Comparison of the Logistic Regression, Decision Tree, and Random Forest to Predict Defaulted Loan

Xueer Zheng

Department of Computer Science, University College London

COMP0050 - Machine Learning with Applications in Finance

Dr. Fabio Caccioli

Feb 20, 2021

I. Abstract

Nowadays, as loans have become part of the main business of commercial banks, predicting whether a customer will default on their loans has become an important issue for commercial banks. This report based on the dataset called "All Lending Club Loan Data" on Kaggle, aiming to build an unsupervised learning model to predict a customer's default risk. After using SMOTE to oversample this unbalanced classification dataset, I compared the prediction capabilities of the three algorithms (logistic regression, decision tree, and random forest). Finally, due to the highest AUC, highest accuracy, highest recall and highest accuracy of the random forest classifier, we can conclude that the random forest classifier has shown the best performance in the defaulted loan prediction.

II. Data

```
[74]: cw=pd.read_csv('~/Desktop/comp50/dataCOMP0050Coursework1.csv', header = 0)
      cw=cw.dropna()
      print(cw.shape)
      print((cw.columns))

      (6802, 21)
      Index(['loan_amnt', 'term', 'installment', 'emp_length', 'home_ownership',
             'verification_status', 'issue_d', 'purpose', 'dti', 'earliest_cr_line',
             'open_acc', 'pub_rec', 'revol_util', 'total_acc', 'application_type',
             'mort_acc', 'pub_rec_bankruptcies', 'log_annual_inc', 'fico_score',
             'log_revol_bal', 'charged_off'],
            dtype='object')
```

The dataset called "All Lending Club Loan Data Version 6, 2018" comes from Kaggle. In this study, the classification purpose is to predict whether customers will default on their loans or not.

This dataset provides 6802 All Lending Club customers' records, including 21 fields: "loan_amnt", "term', 'installment", "emp_length", "home_ownership", "verification_status", "issue_d", "purpose", "dti", "earliest_cr_line", "open_acc", "pub_rec", "revol_util", "total_acc", "application_type", "mort_acc", "pub_rec_bankruptcies", "log_annual_inc", "fico_score", "og_revol_bal", and "charged_off."

1. **Target Variable "y"**

The target variable is "charged_off" which means whether a debtor repay the loan or not, whose value is 0 if the debtor repays, 1 otherwise.

2. **Independent Variables (features)**

Variable name	Numerical or Categorical	Variable Explanation

'loan_amnt'	numerical	the amount of money that a debtor lent
'term'	numerical	the length of time for a loan to be completely paid off
'installment'	numerical	the amount of money the debtor needs to pay regularly
'emp_length'	numerical	employment length of the debtor
'home_ownership'	categorical	("RENT", "OWN, "MORTGAGE")
'verification_status'	categorical	("SourceVerified","Verified", "Not Verified"), whether the income source is verified by the All Lending Club
'issue_d'	categorical	issue date of the loan
'purpose'	categorical	the purpose of loan
'dti'	numerical	debt to income ratio
'earliest_cr_line'	numerical	the earliest credit line
'open_acc'	numerical	the number of open credit lines on the debtor's record
'pub_rec'	numerical	the number of derogatory public records of a debtor
'revol_util'	numerical	revolving utilization ratio
'total_acc'	numerical	total number of historical credit lines on the debtor's record
'application_type'	categorical	('Individual', 'Joint App')
'mort_acc'	numerical	total number of credit lines
'pub_rec_bankruptcies'	numerical	the number of bankruptcies in debtor's public record
'log_annual_inc'	numerical	logarithm of the debtor's annual income
'fico_score'	numerical	Fair Isaac Corporation (FICO) score of a debtor
log_revol_bal'	numerical	logarithm of total credit revolving balance

3. Data Cleaning

Firstly, as the variable "verification_status" has three categories, I calculated the category means of target variable "charge_off". As the category mean of "Source Verified" is close to that of "Verified", I grouped these two categories into "Verified".

