Loan Data Analysis

EE 660 Course Project

Project Type (1) Design a system based on real-world data
Number of student authors: 1

Weiding Huang
weidingh@usc.edu

03/12/2020

1 Abstract

Banks spend a lot of recourses going through loan applications, and having bad debts has negative effects on the banks. This project aims to use machine learning to predict whether a debt is going to be paid off or charged off using different algorithms such as logistic regression, random forest, k nearest neighbor, and EM estimator. The data was divided into pre-training, training, and test. Pre-training was used to observe the data, to find hidden information such as the distribution and the mean of the data and to help decide what algorithms to use. In training, cross-validation was used to eliminate any erratic performance of the algorithms. Having trained the data, logistic regression with L2 norm regularization was found to be the fastest and the most accurate algorithm, having about 80% accuracy. Hence logistic regression was used on the test set.

2 Introduction

2.1 Problem Type, Statement and Goals

Whether to issue a loan to a person requires a lot of investigation and time, and it will cause losses to the loaner or the bank if a lender is behind the due and the debt must be charged off. Typically, banks spend lots of resources and time on reviewing loan applications. In order to help banks to expedite the process, this project is aimed to build a model and classify whether a loan is going to be paid off or charged off. The data set used in the project contains 100514 training data points with 18 feature spaces and 10353 test data points. Among these 18 features, there are 7 categorical data and 11 numerical data. And 6 features contain missing data points. Due to the large data quantity, mixing of categorical and numeric data, and many missing data points, a comprehensive preprocessing step needed to be conducted in order to get the best result.

2.2 Literature Review

The project was previously done by Victor Hugo Pereira on Kaggle (https://www.kaggle.com/panamby/bank-loan-status-dataset). He filled in

missing data points using the mean values and reduced feature dimensions by removing collinear features. For classification models, he used logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, gradient boosting, and random forest. In this project, the best model was gradient boosting, resulting in an 81.7% accuracy on the test set.

2.3 Our Prior and Related Work

None

2.4 Overview of Our Approach

In the project, four different algorithms were experimented with to optimize the result. They were logistic regression, random forest, k nearest neighbor, and EM estimation. In addition, linear regression was used to predict and fill in the missing data points according to other data that did not have missing features. All for algorithms were given multiple parameter values in order to optimize the performance.

3 Implementation

3.1 Data Set

The bank loan dataset contains two sets: a training set and a test set. The training dataset contains 18 different features and 100514 data points (the last 514 are NaN). The test dataset contains 17 features (missing loan status) and 10353 data points (the last 353 are Nan).

Feature name	Categorical/ numerical	Number of missing data points	Description (cardinality, range)
Loan ID	Numerical	0	Serial number (eg.14dd8831-6af5-400b-83ec-68e61888a048)
Customer ID	Numerical	0	Serial number (eg.981165ec-3274-42f5-a3b4-d104041a9ca9)
Loan Status	Categorical	0	Fully Paid, Charged Off
Current Loan Amount	Numerical	0	10.8K – 100M
Term	Categorical	0	Short Term, Long Term

Credit Score	Numerical	1950	585-7510
Annual Income	Numerical	1950	81.092K -17.8M
Years in current job	Categorical	424	<1 year, 2 years, 3 years, 4 years, 5 years, 6 years, 7 years, 8 years, 9 years, 10+years
Home Ownership	Categorical	0	Home Mortgage, Rent, Own Home, HaveMortgage
Purpose	Categorical	0	Debt Consolidation, Home Improvements, Business Loan, Buy a Car, Medical Bills, Buy House, major purchase, Take a Trip, small business, Educational Expenses, moving, wedding, vacation, renewable energy
Monthly Debt	Numerical	0	0 - 436K
Years of Credit History	Numerical	0	3.6 - 70.5 years
Months since last delinquent	Numerical	5224	0 -131 Months
Number of Open Accounts	Numerical	0	1 - 55
Number of Credit Problems	Numerical	0	0 - 10
Current Credit Balance	Numerical	0	0 – 16.2 M
Maximum Open Credit	Numerical	0	1 – 14 M
Bankruptcies	Categorical	21	0 (not bankrupt), 1 (bankrupt)
Tax Liens	Categorical	0	0 (not a tax lien), 1 (tax lien)

