Loan Classification Chioma Nwuzi

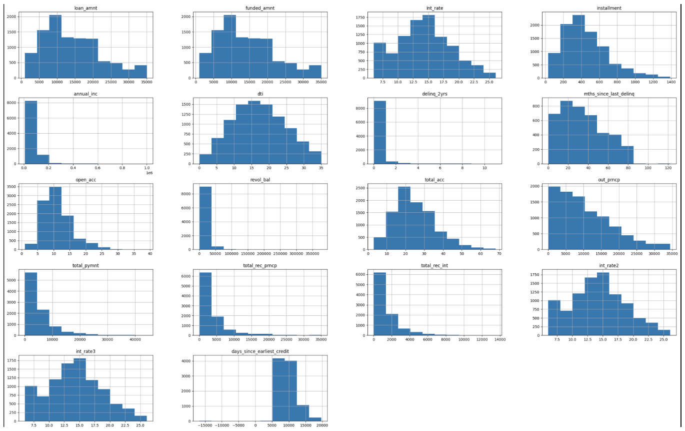
Business Question:

Given the personal data, credit data, and loan history of an applicant's loan, is there a model that will be able to predict whether a loan given to a borrower will end up being “good” or “bad”?

Methodology:

Data Cleanup

The first step of the methodology begins with adjusting the data in order to be suitable for modeling. Certain variables needed formatting transformations such as the variables describing interest rate that were transformed to integer variables and removed of extraneous characters, and the “earliest credit line” variable that was transformed to a “days since first credit” to make it usable for modeling and exploratory analysis.

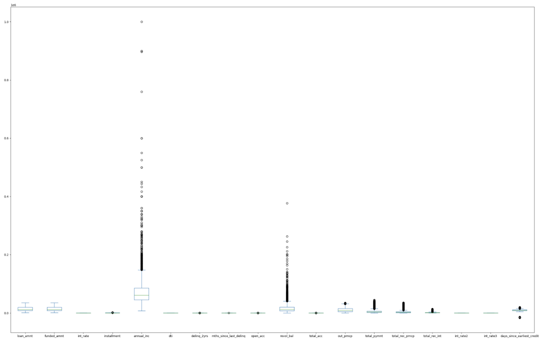


Exploratory Data Analysis

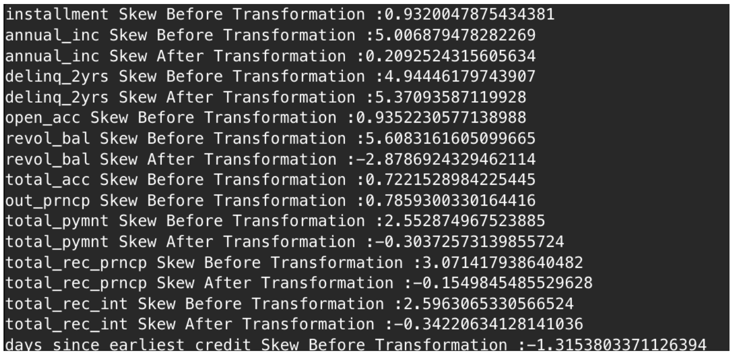
As part of exploratory data analysis, each continuous numeric variable histogram was plotted to assess the distribution of the table (left). This showed that some variables, particularly those that were around a borrower's total income or the total of the receivables from the loan did not have approximately normal distributions. Such variables were flagged for transformations.

