LDA, K-fold, OneHotEncoding

1**Introduction**

The dataset is from the U.S. Small Business Administration (SBA). The U.S. SBA was founded to promote and assist small businesses. Since small businesses are the primary job creation sources in the US, they play a significant role in increasing work opportunities and decreasing unemployment. Hence, it is important to ensure that more and more small businesses are created and that they can survive and thrive.

SBA helps small businesses by acting as an insurer for bank loans provided to small businesses. They guarantee as much as 90% of the loan given to a small business, and if the business defaults on the loan, the SBA becomes responsible for the amount insured. Although this helps to decrease the risks for banks and encourages banks to give out loans to small businesses, it poses a significant risk for the SBA, and hence it is of crucial importance that loans’ default tendencies are accurately vetted before approving them.

Also, since SBA loans only guarantee a portion of the loans, if the small business defaults on the loan, the bank is still exposed to losses. Since small businesses tend to have a high default risk, banks tend to be very hesitant to issue the loans to small businesses. Although SBA guarantees most of the loan amount, banks will still incur some losses if a small business defaults on its SBA-guaranteed loan. Therefore, banks are still faced with a difficult choice as to whether they should grant such a loan or not based on available data. Using machine learning to analyze historical data and predict loan default characteristics can help banks and the SBA make more informed decisions about loan approvals. This paper analyzes historical data and attempts to classify between whether a loan should be approved or not, based on a loan’s risk profile.

1.2**Inspiration**

This dataset was analyzed by multiple individuals on kaggle and was also used as a diversified dataset for teaching statistical modeling to undergraduate and graduate students. Machine learning based loan default prediction can play a significant profitable role in the fintech sector. The peer-to-peer lending industry in China and around the globe heavily makes use of such predictions. Recently, although the Chinese peer to peer lending market has been on a downturn, if the models can be enhanced enough to achieve strong accuracy and applicability, such data models can be used to build highly profitable and efficient peer to peer or consumer to business lending markets across the globe.

1.3**Goal**

The goal is to build a model that can accurately predict whether a loan should be approved or not.

2**Dataset Summary**

The dataset contains historical data from 1984 to 2014 of approved SBA loans. It comprises 899,164 instances and 27 feature variables. Appendix I summarizes some of the characteristics of the data. There are 15 categorical and 12 numeric variables.