3.2 Data Methodology

Firstly, by just observing the data on Kaggle, there are 514 rows of all NaN data. These were dropped. 5% of the data (500 data points) was used as pre-training data in order to determine the algorithms needed for filling the

missing data and for predicting results. 10 % of the data (1000 data points) was used as test data and 85% of the data (8500 data points) was used as training data. The split was done before any preprocessing to prevent data snooping.
In training, cross-validation was used to compute average accuracy for each algorithm. After finishing the preprocessing step, a 2-fold cross-validation was used to exam the results of training models. And the model with the best average accuracy was used in the test dataset that was separated from the credit_train. Also, the same model was used to predict the results in the credit_test set. Both of the separated test sets and the original test set were used only once.

3.3 Preprocessing, Feature Extraction, Dimensionality Adjustment

The first step was to get rid of useless features such as loan ID and customer ID which have no effects on the prediction accuracy.
The second step of preprocessing was to know how many data points are missing in each respective feature. The graph below shows the missing data points.

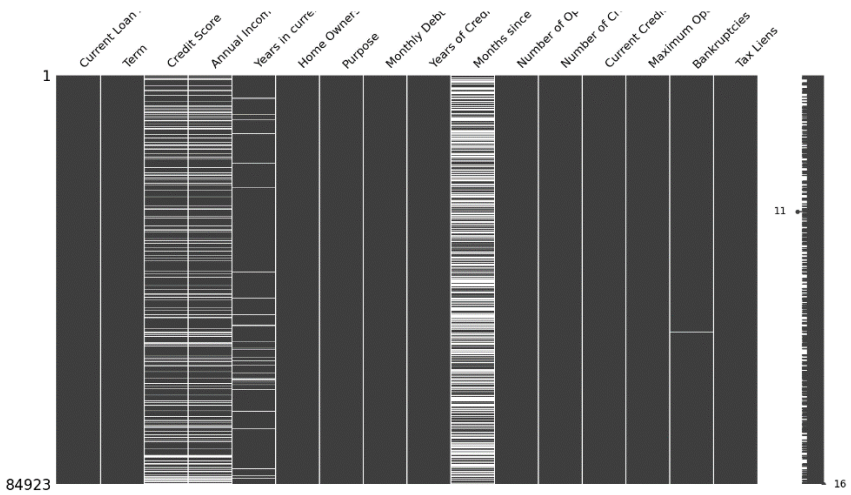

Figure 1: Visualization of missing data points

Feature ‘Months since last delinquent’ was mostly missing, so it was dropped.
The next step was to convert all the categorical data to numerical data.
Term: Short Term: -1, Long Term: 1
Years in current job: < 1 year: 0, 1 year: 1, 2 years: 2, 3 years: 3, 4 years:4, 5 years: 5, 6 years: 6, 7 years: 7, 8 years: 8, 9 years: 9, 10+ years: 10
Home Ownership: Rent: 1, Own Home: 2, Have Mortgage: 3, Home Mortgage:4
After all the data has been converted to numerical data, filling in the missing data points was the next step. The basic concept of how to fill in missing data is to predict the value with information from other complete feature spaces

that do not have missing data points. The algorithm used to predict the value was linear regression because some features have linear patterns. Linear regression has a faster computation time as well compared to other algorithm such as SVM.

3.4 Training Process

Logistic regression: logistic regression is often used in binary classification. It gives a probability of a data point with certain features belonging to a class: $Y(x) = P(Y = C|\, x, D)$. Logistic regression is a suitable algorithm for this project since this project is to predict a two-class result, whether the loan is going to be paid off or charged off. The logistic regression algorithm provided in sklearn allows to choose regularization from none or l2, but the results were almost identical. It might be the optimal solution was already inside the regularized area, so the regularization did not do much.

Parameter	Average accuracy over 2 cross validations
No regularization	81.9434304420527%
L2 norm regularization	81.9446079932173%

Random Forest: random forest consists of multiple decision trees. It uses bagging to average the prediction from every decision tree and taking the majority vote as the class for each stump.

