Vehicle loan default detections

Akhil Sai Peddireddy
Charlottesville, USA
ap3ub@virginia.edu

Akhila
Charlottesville, USA
ar8aq@virginia.edu

Sanatkumar
Charlottesville, USA
sjk6dn@virginia.edu

Goutham
Charlottesville, USA
gp4em@virginia.edu

ABSTRACT

The increased vehicle loan rejection rates by the financial institutions are a direct result of increase in the default of vehicle loans which leads to significant losses for the institutions. Our goal is to identify the clients capable of repayment so that their loan is not rejected. The financial institution will also benefit from knowing which clients are likely to default on a vehicular loan. In order to address this issue we plan to implement a system of data mining algorithms to classify the loanee as defaulter or not a defaulter.

This project uses "LT Vehicle Loan Default Prediction" data set from Kaggle[**?**]. Various information regarding the loan and loanee are provided in the data set such as Loanee Information (Demographic data like age, Identity proof etc.), Loan Information (Disbursal details, loan to value ratio etc.), Bureau data history (Bureau score, number of active accounts, the status of other loans, credit history etc). On this data set, we performed classification using algorithms K-Nearest Neighbor (KNN)[2], Support Vector Machine (SVM)[3], Naive Bayes, Decision Trees, Random Forest[10], XGBoost[8] and Logistic regression. The dataset contains ground truth labels and thus we can find out which classification algorithms can perform better on this type of dataset. We use accuracy, recall and precision.

INTRODUCTION

When there is an increase in the default of vehicle loans, it leads to significant losses for the institutions which is in turn leads to the increase in the rejection rates of loans. To identify the clients capable of repayment so that their loan is not rejected, we use data mining techniques to improve the efficiency of vehicle loan approval process. This has major advantages like increasing customer satisfaction and reducing bad loans.

The data set being used is from a financial institution named LT. The data set is "LT Vehicle Loan Default Prediction" from Kaggle. It has about 233,000 Training Samples and 112,000 Test Samples. This data set has 40 features. Some of the important features in this data set are - Perform Cns Score (Bureau Score), Disbursed Amount (total amount that was disbursed for all the loans at the time of disbursement), ltv (Loan to Value of the asset), Current pincode (Current pincode of the customer), Primary Sanctioned Amount (total amount that was sanctioned for all the loans at the time of disbursement) and Primary overdue accounts (count of default accounts at the time of disbursement).

The main goal of the project is to identify loan defaulters.

This project has three major parts

1. Preliminary Analysis of the Data.
2. Using Data Mining Techniques for Prediction
3. Improvement of Prediction using advanced techniques

METHODOLOGY

The following are the main implementation steps in the project:

1. Exploratory Data Analysis
2. Data pre-processing
3. Classification using Data Mining algorithms
4. Using advanced techniques

Exploratory Data Analysis

Exploratory Data Analysis was performed in two stages

1. Analysing the features of the data set 2. Finding the most and least important features in the data set

Analysis of some Features

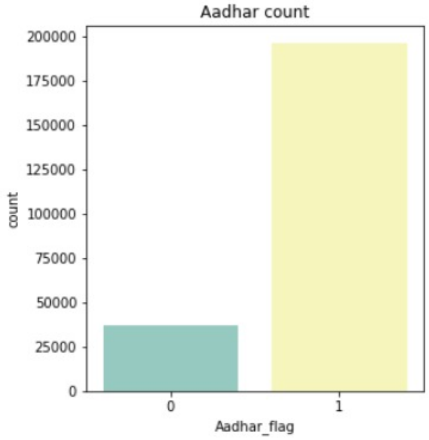
First we have analyzed the some features like Aadhaar flag, Voter flag and PAN flag.

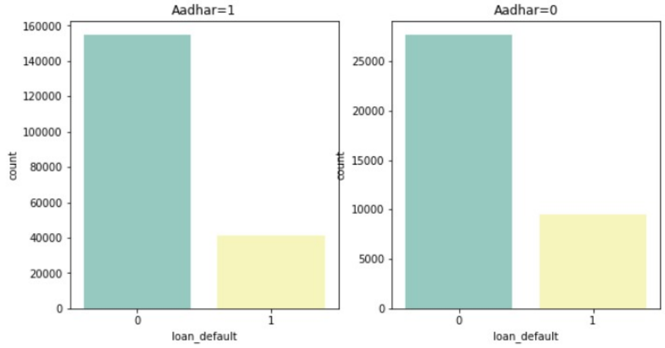
An Aadhaar card is a unique number issued to every citizen in India. This data set has a flag that tells if a client has aadhar card or not. We analyzed Aadhaar flag and found that in training set, the number of people having aadhar card is greater than the number of people not having aadhar card. Further analyzing it against the loan defaulters, we find that among the people having aadhar card, the percentage of them

ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

defaulting is about 21 percent whereas, among the people who do not have aadhar card, the percentage of people defaulting is around 27 percent



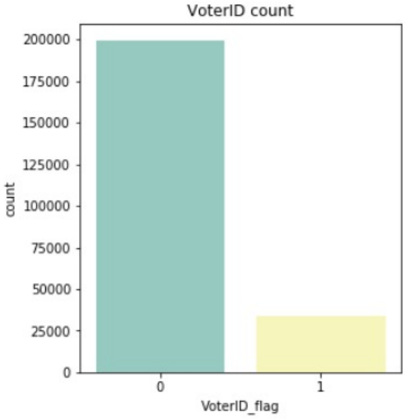


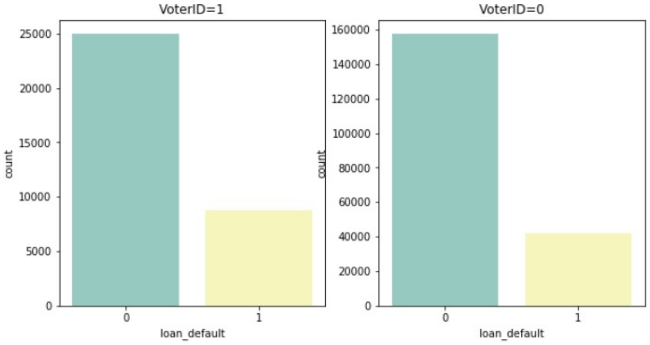
We see that this case is flipped in case of voter id. When analyzing Voter flag, we find that in training set, the number of people not having voter id is greater than the number of people having voter id. Further analyzing it against the loan defaulters, we find that among the people having voter id, about 25 percent default loans whereas among the people not having a voter id, around 20 percent of them default.

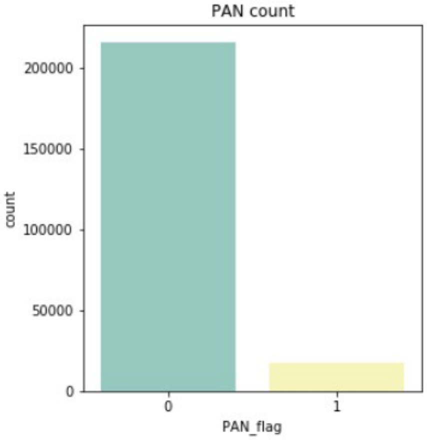
We further analyzed the PAN flag. A permanent account number is a ten-character alphanumeric identifier, issued in the form of a laminated "PAN card", by the Indian Income Tax Department. The PAN flag, in this data set tells whether a person has PAN card or not. When analyzing PAN flag, we find that in training set, the number of people not having PAN card is greater than the number of people having PAN card. Further analyzing it against the loan defaulters, we find that among the people having PAN card, about 23.5 percent default loans whereas among the people not having a PAN card, around 20 percent of them default loans.

Feature Importance

Next, we analyzed the importance of the features in predicting the loan defaulters and the following is what we found.





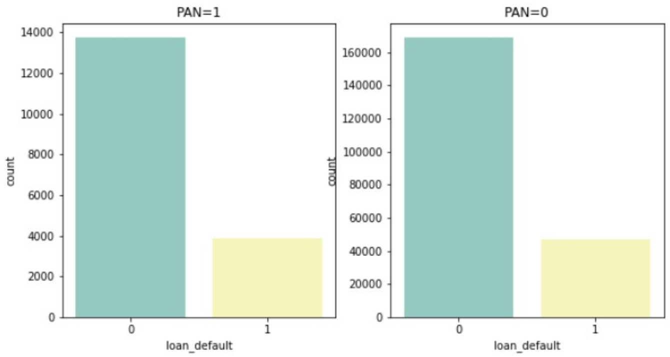


1. The top 4 Most Important features are CNS.score, Disbursed amount, LTV and Zip code

2. The top 3 Least Important features are DOB, New account in last 6 months and Secondary accounts

Data pre-processing

Data pre-processing is an important step when applying Data Mining techniques. The following are the various processing done on our dataset.



Dropped columns

We dropped certain columns that have all unique values or all rows have same value as these features do not add value to the prediction. UniqueID, Employee Code ID, Perform cns score description, Mobile No avl flag

Date to months

Converted date to months for certain features Average account age, Credit history length .

