Analysis of Inquired Loan Repayment in Full Based on Buyer's History

Aya Salka

Introduction:

The data set acquired for the project in question was received from LendingClub.com, a public company which screens borrowers as well as takes part in the transactions and facilitations of loans between different parties. The data in specific, was taken between the years 2007 and 2012 and includes a standard sized list of information of different buyers histories. The features are explained in more detail below.

Feature Description*(Note these descriptions were provided by LendingClub.com)*:

credit.policy: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
purpose: The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").
int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
installment: The monthly installments owed by the borrower if the loan is funded.
log.annual.inc: The natural log of the self-reported annual income of the borrower.
dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).
fico: The FICO credit score of the borrower.
days.with.cr.line: The number of days the borrower has had a credit line.
revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months.
delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

This project was aimed to fit and analyze different types of models to the data and find the best model which resembles the data in question. The independent variables in this dataset were all of the given non-categorical data with the exception of not fully paid, which was used as the dependent variable.

Methodology:

Upon receiving the data set and loading it into jupyter, some data cleaning was required to take out all of the null values as well as deal with the one categorical feature.

```
credit.policy        0
int.rate             0
installment          0
log.annual.inc       0
dti                  0
fico                 0
days.with.cr.line    0
revol.bal            0
revol.util           0
inq.last.6mths       0
delinq.2yrs          0
pub.rec              0
dtype: int64
```

After realizing that there were 0 null values in the data set, further exploration was granted as numerous plots and graphs were made to signify or note any patterns or relationships within the data.

After the exploratory phase, it was necessary to begin implementing models to test the fit of the data to numerous examples. This paper identifies and implements the models and techniques of Logistic Regression, Feature Selection and Scaling, Precision Recall Curves, Principal Component Analysis, Grid Search, Decision Trees, Random Forest Classifier Trees, and Gaussian Naives Bayes classifier. Implementation Descriptions can be found below.

Logistic Regression:
Specifically, this was the model that was mostly going to be implemented in the data as a binary logistic regression model. Below were the results after one was implemented:

Parameters

```
[[ 0.29522525 -0.07144138 -0.00705662  1.11042886  0.03221686  0.08960756
   0.13604877  0.05752726  0.00260927  0.03380929  0.16915648 -0.07525497
   1.78174461  0.06137697 -0.05746979 -0.12627367 -0.11294449 -0.19016788
   0.01739149  0.05584491]]
```

Classification Report

	precision	recall	f1-score	support
0	0.94	0.88	0.91	153
1	0.88	0.94	0.91	147
avg / total	0.91	0.91	0.91	300

This model's accuracy was close to 90.6%. To improve accuracy, two methods were implemented Feature Selection and Scaling.

Scaling:
This was used primarily because the data was not normalized. FICO scores were close to the hundreds ranges while interest rates were decimal points therefore, implementing scaling would be not only important in improving accuracy but in establishing data that was not too skewed.

Parameters

```
[[ 0.40787451 -0.07057281 -0.00770694  1.56553829  0.03166561  0.08982892
   0.1342854   0.05512562  0.00323899  0.03122193  0.16677645 -0.07456801
   2.08589006  0.06175621  0.01496563 -0.13043532 -0.10832297 -0.18588641
   0.01741625  0.05540212]]
```

Classification Report

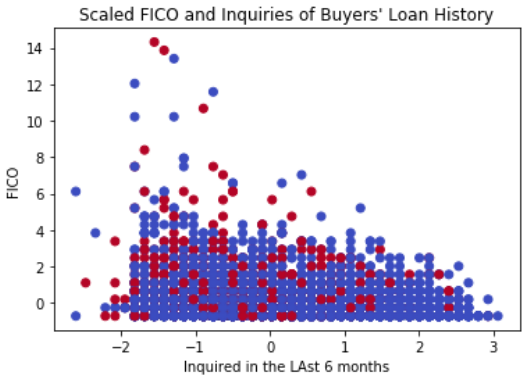
	precision	recall	f1-score	support
0	0.95	0.87	0.91	153
1	0.88	0.95	0.91	147
avg / total	0.91	0.91	0.91	300

The implementation was done and, surprisingly, the accuracy of the model only increased by a minute amount after Logistic Regression was performed again. Specifically, it increased to exactly 91%. Therefore, another implementation was performed: Feature Selection.