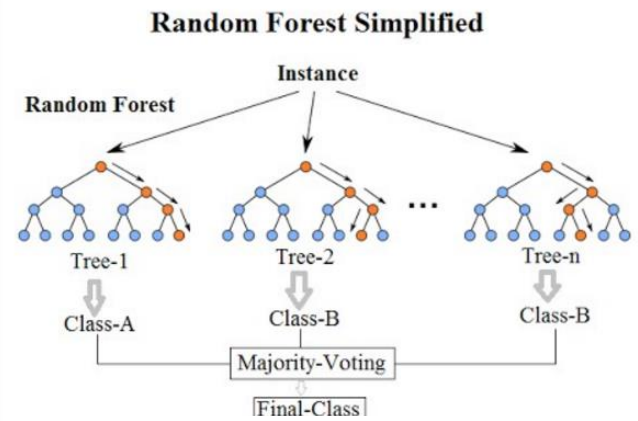

Figure 2 simplified random forest

In random forest algorithm, the number of decision trees will affect the performance. If there are too many decision trees, it might cause the data to overfit. Or there too few decision trees, it has lower accuracy. So multiple numbers of trees (500, 250, 100, 50) were chosen to find out the best parameter.

Parameter	Average accuracy over 2 cross validations

500 trees	81.92576717458373%
250 trees	81.86571206518923%
100 trees	81.80683450695932%
50 trees	81.73147123242505%

K nearest neighbor: k nearest neighbor algorithm has a similar methodology to the random forest algorithm. An unknown point is classified by the majority vote of its k nearest neighbor. In order to get the best performance, value k was chosen from 5, 10, 15, 20.

Parameter	Average accuracy over 2 cross validations
5 nearest neighbors	75.18193165493041%
10 nearest neighbors	75.80956642566119%
15 nearest neighbors	77.16728291844281%
20 nearest neighbors	77.15315230446762%
25 nearest neighbors	77.38512988389346%
50 nearest neighbors	77.43929723746497%
100 nearest neighbors	77.44047478862956%

EM estimator: EM algorithm iteratively improves the model to come up with the best fit to the data. The E step calculates the expected value of log-likelihood, and the M step calculates the parameters maximizing the log-likelihood. EM was selected because the data might have latent variables that are unknown. Some of the features might show patterns of normal distribution, but due to the high dimensionality, it is hard to be visualized. Two cross-validation sets were tested, and the average accuracy rate is the overall performance of two EM estimators.

3.5 Model Selection and Comparison of Results

The results showed that logistic regression and random forest are the two most promising models.

	Logistic regression	Random Forest	K nearest neighbor	EM estimator
Paramete	L2 norm	500 trees	100 trees	
Average	81.9446079932173%	81.92576717458373%	77.43929723746497%	49.97644897670804%

Dataset	Validation	Validation	Validation	Validation

Among all the results, both logistic regression and random forest provided similar results. This was unexpected and interesting. The two algorithms had similar results meaning that their decision boundaries are similar. But logistic regression gives a continuous, differentiable boundary while random forest does not. Both algorithms have similar accuracy meaning that their decision boundaries are similar.

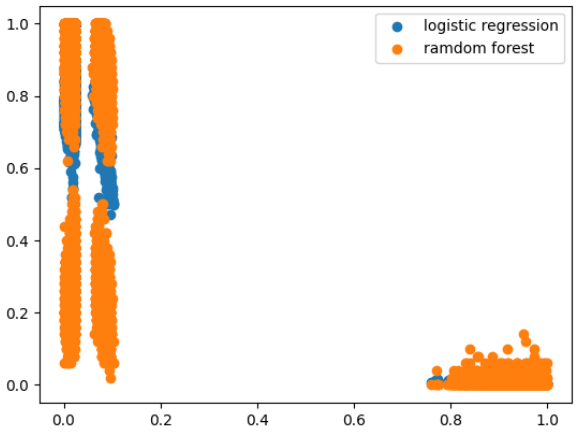

Figure 3: Both logistic regression and random forest have similar decision boundaries on credit score feature

EM estimator did not perform as well as others, this is expected since there are only a few features that are normal distributions. Also, when converting from categorical data to numerical data, there was no way the data distributions were normal. As a result, even some of the features are normally distributed, EM could not correctly classify features that are not normally distributed.