Outlier Analysis

As the first step to understanding outliers, the box plots (right) of the continuous numeric variables were plotted to visualize the extent of the outliers. The “annual income” ,“revolving balance”, and variables around total payments had obvious outliers. The interquartile ranges in combination with the skew(left-below) of the variables are used as the final evaluation for outlier understanding. These variables were then

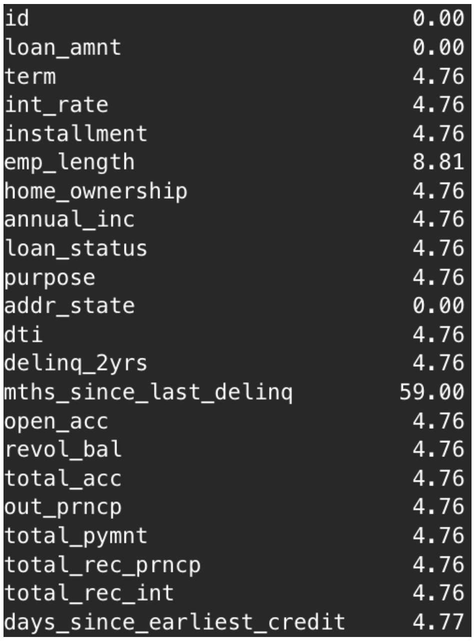


adjusted with a log_transformation to address the skew and the long-tailed outliers for these variables.



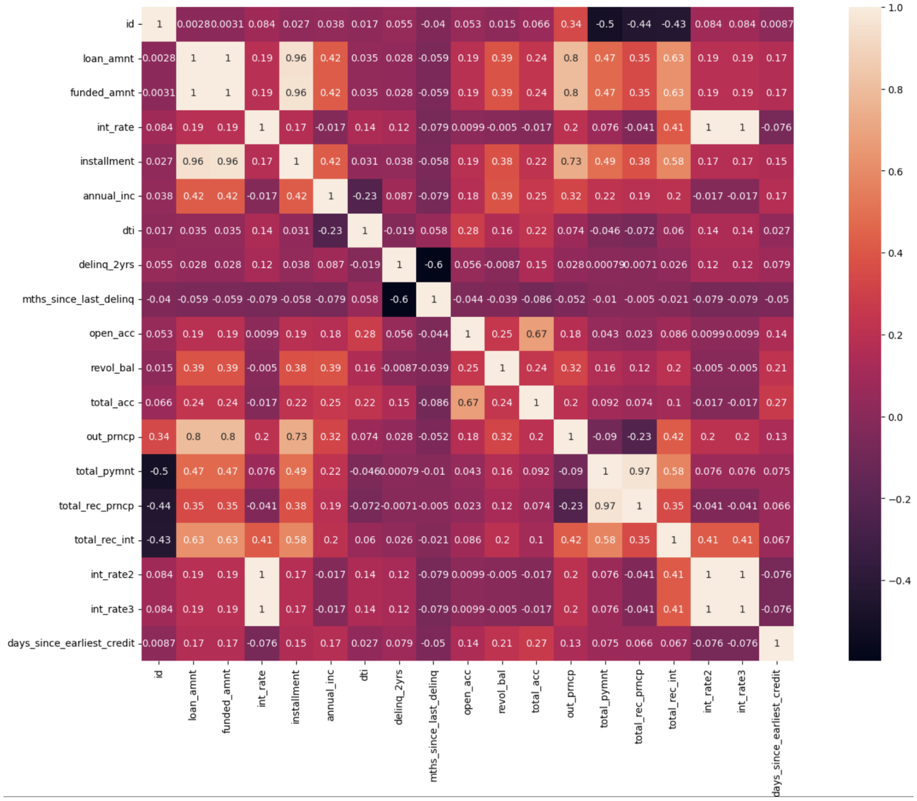
```
installment Skew Before Transformation :0.9320047875434381
annual_inc Skew Before Transformation :5.006879478282269
annual_inc Skew After Transformation :0.2092524315605634
delinq_2yrs Skew Before Transformation :4.94446179743907
delinq_2yrs Skew After Transformation :5.37093587119928
open_acc Skew Before Transformation :0.9352230577138988
revol_bal Skew Before Transformation :5.6083161605099665
revol_bal Skew After Transformation :-2.8786924329462114
total_acc Skew Before Transformation :0.7221528984225445
out_prncp Skew Before Transformation :0.7859300330164416
total_pymnt Skew Before Transformation :2.552874967523885
total_pymnt Skew After Transformation :-0.30372573139855724
total_rec_prncp Skew Before Transformation :3.071417938640482
total_rec_prncp Skew After Transformation :-0.1549845485529628
total_rec_int Skew Before Transformation :2.5963065330566524
total_rec_int Skew After Transformation :-0.34220634128141036
days_since_earliest_credit Skew Before Transformation :-1.3153803371126394
```