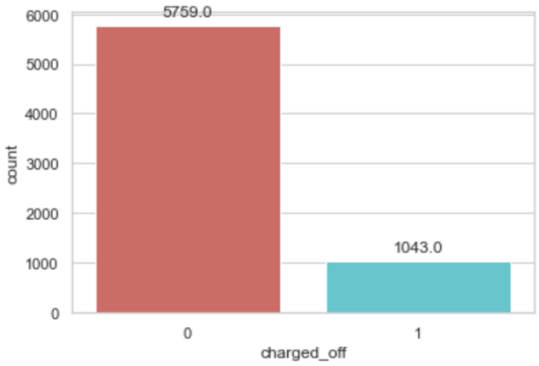
Secondly, because the variable called "issue_d" has lots of (= 67) categories, I dropped this variable from the dataset. Moreover, variable "purpose" was dropped as the only purpose of loan is "car."

Finally, for better modeling, the categorical variables "home_ownership", "verification_status", and "application_type" were converted into dummy variables. After that, the dataset would have 22 independent variables :"loan_amnt", "term", "installment', "emp_length", "dti", "earliest_cr_line", "open_acc" , "pub_rec", "revol_util", "total_acc",

"mort_acc", "pub_rec_bankruptcies", "log_annual_inc", "fico_score", "log_revol_bal", "RENT", "MORTGAGE", "OWN", "Verified", "Not Verified", "Individual", and "Joint App."

III. Methodology

1. SMOTE for balancing data



The count plot shows the number of repayments in original data 5759 and the number of charge offs is 1043, which implies the original dataset has a severe class imbalance. In this case, the SMOTE (Synthetic Minority Oversampling Technique) was used to oversample this imbalanced classification dataset. Specifically, based on the random instances from the minority class of the original dataset, the SMOTE algorithm creates synthetic data points to balance a dataset (Ma & Fan, 2017).

The SMOTE algorithm is only used to balance the training set, because we want to make sure that no data in the test set involves creating synthetic data points. In other words, the training set and the test set are completely independent of each other. After applying SMOTE algorithm, we get a completely balanced dataset which contains 4021 repayments and 4021 charge-offs.

2. Recursive Feature Elimination (RFE) for selecting features

Recursive Feature Elimination (RFE) is a feature selection algorithm whose purpose is to retain more significant features and eliminate secondary features by repeatedly building models based on the balanced training data. In this study, RFE was used to select features in logistic regression Classifier, decision tree classifier, and random forest classifier. Moreover, all 22 features are ranked according their importance, and the selected features are assigned rank 1 which are saved for model fitting in next step.

3. Model Fitting (Logistic Regression, Decision Tree, and Random Forest)

Based on the selected features by RFE, three unsupervised learning algorithms (Logistic Regression, Classification Tree, and Random Forest) are used to fit the data and predict the

target variable "charge_off." The reasons that I chose these three algorithms is that they do not require feature scaling, and they can work well with large-scale dataset (Pranckevičius & Marcinkevičius, 2017).

4. Model Comparison

Finally, to compare the predictive performances among models, I generated the classification report for each classification model, draw ROC curve (Receiver Operating Characteristic curve), and compute the corresponding AUC (Area under the ROC Curve). In general, a model with good predictive power should have high precision and high recall, so it will have high AUC as well (Seshan & Begg, 2013). Based on these criteria, the best model among logistic regression, decision tree, and random forest will be selected.

IV. Results

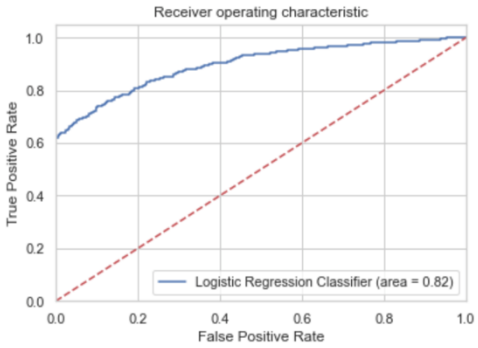
1. Logistic Regression

```
print(confusion_matrix(y_test, y_pred_lg))
print(classification_report(y_test, y_pred_lg))
```

```
[[755  67]
 [227 560]]
              precision    recall  f1-score   support

           0       0.77      0.92      0.84       822
           1       0.89      0.71      0.79       787

    accuracy                           0.82      1609
   macro avg       0.83      0.82      0.81      1609
weighted avg       0.83      0.82      0.82      1609
```