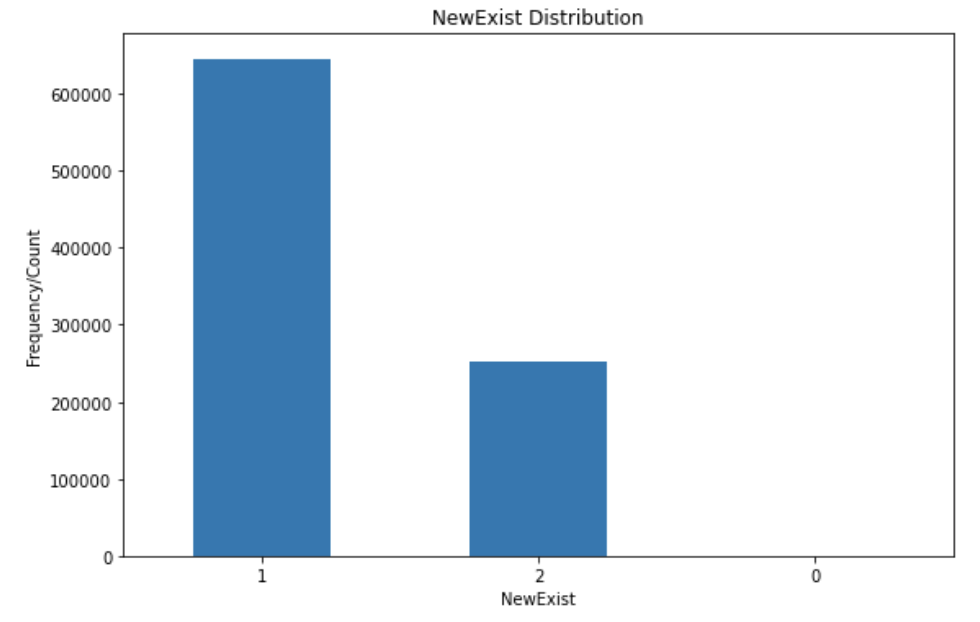


Figure 1-a: Distribution of NewExit

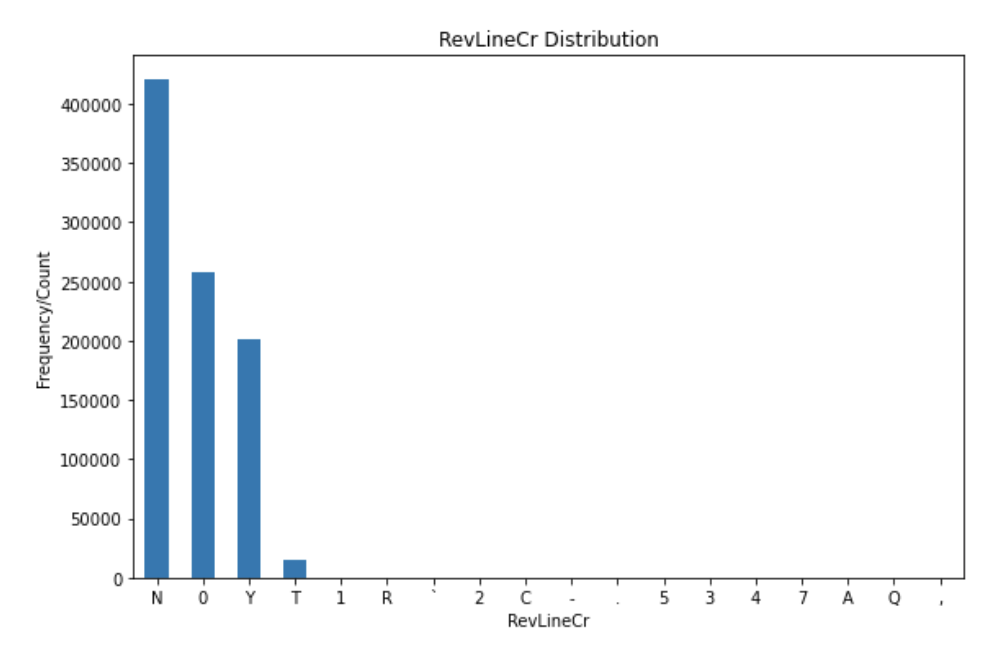


Figure 1-b: Distribution of RevLineCr

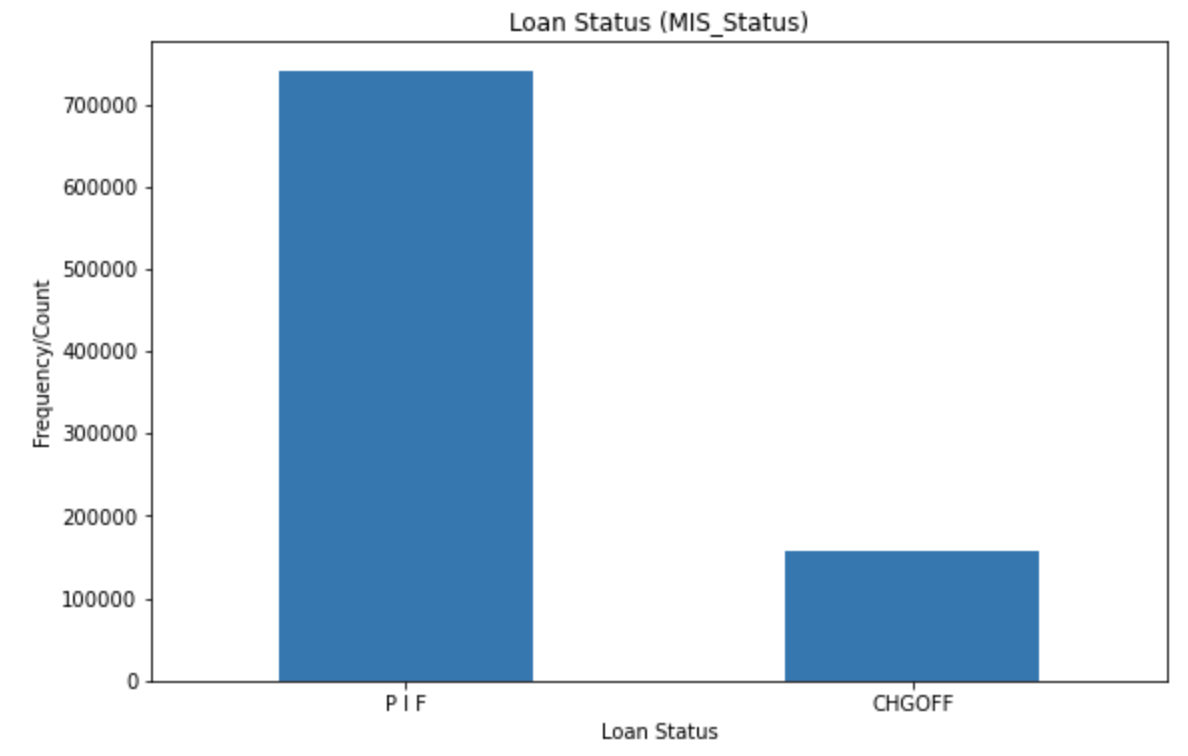


Figure 1-c: Distribution of MIS\_Status

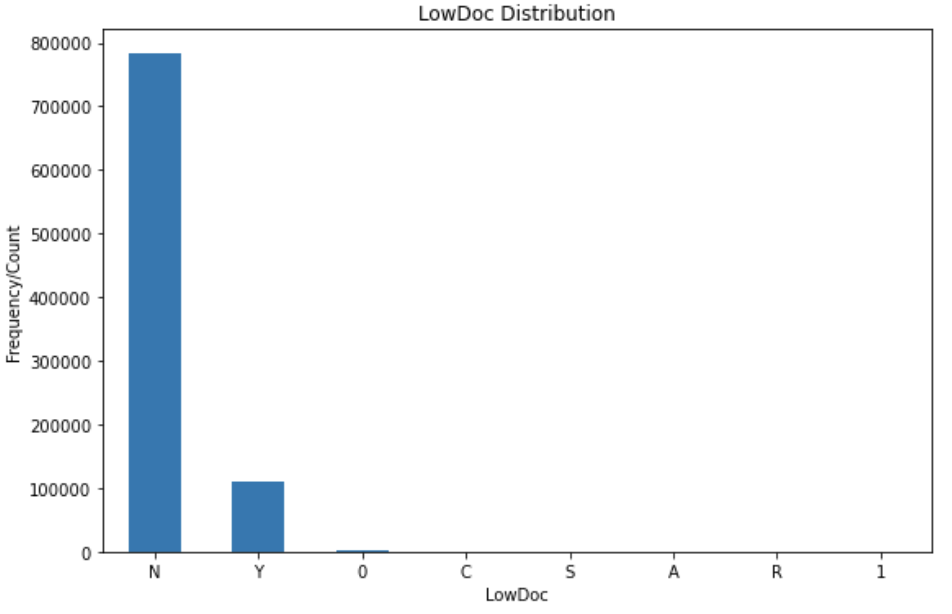


Figure 1-d: Distribution of LowDoc

Figures 1 (a-d) above show the distributions of several categorical variables. From the plots, it can be seen that there are variables with missing or invalid values, and imbalances in the distributions.

The target variable, the loan status, has an imbalanced distribution. The imbalance could possibly be a natural occurrence given that the dataset comprises approved loans which were already considered to be creditworthy. To overcome the class imbalance, a random sample, of equal number of instances, from each category (PIF, CHGOFF) will be taken to create the dataset for analysis, followed by a 70/30 split of this dataset for training and validation data. Given the large sample size of data, it can be assumed that this approach would compensate the imbalanced target variable distribution.

