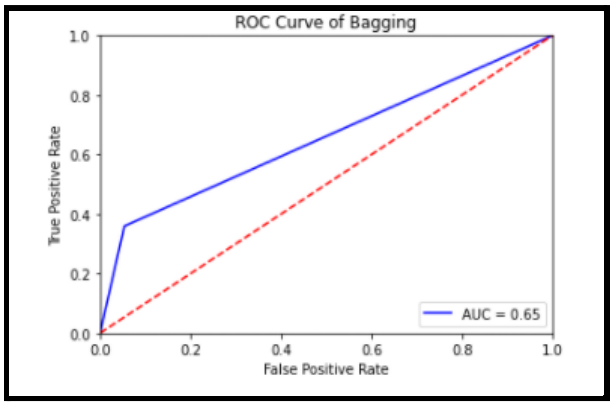
Credit Default

In our design, we originally considered many potential predictive models available for classification. We calculated and observed the KFolds cross validation of Logistic Regression, Naive Bayes, Random Forest, Support Vector, kNN, and Decision Tree. After considering the accuracy of certain models, we attempted to build ensemble models based off of max voting, bagging, and random forest. While we received comparable results with bagging and random forest, we ultimately decided to move forward with the bagging method.

We cleaned up the data by finding variables that were encoded in a confusing way, like repeated categories of "other" for the education and marriage status. After running some quick models, we found that our data responded well to bagging methods like random forest, so we decided to run those more in depth. A Bagging Classifier model slightly outperformed the Random Forest model, so we are including the ROC/AUC curves and results for that model here.




```
Precision: 0.7467673903300256
Recall: 0.6529179071741349
F1: 0.6966963252519573
              precision    recall  f1-score   support

           0       0.84      0.95      0.89      4684
           1       0.65      0.36      0.46      1316

    accuracy                           0.82      6000
   macro avg       0.75      0.65      0.68      6000
weighted avg       0.80      0.82      0.80      6000
```

Considering evaluation metrics, we took on the mindset of a credit agency. We discussed which outcomes we are trying to minimize or maximize. We concluded that minimizing the number of false negatives (always catching a true default, even if they did not default) is most important.

Our bagging ensemble method provided the best results, based on the metric that is associated with minimizing the number of false negatives (we said they wouldn't default, but then they did). Precision lets us know how well our model correctly predicts true positives (they defaulted when we predicted they would) over the number of all predicted positives. Precision is good to use when the price of a false positive is high. Recall lets us know how well our model predicts true positives over the number of actual positive users. This is good when the price of a false negative is high. F1 score is a balance between the two scores, and is a better reflection of actual model accuracy.

Since we want to reduce false negatives, we will use the recall score as our metric for effectiveness. Random Forest resulted in a score of 0.63 and 0.81 for average and weighted average recall, while Bagging Classifier resulted in 0.65 and 0.82. Since our bagging was more

effective, we would recommend that one be used for future models. For future research, we would create more engineered features based on the demographic information, and attempt to understand the payment and billing patterns more, perhaps with a time series.