The final model that was used in the test set and used for prediction was logistic regression with l2 regularization. Though random forest gave a similar accuracy, the compute time was significantly slower than logistic regression due to the large tree size.

4 Final Results and Interpretation

Logistic regression was used for predicting results of the test set. In the test set, the missing data (credit score, annual income, years in current job, bankruptcies, maximum open credit, and tax liens) was filled in using the linear regression parameters learned from the training data. And then using logistic regression with l2 norm regularization, and according to feature current loan amount, term, credit score, annual income, years in current job, home ownership, monthly debt, years of credit history, number of open accounts, number of credit cards, current credit

balance, maximum open credit, bankruptcies, and tax liens, the prediction has about 80% accuracy, which is similar to the result from the cross-validations. Compared the result with the Victor Hugo Pereira's, the results are very similar. Below shows a chart of results from Victor Hugo Pereira's project.

Model	Accuracy
Logistic Regression	81.97%
K nearest neighbor	79.03%
Naïve Bayes	41.93%
Random Forest	80.17%
Gradient Boosting	81.98%

Although in Victor Hugo Pereira's project, he used medians to fill out all the missing numerical data and modes to fill out the missing categorical data and this project used linear regression to predict the missing data, which previously thought was more accurate, the results are very similar.
Logistic regression in both projects provided good results. Having observed the whole data, the reason might be that majority of the feature are linear separable, which was thought unlikely because real-world data usually is not linearly separable.
Random forest shows good results as well and less prone to overfitting even when having a large number of trees. But due to the nature of the data (most data are not tubular and some have strong correlations), random forest algorithm is not as efficient as logistic regression in this case.
For feature improvement, a more comprehensive understanding of data distribution from pre-training data can be very useful. Moreover, there might be a more appropriate way to predict values of missing data, such as EM estimator or other probability estimators. More precise predictions of missing values could lead to higher accuracy. Moreover, the feature dimension can be adjusted to allow faster computation time because some of the features seem to have a high correlation such as credit score and annual income.

5 Contributions of each team member

It is an individual project

6 Summary and conclusion

In conclusion, logistic regression was the best algorithm in this project. But it was not always the case since most of the real-world data are not linearly separable. This project also gives a glimpse of how to fill missing data points. A complex algorithm does not necessarily mean it will have a better prediction. The algorithm used must be thought through according to its distribution, its type, and other parameters that might not even be included in the data. pre-training is a

very important step to learn the data distribution and come up with appropriate algorithms. Although pre-training itself does not improve the model performance, it does help to choose appropriate algorithms. Especially when doing a real-world project whose data is often to be a jumbled mess. Also, feature reduction has a great effect on algorithm efficiency. If the features were to undergo a feature reduction step, random forest algorithms might out-perform logistic regression because there are fewer features spaces to be divided.

7 References

Kaggle.com. 2020. *Treating Categorical Data Before Feeding To Model | Data Science And Machine Learning*. [online] Available at: <https://www.kaggle.com/getting-started/55836> [Accessed 3 December 2020].

Regression?, W. and Learning?, W., 2020. *What Are The Advantages And Disadvantages Of Logistic Regression? | I2tutorials*. [online] i2tutorials. Available at: <https://www.i2tutorials.com/what-are-the-advantages-and-disadvantages-of-logistic-regression/> [Accessed 3 December 2020].

Kaggle.com. 2020. *Missing Data Imputation Using Regression*. [online] Available at: <https://www.kaggle.com/shashankasubrahmanya/missing-data-imputation-using-regression> [Accessed 3 December 2020].

Forest?, W., 2020. *When To Avoid Random Forest?*. [online] Cross Validated. Available at: <https://stats.stackexchange.com/questions/112148/when-to-avoid-random-forest#:~:text=Random%20forests%20basically%20only%20work,approximated%20by%20many%20rectangular%20partitions.> [Accessed 3 December 2020].