2.1**Data Quality Issues**

When cleaning the data, many data quality issues were found. The notebook documents all data quality issues and how and why each instance was cleaned.

Here are some of the issues we had to address:

* Remove the dollar signs and commas from the currency columns.
* Date column contradictions - There were some instances where the DistributionDate was before the ApprovalDate. There were other instances where the ChgOffDate was before the DistributionDate.
  + When the dates were imported, they were in the format dd-mmm-yy, this was an issue for dates, for instance, 16-Aug-66. Since the century is not provided, it is assumed that the century is the current century so the year 66 becomes 2066 instead of 1966. This is the primary reason why negative values were encountered between the charge-off dates and the disbursement dates.
* Erroneous data - some of the columns had values that didn’t mean anything based on the data description provided. RevLinCr, LowDoc, NewExist all had erroneous data.
  + To fix erroneous values/categories, it was decided to first replace all erroneous values with null and then treat them as missing values.
* Missing data - A lot of the columns had missing data. Some were able to be estimated based on the data in other columns. Others were imputed using the mode.
* Data mismatch: There were some instances where the loan was marked as CHGOFF but have a zero ChgOffPrinGr. Other instances were marked as PIF and had a non-zero ChgOffPrinGr.
* Unbalanced data: There were issues with some variables’ data being unbalanced, for example, MIS\_Status, State, etc.

2.2**Bias**

Since this paper uses a dataset derived from another paper, there may be inherent biases in the dataset. For example, some bias may have been intentionally introduced because the dataset was put together for introductory and advanced statistics/data analysis courses.

Secondly, this dataset is a retroactive look at loan default patterns. That is, it only looks at loans that were already approved as credit worthy based on loan data criteria and were later deemed to be defaulted or paid in full. Hence, most of the data instances in this dataset has loans that were paid in full and very little instances of loans that defaulted. This leads to an unbalanced dataset. However, this unbalance was corrected by taking an equal number of random samples from both instances - defaulted and paid in full.

Other inherent biases are as follows:

1. SBA loan quantities fluctuate through economic cycles. They tend to be more common during recessions. Also, loans tend to have a higher probability of defaulting during economic recessions.
2. Loans’ default risks also depend sometimes on the industry. Some industries, particularly those with barriers to entry like Medicine practice, law office, etc much safer industries for loans than other industries.

2.3**Data Manipulation**

The libraries used for the data manipulation, processing and modeling were Pandas, Sklearn, Matplotlib and NumPy.

To begin to explore the dataset, it was first imported in the CSV format and then loaded into a jupyter notebook for better manipulation and identifying any potential categorical variable that would require additional manipulation.

After thorough examination it was decided to skip the following columns because they do not provide any useful information on whether the loan will default or not. The columns removed were: LoanNr\_ChkDgt, Name, Bank, and BankState.

Some columns were found to have too many categories which would have created lots of issues for our ML algorithms and thus were dropped. These columns were dropped: City, State, and Zip.

Efforts were taken to convert certain categorical columns from a high number of categories to the least possible number of categories by using the values to create a new column and drop the old columns. In other cases, numerical columns were used to create a new column or compute to form a new numerical column.

* NAICS→Industry: First 2 digits were removed, and the rest were presented as categories of the Industry column while dropping the NAICS column completely. Industry was dropped in the end as it still had a lot of categories which might have affected the model.
* FranchiseCode→IsFranchise: In this column, observations with 0 or 1 had no observations and rest were Franchise codes. So it was decided to consider a new column IsFranchise, wherein if a FranchiseCode is greater than 1, then it is considered a franchise. If a FranchiseCode is less than or equal to 1, then it's considered not a franchise.
* Term→ RealEstate: “Term” was the only column which helped form a new column to help add a new feature to better fit a model with “Term” keeping its relevance intact. From the Loan Term Distribution plot, there seemed to be a pretty large range of term values. The majority of the loans seem to be around 84 months. It was considered breaking this down into subranges or whether the loan is backed by real estate. The original paper classified loans 20 years or greater (>=240 months) to be backed by real estate. Thus, it was decided to create a new RealEstate column that indicates whether the loan is backed by real estate. 1 = backed by real estate, 0 = not backed by real estate.
* DisbursementDate, ChgOffDate→ TimeToDefault: 0 if paid in full, number of days from disbursement to charge off if defaulted.

