Customer Default Identification Report

Problem:
An increase in customer default rates is bad for Credit One since its business is approving customers for loans in the first place. This is likely to result in the loss of Credit One's business customers.

Results

1. We can identify customers who will default 66% of the time.

Figure 1: Actual Results vs Model

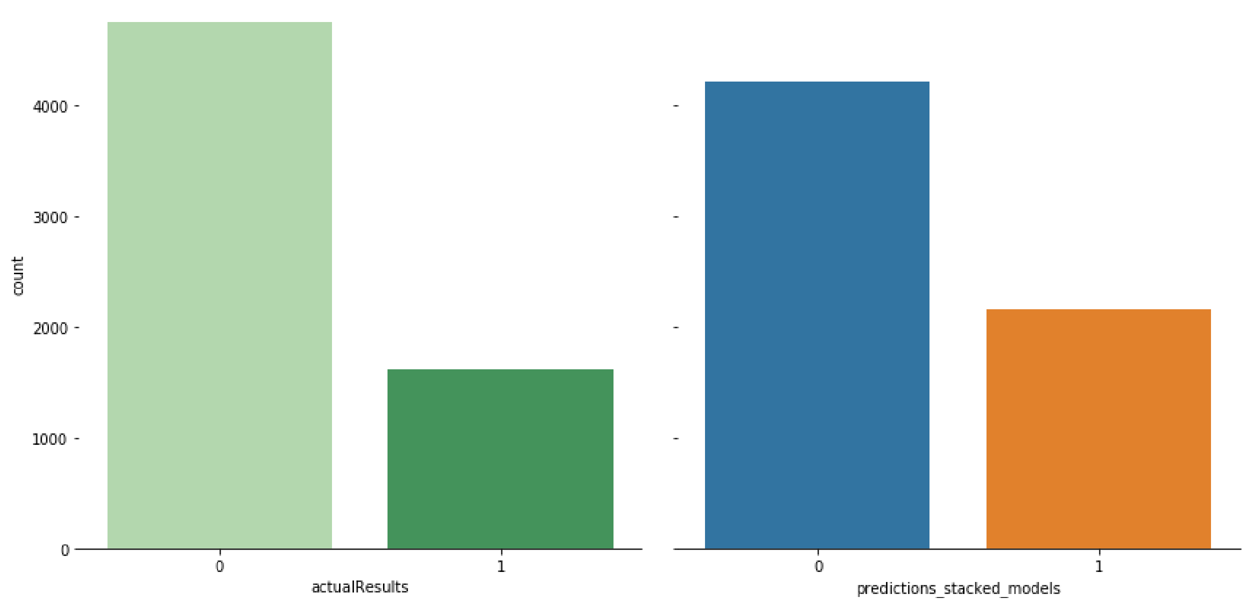
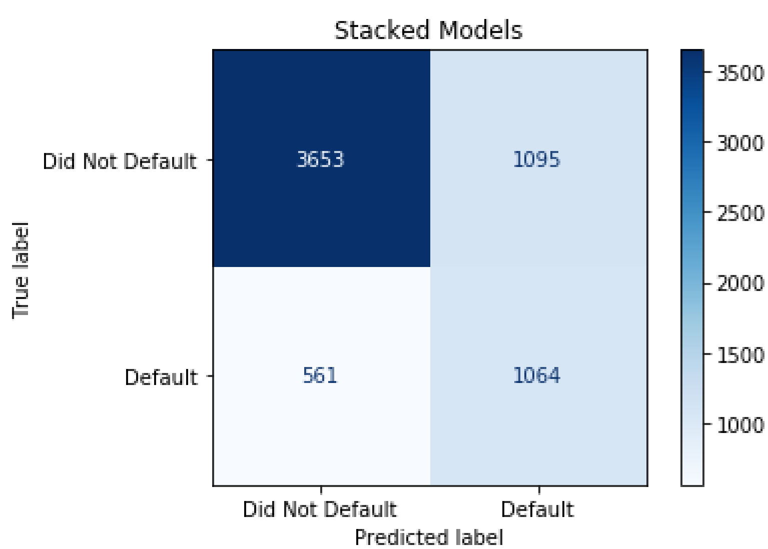