Data Wrangling



```
id                            0.00
loan_amnt                     0.00
term                          4.76
int_rate                      4.76
installment                   4.76
emp_length                    8.81
home_ownership                4.76
annual_inc                    4.76
loan_status                   4.76
purpose                       4.76
addr_state                    0.00
dti                           4.76
delinq_2yrs                   4.76
mths_since_last_delinq       59.00
open_acc                      4.76
revol_bal                     4.76
total_acc                     4.76
out_prncp                     4.76
total_pymnt                   4.76
total_rec_prncp               4.76
total_rec_int                 4.76
days_since_earliest_credit    4.77
```

The variables that included no values for any observations were removed as well as the observations that did not have any data outside of the borrower identification number and the amount of loan (22 missing variables of 25).This left 95% of the original observations from the dataset. The variable describing months since last delinquency was also removed since about 59% of this data was missing (right-above). This could either be from true missing data, or if the borrower has never been delinquent, then the correct value should be 0. But looking at the Pearson correlation coefficient for the two variables does not reveal a correlation (left). More data may be able to reveal the correlation but for this analysis the variable was removed. After these measures, the observations that still had missing variables had the values either imputed for the mode for the categorical/ordinal variables or the median for continuous variables. The last step for the data wrangling was the correlation analysis which showed that two additional variables for loan interest were perfectly correlated and therefore two of the three variables were removed.