Date to Quarter

Extracted the quarters from the feature Disbursal date, which is a date feature

Categorized Date of Birth

Made buckets of 5 years for the feature "Date of Birth"

One hot encoding

One hot encoded some features like Disbursal date, Date of Birth

Scaling

Scaled the features when required. Certain algorithms like k nearest neighbours which is dependant on the distance, requires scaling and performs better prediciton when scaled. We have used Standard Scaler and Min Max scaler.

Classification using ML algorithms

Classification is the process of assigning data points to predefined classes or categories. In this project we have implemented six classification algorithms and a neural network which are described in the following sections.

Naive Bayes Classifier

Bayesian classifiers are statistical classifiers which predict class membership probabilities. They use Bayes theorem to calculate posterior probability. Naive Bayes is a simple classifier which assumes that the attributes are conditionally independent to each other, which simplifies the calculations.

Logistic Regression

Logistic Regression is a Statistical Learning technique. It is one of the Supervised Machine Learning methods used in Classification tasks.

K Nearest Neighbour Classifier

Nearest Neighbour Classification Algorithm is one of the classification algorithms that classifies the new data points based on the majority vote of it's nearest neighbours from the training data set.

SVM

SVM whose full form is Service Vector Machine, is another statistical learning technique. It is used to solve both linear and non-linear problems because it has non-linear kernels in addition to linear kernels such as rbf, polynomial, etc. The idea is to create a line or hyper plane that separates the data into classes and it tries to maximize the margin around it. It also has error factor which decides how much it should penalize the wrongly classified values.

Decision tree

Decision Tree is a tree flowchart like structure that divides the data into different subgroups based on conditions in order to classify the data. A condition is selected such that the classification is as pure as possible. At each node of the tree a decision is made about how to split the data and how to get the purest nodes. In order to calculate what attribute to split on, we can use different measures like Gini, entropy or misclassification error. When you travel down the tree, finally at leaf nodes we find the labels of the data of a particular sample.

XGboost

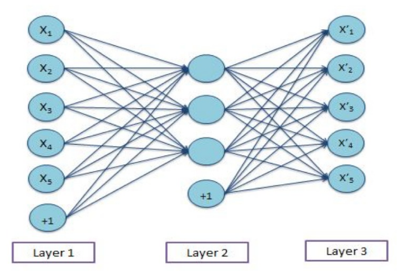
In boosting, we train multiple models and use weighted voting methods to classify the data. At every iteration we train a model with samples and check for wrongly classified samples. The weight for these samples is increased for the next iteration and new model is trained. After this training ends, all the weak learners are combined and finally used to classify a sample by running it through all the models. The predictions from these models are used as a weighted voting mechanism to classify the data.

Random Forest

For random forest, we select random features in order to check for best split attribute. And we build multiple such tress and use max voting classifier to classify the data.

Neural Networks

Neural Network algorithms are loosely based on our brain. They are multi-layer networks of neurons that are used for classification. The following figure shows how a basic neural network looks like. This image shows a neural network with 6 inputs, a middle layer containing 4 neurons and the layer 3 is output layer.



Use advanced techniques

SMOTE

SMOTE is the short form of Synthetic Minority Oversampling Technique. It is implemented by finding the k-nearest-neighbors for minority class observations (finding similar observations) and Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observation.

PCA

PCA, shortform for Principal Component Analysis is a statistical procedure which is useful for Dimensionality Reduction. Reducing the number of features is called Dimensionality Reduction. It transforms the input features and then we can drop the less important ones from the transformed features while still using the valuable parts of our original features.

EXPERIMENTS

1. Preliminary Results After data pre processing, we performed classification using algorithms K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, Decision Trees, Random Forest, XGBoost and Logistic regression. The following are the initial results of running these classification techniques on our data set.

   Naive Bayes
   Accuracy - 77.47 Precision - 36.5

   Recall - 4.65

   Logistic Regression
   Accuracy - 78.17 Precision - 40.48

   Recall - 0.49

   KNN
   Accuracy - 74.55 Precision - 29.074

   Recall - 11.72

   SVM
   Accuracy - 78.0 Precision - 28. 9

   Recall - 12.3

   XGboost
   Accuracy - 78.22 Precision - 53.1

   Recall - 7.3

   Random Forest
   Accuracy - 78.22 Precision - 50.13

   Recall - 1.12

   Decision tree
   Accuracy - 77.14 Precision - 31.5

   Recall - 4.25

2. Improvements

   From the preliminary results, we can see that the precision is really low, which is because it is not classifying one class of records properly.