Feature Selection:
Feature selection was used to figure out the top two features in the set and run those against the not fully paid binary column. The output of ranking was as follows:

[5 4 2 3 9 1 11 6 10 1 7 8]

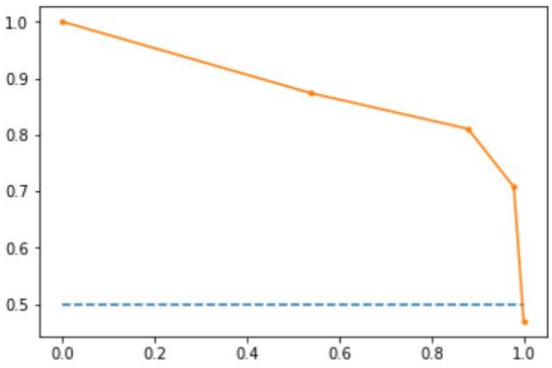
This concluded that the top two features, in terms of importance, were FICO score and Inquiry by creditors in the last 6 months. Therefore, these two were selected to train a model for logistic regression against those two terms.



The following features were used to gain an accuracy of 32%. This model was deemed unfit as it had a low accuracy and any method to fix the accuracy such as the number of iterations or features or changing the theta or learning rate were null and void in improving the accuracy of the model. Therefore, the implementation was not included in the final report as it was extremely and grossly a terrible model.

Precision Recall Curve:

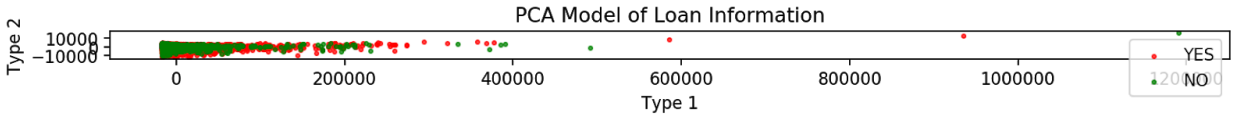
It takes the actual outcomes and the predicted probabilities and returns precision, recall, and threshold values. The graph below shows the precision and recall (orange line) for the threshold (blue line)



0.8435374149659864
0.8794715401570954
0.8270245226220369

Grid Search:
Grid search was used to find the closest estimation of the best parameters. The parameters[0.0001, 0.001, 0.01, 0.1, 1] were used and the final output determined that 1.0 was the best parameter.

PCA Model:
Another model implemented was a Principal Component Analysis. Starting again, this method was used, logically, to help speed up and improve relations concerning the fit of the data. What was implemented was a 2 component analysis to better visualize the data in terms of loans being not fully paid or fully paid. This can be seen below.



To better improve, we now apply this PCA to logistic regression and therefore receive an accuracy of 84%.

Decision Trees:
Interestingly enough, when Decision Trees were implemented, the accuracy was 100%. Although the model was not overfitting this could be attributed to a number of issues such as an error in the implementation or problems with the data. Nonetheless, this modified to a certain extend to fix the accuracy issue and the accuracy become close to 84% or 85%. However, this model will not be included in the final report as it gives way to other issues and cannot be the best model for the data.

Classification Report

	precision	recall	f1-score	support
0	0.81	0.84	0.82	153
1	0.82	0.80	0.81	147
avg / total	0.82	0.82	0.82	300

Confusion Matrix

```
[[1976  450]
 [ 338  110]]
```

Random Forest Classifier Trees:
This classifier had the best accuracy of over 99% and therefore did not need any improvements or modifications.
Classification Report

	precision	recall	f1-score	support
0	0.92	0.86	0.89	153
1	0.86	0.92	0.89	147
avg / total	0.89	0.89	0.89	300

Confusion Matrix

```
[[131  22]
 [ 12 135]]
```

According to the confusion matrix above, it predicted 131 instances of not full paid correctly of the class 1 and the model correctly predicted 135 instances of class 0 in not fully paid.

Gaussian Naives Bayes Classifier:
The last method implemented was naive bayes. The Accuracy presented was much lower than the other two classification trees used earlier and in the classification report, which included f1 score, recall, and precision, as report.
Classification Report

	precision	recall	f1-score	support
0	0.85	0.96	0.90	2426
1	0.32	0.10	0.15	448
avg / total	0.77	0.83	0.79	2874

Conclusion:

Overall, the implementations and methods described above performed exceedingly well and beyond expectation, with the exception of Decision Trees. Nonetheless, the implementation or model that performed the best was the Random Forest Classifier with an accuracy of around 99%. Therefore, no further action was needed on this part to improve the model. Although there are no parameters or coefficients that can be provided like an ordinary least squares or logistic regression model, it is safe to say that this model exceeds all expectations concerning its precision, recall, accuracy, and f1 score.