	id	loan_amnt	funded_amnt	int_rate	installment	annual_inc	dti	delinq_2yrs	mths_since_last_delinq	open_acc	revol_bal	total_acc	out_prncp	total_pymnt	total_rec_prncp	total_rec_int	int_rate2	int_rate3	days_since_earliest_credit
id	1	0.0028	0.0031	0.084	0.027	0.038	0.017	0.055	-0.04	0.053	0.015	0.066	0.34	-0.5	-0.44	-0.43	0.084	0.084	0.0087
loan_amnt	0.0028	1	1	0.19	0.96	0.42	0.035	0.028	-0.059	0.19	0.39	0.24	0.8	0.47	0.35	0.63	0.19	0.19	0.17
funded_amnt	0.0031	1	1	0.19	0.96	0.42	0.035	0.028	-0.059	0.19	0.39	0.24	0.8	0.47	0.35	0.63	0.19	0.19	0.17
int_rate	0.084	0.19	0.19	1	0.17	-0.017	0.14	0.12	-0.079	0.0099	-0.005	-0.017	0.2	0.076	-0.041	0.41	1	1	-0.076
installment	0.027	0.96	0.96	0.17	1	0.42	0.031	0.038	-0.058	0.19	0.38	0.22	0.73	0.49	0.38	0.58	0.17	0.17	0.15
annual_inc	0.038	0.42	0.42	-0.017	0.42	1	-0.23	0.087	-0.079	0.18	0.39	0.25	0.32	0.22	0.19	0.2	-0.017	-0.017	0.17
dti	0.017	0.035	0.035	0.14	0.031	-0.23	1	-0.019	0.058	0.28	0.16	0.22	0.074	-0.046	-0.072	0.06	0.14	0.14	0.027
delinq_2yrs	0.055	0.028	0.028	0.12	0.038	0.087	-0.019	1	-0.6	0.056	-0.0087	0.15	0.028	0.00079	-0.0071	0.026	0.12	0.12	0.079
mths_since_last_delinq	-0.04	-0.059	-0.059	-0.079	-0.058	-0.079	0.058	-0.6	1	-0.044	-0.039	-0.086	-0.052	-0.01	-0.005	-0.021	-0.079	-0.079	-0.05
open_acc	0.053	0.19	0.19	0.0099	0.19	0.18	0.28	0.056	-0.044	1	0.25	0.67	0.18	0.043	0.023	0.086	0.0099	0.0099	0.14
revol_bal	0.015	0.39	0.39	-0.005	0.38	0.39	0.16	-0.0087	-0.039	0.25	1	0.24	0.32	0.16	0.12	0.2	-0.005	-0.005	0.21
total_acc	0.066	0.24	0.24	-0.017	0.22	0.25	0.22	0.15	-0.086	0.67	0.24	1	0.2	0.092	0.074	0.1	-0.017	-0.017	0.27
out_prncp	0.34	0.8	0.8	0.2	0.73	0.32	0.074	0.028	-0.052	0.18	0.32	0.2	1	-0.09	-0.23	0.42	0.2	0.2	0.13
total_pymnt	-0.5	0.47	0.47	0.076	0.49	0.22	-0.046	0.00079	-0.01	0.043	0.16	0.092	-0.09	1	0.97	0.58	0.076	0.076	0.075
total_rec_prncp	-0.44	0.35	0.35	-0.041	0.38	0.19	-0.072	-0.0071	-0.005	0.023	0.12	0.074	-0.23	0.97	1	0.35	-0.041	-0.041	0.066
total_rec_int	-0.43	0.63	0.63	0.41	0.58	0.2	0.06	0.026	-0.021	0.086	0.2	0.1	0.42	0.58	0.35	1	0.41	0.41	0.067
int_rate2	0.084	0.19	0.19	1	0.17	-0.017	0.14	0.12	-0.079	0.0099	-0.005	-0.017	0.2	0.076	-0.041	0.41	1	1	-0.076
int_rate3	0.084	0.19	0.19	1	0.17	-0.017	0.14	0.12	-0.079	0.0099	-0.005	-0.017	0.2	0.076	-0.041	0.41	1	1	-0.076
days_since_earliest_credit	0.0087	0.17	0.17	-0.076	0.15	0.17	0.027	0.079	-0.05	0.14	0.21	0.27	0.13	0.075	0.066	0.067	-0.076	-0.076	1

Feature Transformation

The categorical and ordered variables were encoded and the continuous numeric variables were scaled. The target variable that was created to proxy for a “good” and “bad” came from the “loan status” variable. All fully paid loans and current loans were coded as “good” and all other statuses were coded as “bad”. This method revealed that the data was very imbalanced with the majority of loans being flagged as “good”

Data Split and Sampling

The data was split into 70% training data and 30% testing data. Along with this split, a testing and training split that is downsampled from the majority class (“good” loans) was created. Both splits were tested on the models and the difference in evaluation metrics compared.

Modeling

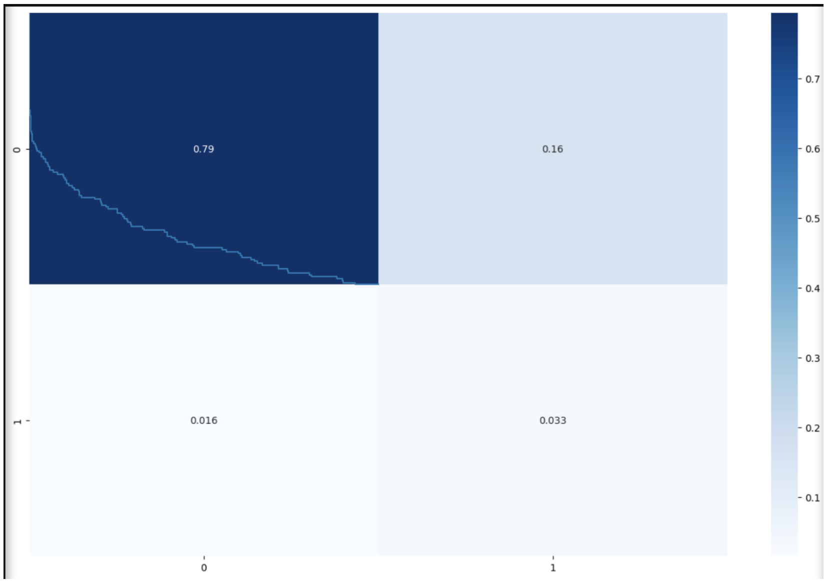
The three models tested were a logistic regression model, random forest model, and xg-boost model. Before modeling, a grid search to find the most optimal hyperparameters for each model is performed. This includes the type of penalty best used to prevent the model from overfitting too heavily to the training data, the maximum number of splits each decision tree can make in the random forest, and the minimum weight for the boosting model to stop splitting once a certain level of homogeneity is reached within a node. The best combination of hyper-parameters were determined by the best produced auc score. These values are then fit to the models