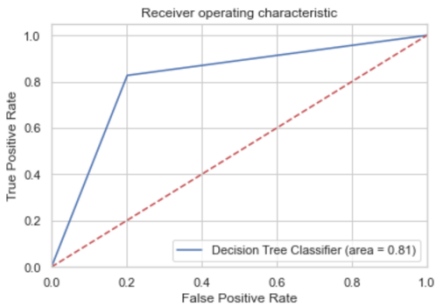
The features selected by RFE are: “term”, “open_acc”, “mort_acc”, "pub_rec_bankruptcies", "log_annual_inc", "log_revol_bal", "RENT", "MORTGAGE", "OWN", "Verified", and “Not Verified.” Based on these features, a logistic regression classifier is constructed, with a precision of 0.77 and a recall of 0.92 for the "repayment" category, a precision of 0.89 and a recall of 0.71 for the "charge off" category, an accuracy of 0.82, and an AUC of 0.82.

2. Decision Tree

```
[[ 940  259]
 [ 209 1005]]
              precision    recall  f1-score   support

           0       0.82      0.78      0.80      1199
           1       0.80      0.83      0.81      1214

    accuracy                           0.81      2413
   macro avg       0.81      0.81      0.81      2413
weighted avg       0.81      0.81      0.81      2413
```



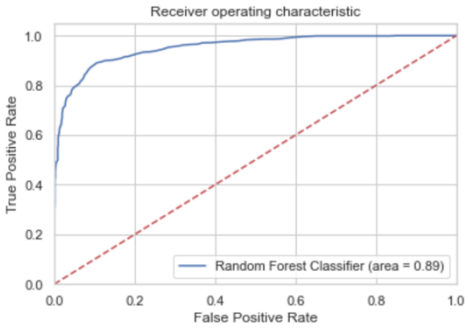
The features selected by RFE are : “loan_amnt”, “installment”, “emp_length”, “dti”, “revol_util”, “log_annual_inc”, “fico_score”, “log_revol_bal”, “RENT”, “MORTGAGE”, and “OWN.” Based on these features, a decision tree classifier is constructed, with a precision of 0.82 and a recall of 0.78 for the "repayment" category, a precision of 0.80 and a recall of 0.83 for the "charge off" category, an accuracy of 0.81, and an AUC of 0.81.

3. Random Forest

```
[[752  70]
 [108 679]]
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       822
           1       0.91      0.86      0.88       787

    accuracy                           0.89      1609
   macro avg       0.89      0.89      0.89      1609
weighted avg       0.89      0.89      0.89      1609
```



The features selected by RFE are : "loan_amnt", "installment", "emp_length", "dti", "revol_util", "mort_acc", "log_annual_inc", "fico_score", "log_revol_bal", "RENT", and "Not Verified." Based on these features, a random forest classifier is constructed, with a precision of 0.82 and a recall of 0.78 for the "repayment" category, a precision of 0.80 and a recall of 0.83 for the "charge off" category, an accuracy of 0.81, and an AUC of 0.81.

V. Conclusion

The goal of this study is to find the best model that can predict whether a customer will default on their loans. By comparing the AUC-OUC curves, the Random Forest Classifier has the highest AUC (=0.88881) and has the closest OUC curve to the upper left corner. Moreover, by comparing the classification reports of those three classifiers, the random forest classifier has the highest precision score for both cases ("repayment"=0.87, and "charge off"=0.91), and it also has the highest recall score for both cases ("repayment"=0.91, and "charge off"=0.86). Overall, the random forest classifier has the highest accuracy of prediction (=0.89).

Due to the highest AUC, highest accuracy, highest recall and highest accuracy of the random forest classifier, we can conclude that the random forest classifier has the best predictive performance compared to the logistic regression classifier and the classification tree classifier.

Reference

Ma, Li, & Fan, Suohai. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics, 18*(1), 169.

Pranckevičius, Tomas, & Marcinkevičius, Virginijus. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing, 5*(2), Baltic Journal of Modern Computing, 2017, Vol.5 (2).

Seshan, Venkatraman E, Gönen, Mithat, & Begg, Colin B. (2013). Comparing ROC curves derived from regression models. *Statistics in Medicine, 32*(9), 1483-1493.