   Data Imbalance

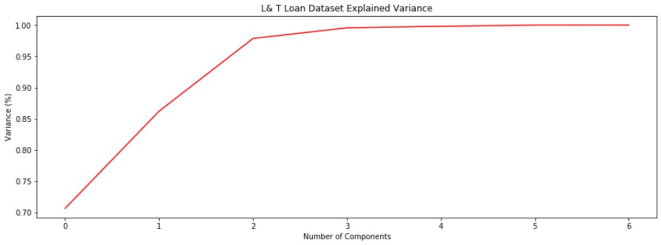
   We found that this is because of the data imbalance in the data set. The loan defaulter, which is our class label was imbalanced. In order to overcome this and to improve our precision (which is important because we have to correctly classify loan defaulters as loan defaulters, otherwise the company will incur losses), we use some advanced techniques as shown below

   SMOTE

   We have fixed the imbalanced data only on the training data. This will ensure that the generated data won't bleed into testing data and we can ensure that the result we get out of this can be generalised.

   PCA

   We have applied PCA on our data set for performing classification using SVM. This is beacause, SVM takes a lot of time given the size of our dataset and having around 40 features and PCA is very useful in such case. The following shows how the variance changes with the increase in the number of components. We can see that after 5 components, there is not much change in the variance percentage. We have used top 7 significant features returned by PCA to perform prediction. This would decrease the amount of time taken for training the model.



3. current results The following are the results obtained as a result of using SMOTE[5] and PCA[6] and then performing prediction. First, we display the confusion matrix results for each classification technique and then present the accuracy, precision and recall in the tabular format.

Naive Bayes

	0	1
0	17124	9162
1	4248	8405

Logistic Regression

	0	1
0	20238	6048
1	5162	7491

Random Forest

	0	1
0	25418	868
1	2779	9874

KNN

	0	1
0	19802	6484
1	4383	8270

Decision Tree

	0	1
0	25301	985
1	911	11742

SVM

	0	1
0	23098	3188
1	4966	7687

Gradient Boosting

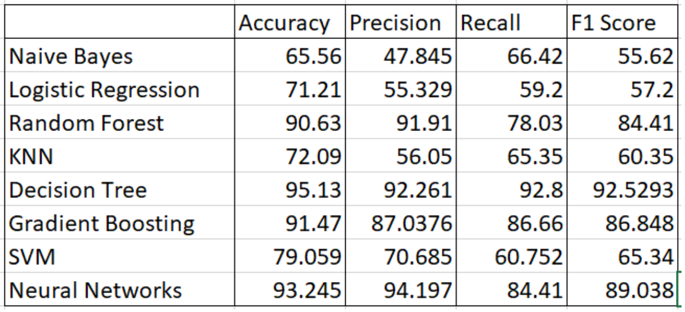
	0	1
0	24653	1633
1	1688	10965

Neural Networks

	0	1
0	25628	658
1	1972	10681

Final Results summarized

The following shows the final results of all the Algorithms run so far. We can see that Neural Networks gives the best prediction with a high precision of 94.2. Decision Tree is also comparable with a high accuracy of 95.13.



	Accuracy	Precision	Recall	F1 Score
Naive Bayes	65.56	47.845	66.42	55.62
Logistic Regression	71.21	55.329	59.2	57.2
Random Forest	90.63	91.91	78.03	84.41
KNN	72.09	56.05	65.35	60.35
Decision Tree	95.13	92.261	92.8	92.5293
Gradient Boosting	91.47	87.0376	86.66	86.848
SVM	79.059	70.685	60.752	65.34
Neural Networks	93.245	94.197	84.41	89.038

CONCLUSION

We have significantly improved the accuracy and precision of predicting a loan defaulter. We also found that Neural Networks gives the best prediction with a high precision of 94.2. Decision Tree is also comparable with a high accuracy of 95.13. We have some ideas on how this can be further improved. This could be the future work for the project. The accuracy can be further improved by creating an ensemble. We can also use data and feature engineering techniques to further improve accuracy. The Neural Networks can be further fine tuned and more models can be explored.

References

[1] 2019. LT Vehicle Loan Default Prediction. (2019). https://www.kaggle.com/mamtadhaker/lt-vehicle-loan-default-predictiondata_dictionary.csv

[2] T. Cover &P. Hart Mickey Haggblade.Nearest neighbor pattern classification, IEEE Transactions on Information Theory 2013.

[3] C.J.C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” submitted to Data Mining and Knowledge Discovery, 1998.

[4] Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.

[5] Nitesh V. Chawla , Kevin W. Bowyer. SMOTE: synthetic minority oversampling technique, Journal of Artificial Intelligence Research archive Volume 16 Issue 1, January 2002

[6] Hervé Abdi, Lynne J. Williams. Principal component analysis, WIREs Computational Statistics, 2010

[7] Aadhar, https://uidai.gov.in/

[8] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System, KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785-794

[9] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System, KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785-794

[10] Leo Breiman, Random Forests, Journal Machine Learning, Volume 45 Issue 1, October 1 2001, Pages 5-32