Results:

The measures used to assess the performance of the models were accuracy, recall, f1-score, and ROC Score. All of these metrics are compared to ensure the models are evaluated on more than one metric where certain metrics are biased towards certain types of data.

For example, the accuracy measure in the “standard” sampled logistic regression is quite high at .826. Since accuracy measures how well the model was able to predict overall, even if the prediction is biased towards one level, the accuracy may be inflated, which is what is shown for this model. Recall is a more important metric in this case since in the business case, it would be better to minimize the number of loans the model predicted to be good but were actually bad - false negative than it would be to have a false positive that the model predicted a loan will be bad but was good.

Logistic Regression Confusion Matrix (below)



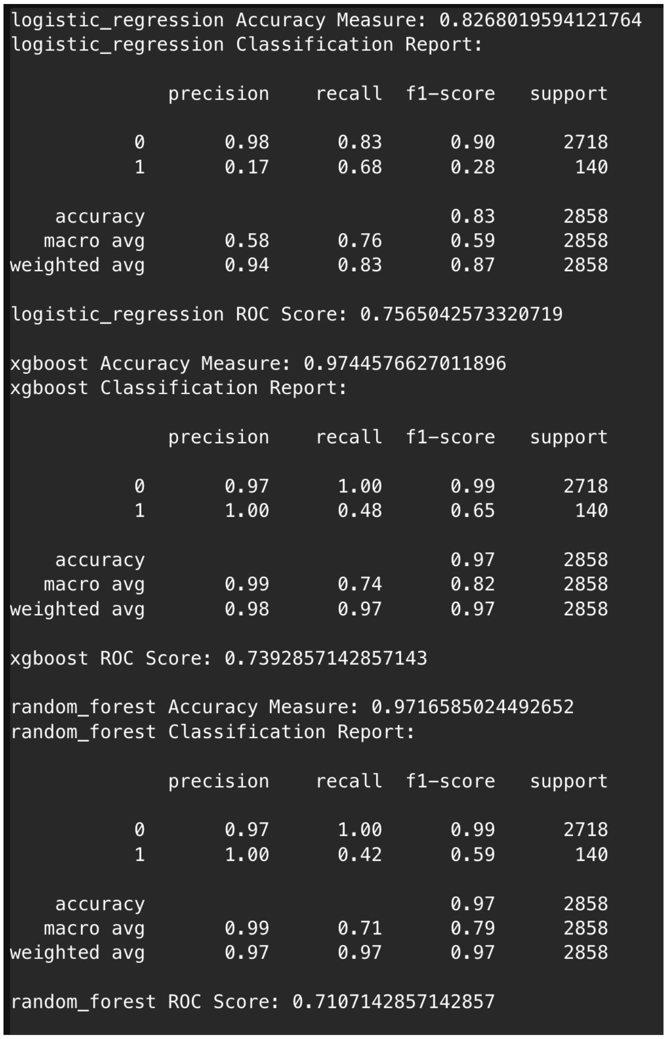
79% of the bad loans were predicted correctly

3.3% of the good loans were predicted correctly

16% of the loans were incorrectly predicted to be bad loans

1.6% of loans were incorrectly predicted to be good loans

Here the recall for the “bad” loans was .83 which does relatively well. On the other hand, the ability for the model to predict good loans is quite a bit lower at .68 (left). This may show that the indicator for a good loan needs to be adjusted, further data collection is necessary, or another model would be a better fit. The f1 score for the good loans was significantly higher than that of the bad loans further showing the bias of the model towards the bad loans. The ROC score for the overall model is .75 which indicates room for improvement in the overall model.