UrbanRural column had an unusually high number of undefined values and thus it was dropped.

Columns like GrAppv and SBA\_Appv were found to not provide any information about the borrower's credit worthiness and thus were dropped.

Certain steps were taken to fix some issues with Date Time Columns.

* All bad DisbursementDate values were replaced with the ApprovalDate.
* Assuming the loan has to be approved before disbursing the funds, all null DisbursementDate values were replaced with the ApprovalDate.

Following steps were taken to validate MIS\_Status:

1. Replace null MIS\_Status values with 'PIF' if the ChgOffPrinGr equals zero and ChgOffDate equals the DisbursementDate.
2. Replace null MIS\_Status values with 'CHGOFF' if the ChgOffPrinGr is greater than zero and ChgOffDate is greater than the DisbursementDate.

2.4**Variable Selection**

Variable selection was based on whether the data would provide information on the creditworthiness of a borrower. Some columns were removed because they had too many categories and drastically increased the computational complexity. All other variables were extracted using Dimensionality techniques like PCA, LDA and Kernel PCA.

3**Prediction Models**

In Modeling, the previously cleaned dataset was fitted with different Classifiers with and without Dimensionality Reductions. Before that, the data has to be prepared. The following steps are performed to match the requirements of all the machine learning algorithms and also to remove bias in the results:

* OneHotEncoding: In some Machine Learning algorithms the input and the output variables should be numerical, therefore it is required to perform OneHotEncoding so as to convert the categorical variables in the dataset. Here those categorical variables are NewExist, RevLineCr and LowDoc.
* Removal of PIF rows: The plot after code cell no. 13 shows the count of each categories, PIF and CHGOFF which is hugely imbalanced. The extra PIF rows have been removed so that each new predicted observation will have an equal chance of being PIF or CHGOFF.
* Dimensionality Reduction, Cross Validation and Model Selection: In order to find the best ML model which fits the dataset, some concepts have been put to use. 3 types of Classifiers were selected namely Logistic Regression, K-NN and Kernel SVM. Each of the classifiers used no Dimensionality Reduction or Feature Selection which were performed during previous data cleaning processes as the first case and PCA with 2 dimensions and LDA with 1 dimension as Dimensionality Reduction techniques in other cases. Then GridSearch was performed to select optimal parameters in each case and then K-fold validation on the Training sets in each case. Then the Mean and Standard Deviation of Model Accuracies for each case was computed under each Classifier and compared to find the best ML model to fit.

**3.1 Feature Selection and Extraction**

First the dataset was built with all possible features by selecting the columns with the least number of categories and those with high relevance to our class variable. Next step was applying dimensionality reduction techniques like PCA, LDA and Kernel PCA and then the dataset was fit into the model.

**3.2 Results**

|  |  |  |
| --- | --- | --- |
| ***Model*** | ***Mean Model Accuracy*** | ***Std Model Accuracy*** |
| LogisticRegressio1 | 0.993156 | 0.000448 |
| LogisticRegression-PCA | 0.712371 | 0.003501 |
| LogisticRegression-LDA | 0.985784 | 0.000921 |
| kNN | 0.994429 | 0.000151 |
| kNN-PCA | 0.845257 | 0.001491 |
| kNN-LDA | 0.985980 | 0.000031 |
| SVC | 0.995733 | 0.000053 |
| SVC-PCA | 0.822414 | 0.000336 |
| SVC-LDA | 0.986915 | 0.000080 |

The model that provided the best performance was SVC with a linear kernel. The mean was 0.9957 with a standard deviation of 0.00005.

Using the confusion matrix of test values, the test accuracy from our optimal model is 0.996. This matches the results we got from the previous section.