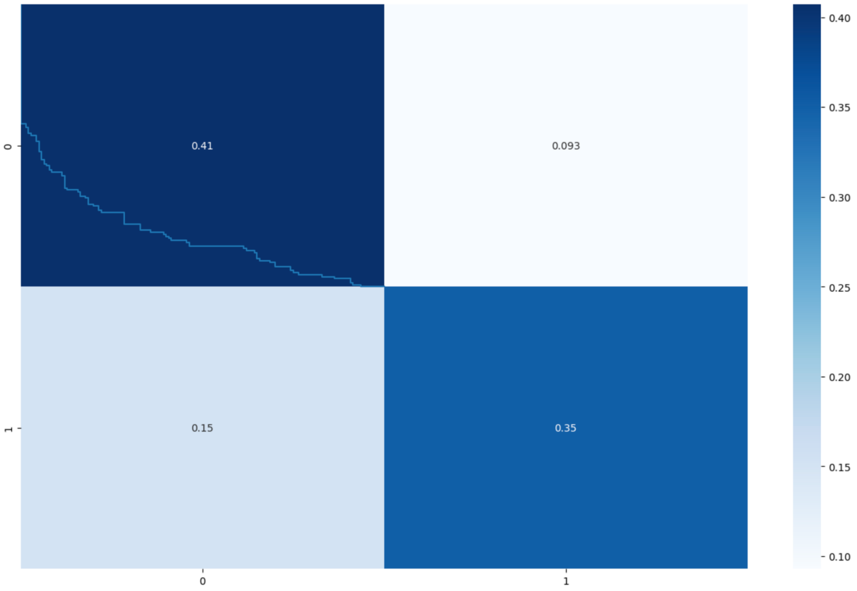
```
logistic_regression Accuracy Measure: 0.8268019594121764
logistic_regression Classification Report:

              precision    recall  f1-score   support

           0       0.98      0.83      0.90      2718
           1       0.17      0.68      0.28       140

    accuracy                           0.83      2858
   macro avg       0.58      0.76      0.59      2858
weighted avg       0.94      0.83      0.87      2858

logistic_regression ROC Score: 0.7565042573320719

xgboost Accuracy Measure: 0.9744576627011896
xgboost Classification Report:

              precision    recall  f1-score   support

           0       0.97      1.00      0.99      2718
           1       1.00      0.48      0.65       140

    accuracy                           0.97      2858
   macro avg       0.99      0.74      0.82      2858
weighted avg       0.98      0.97      0.97      2858

xgboost ROC Score: 0.7392857142857143

random_forest Accuracy Measure: 0.9716585024492652
random_forest Classification Report:

              precision    recall  f1-score   support

           0       0.97      1.00      0.99      2718
           1       1.00      0.42      0.59       140

    accuracy                           0.97      2858
   macro avg       0.99      0.71      0.79      2858
weighted avg       0.97      0.97      0.97      2858

random_forest ROC Score: 0.7107142857142857
```

The xg-boost model was shown to be very biased towards the bad loans as the model was able to predict the bad loans at about a 99% level of the test data. The recall of the good loans comparatively was almost half these levels at about .48. . The ROC score for this model was .739 indicating that although the recall level for the bad loans was better for this model, the logistic regression model overall is better.

The random model showed similar results to that of the random forest model with recall of the bad loans about .99 while the recall of the good loans almost half that at .42. The ROC score of this model was also the lowest at .71.