4.**Conclusion**

From our analysis we were able to identify a model with 99.6% accuracy. This is a very high accuracy level given that only a limited number of external features were used as variables for building the model. However, it is clear that the analysis was missing something very important: the financial statements of the businesses themselves. Although machine learning helped produce a model with great accuracy without business cash flow information, a more expanded analysis that takes business cash flows into account may yield prediction models with even greater accuracy and precision. It may also help to determine which metrics / variables carry the most significance for determining a business’ short- and long-term financial performance.

**ACKNOWLEDGMENTS**

We would like to thank Dr. Manjeet Rege for everything we have learned so far about Machine Learning.

This project consumed a huge amount of work, research and dedication. Still, implementation would not have been possible without clear instruction and proper guidance throughout the course.

**REFERENCES**

[1] Mirbek Toktogaraev. 2020. Should This Loan be Approved or Denied? (March 2020). Retrieved December 10, 2020 from https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied.

[2] All authors and Min Li. "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines. Retrieved December 10, 2020 from https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342.

[3] Kevinm6720. 2020. SBA Loan Approval Analysis. (May 2020). Retrieved December 10, 2020 from <https://www.kaggle.com/kevinm6720/sba-loan-approval-analysis>

[4] Beibei Niu, Jinzheng Ren, Ansa Zhao, and Xiaotao Li. 2020. Lender Trust on the P2P Lending: Analysis Based on Sentiment Analysis of Comment Text. (April 2020). Retrieved December 10, 2020 from <https://www.mdpi.com/2071-1050/12/8/3293/htm>.

[5] Alishba Imran. 2020. Lending Models using Machine Learning. (January 2020). Retrieved December 10, 2020 from <https://medium.com/analytics-vidhya/lending-models-using-machine-learning-d59d57f28ad7>.

[6] Anon. Deep Dense Convolutional Networks for Repayment Prediction ... Retrieved December 10, 2020 from<http://sclab.yonsei.ac.kr/publications/Papers/IC/2018_SOCO_JYK.pdf>.

**SUPPLEMENTAL**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Classification** | **# Missing Values** |
| LoanNr\_ChkDgt | Identifier Primary key | Categorical | 0 |
| Name | Borrower name | Categorical | 14 |
| City | Borrower city | Categorical | 30 |
| State | Borrower state | Categorical | 14 |
| Zip | Borrower zip code | Categorical | 0 |
| Bank | Bank name | Categorical | 1559 |
| BankState | Bank state | Categorical | 1566 |
| NAICS | North American industry classification system code | Categorical | 0 |
| ApprovalDate | Date SBA commitment issued | Numerical | 0 |
| ApprovalFY | Fiscal year of commitment | Categorical | 0 |
| Term | Loan term in months | Numerical | 0 |
| NoEmp | Number of business employees | Numerical | 0 |
| NewExist | 1 = Existing business,  2 = New business | Categorical | 136 |
| CreateJob | Number of jobs created | Numerical | 0 |
| RetainedJob | Number of jobs retained | Numerical | 0 |
| FranchiseCode | Franchise code,  (00000 or 00001) = No franchise | Categorical | 0 |
| UrbanRural | 1 = Urban,  2 = rural,  0 = undefined | Categorical | 0 |
| RevLineCr | Revolving line of credit:  Y = Yes, N = No | Categorical | 4528 |
| LowDoc | LowDoc Loan Program:  Y = Yes, N = No | Categorical | 2582 |
| ChgOffDate | The date when a loan is declared to be in default | Numerical | 736465 |
| DisbursementDate | Disbursement date | Numerical | 2368 |
| DisbursementGross | Amount disbursed | Numerical | 0 |
| BalanceGross | Gross amount outstanding | Numerical | 0 |
| MIS\_Status | Loan status:  charged off = CHGOFF,  Paid in full = PIF | Categorical | 1997 |
| ChgOffPrinGr | Charged-off amount | Numerical | 0 |
| GrAppv | Gross amount of loan approved by bank | Numerical | 0 |
| SBA\_Appv | SBA’s guaranteed amount of approved loan | Numerical | 0 |

Appendix I: Table with features, definitions along with data type and total of missing values.