Down-sampled models:

For the down-sampled models, the disparities that were seen in the evaluation metrics due to the imbalanced nature of the response are no longer seen. For the logistic regression model, the recall for the "bad" and "good" loans are .81, and .71 respectively. The accuracy of the logistic regression down-sampled model has more interpretability since there is not a bias towards one level of the response in the sampling. The accuracy and the ROC curve of the logistic regression model at both .75 which indicates improvements could be made in terms of tuning the hyperparameters.



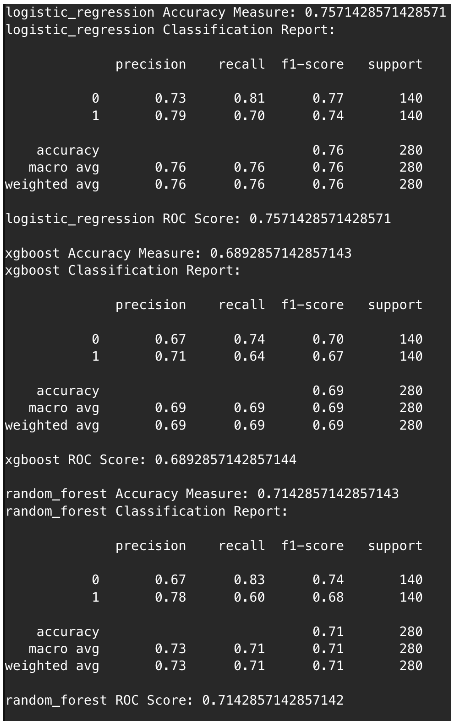
Logistic Regression Down-Sampled Confusion Matrix (above)

41% of the bad loans were predicted correctly

35% of the good loans were predicted correctly

9.3% of the loans were incorrectly predicted to be bad loans

15% of loans were incorrectly predicted to be good loans



```
logistic_regression Accuracy Measure: 0.7571428571428571
logistic_regression Classification Report:

              precision    recall  f1-score   support

           0       0.73      0.81      0.77       140
           1       0.79      0.70      0.74       140

    accuracy                           0.76       280
   macro avg       0.76      0.76      0.76       280
weighted avg       0.76      0.76      0.76       280

logistic_regression ROC Score: 0.7571428571428571

xgboost Accuracy Measure: 0.6892857142857143
xgboost Classification Report:

              precision    recall  f1-score   support

           0       0.67      0.74      0.70       140
           1       0.71      0.64      0.67       140

    accuracy                           0.69       280
   macro avg       0.69      0.69      0.69       280
weighted avg       0.69      0.69      0.69       280

xgboost ROC Score: 0.6892857142857144

random_forest Accuracy Measure: 0.7142857142857143
random_forest Classification Report:

              precision    recall  f1-score   support

           0       0.67      0.83      0.74       140
           1       0.78      0.60      0.68       140

    accuracy                           0.71       280
   macro avg       0.73      0.71      0.71       280
weighted avg       0.73      0.71      0.71       280

random_forest ROC Score: 0.7142857142857142
```

The xg-boost model and the random forest model both performed poorly in comparison to the logistic regression model(right). The bias that was shown towards the bad loans in the "standard" sampled model results is not present in the down-sampled model as the recall for the xg-boost model and the random forest at .64 and .60 respectively. But these metrics and the ROC score remain lower than that of the logistic regression.

Conclusion: The conclusion for the “standard” sampling model evaluation is that the logistic regression model is the best for this data especially for optimizing for a model that can better correctly predict a “bad loan”

Limitations and Future Improvements:

In the future, if more data is available, instead of current loans being an indicator for a good loan, more loans that were fully paid can be used to offset the imbalance from the good loans to bad loans without the loss of information.

Although the down-sampled dataset showed improvements to the evaluations of the “standard” sampled data, there is still room for model improvement. Particularly where this would best be seen is which an increased sample size.