Figure 2: Confusion Matrix

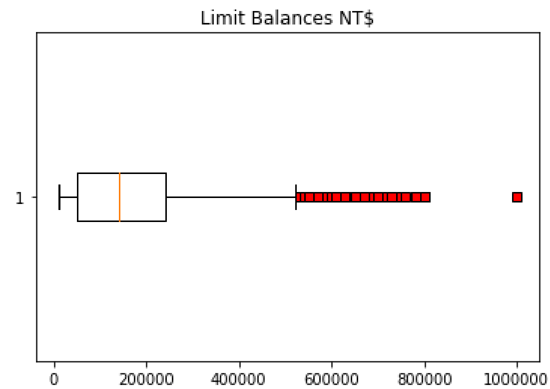


2. Our model overestimates the customers who willd default. It fits the term "Better Safe than Sorry". However, out of the customers who did not default, our model predicted that

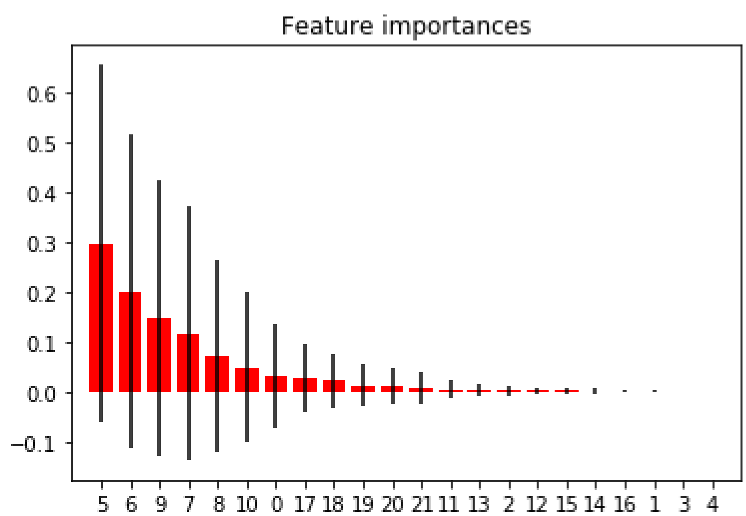
23% of them defaulted. This can be a significant loss of revenue. However, our biggest concern right now is approving customers who end up defaulting. We can find them 66% of the time. This means that about 33% of our customers that we approve based on this model will default.

3. We removed outliers. These are customers with limit amounts greater than $525,000, and with monthly payments generally above $9,500.

Figure 3: Boxplot of Limit distribution. Most customers have balances below $250,000.



4. Statistically significant attributes in the data:
    a. The customer's pay status in the last 3 months
    b. The amount of the first bill
    c. Limit balance



Feature ranking:	Feature Number:	Feature Name	Percent Used
1	5	pay1	21.12%
2	6	pay2	9.15%
3	7	pay3	6.85%

4	11	bill1	5.27%
5	0	limit	4.74%
6	17	paid1	4.70%
7	18	paid2	4.48%
8	8	pay4	4.46%
9	12	bill2	4.18%
10	19	paid3	3.81%
11	20	paid4	3.68%
12	13	bill3	3.55%
13	14	bill4	3.37%
14	15	bill5	3.24%
15	16	bill6	3.23%
16	10	pay6	3.21%
17	4	age	3.12%
18	9	pay5	2.89%
19	21	paid5	2.65%
20	2	edu	0.96%
21	3	marriage	0.71%
22	1	sex	0.62%

It is interesting that age, education, marriage, and sex were not the most important features. Rather, a customer's payment history and loan amount was. It is crucial to collect customer payment history in order to use this model to approve a customer.

For future studies, we recommend preparing more customer data that reflects customer behavior over time. This may strengthen the model since the most important features were those that described customer behavioral patterns (payment status, and the amount that a customer chose to pay).

Model Performance Metrics

Gradient Tree Boosting

Gradient Tree Boosting				
	F1	**Recall**	**Precision**	**Accuracy**
Out of the Box:	0.46220	0.35408	0.66539	0.82056
Multicollinearity:	0.46220	0.35408	0.66539	0.82056
RFE:	N/A	N/A	N/A	N/A
Outliers:	0.49413	0.38831	0.67922	0.79727
Oversampling	0.52643	0.61990	0.45745	0.75711
Undersampling	0.52537	0.62602	0.45260	0.75367
Outliers, Undersampling	0.55867	0.63877	0.49641	0.74266
Outliers, Undersampling, Bagged	0.56186	0.64554	0.49739	0.74329

Ada Boost

Ada Boost				
	F1	**Recall**	**Precision**	**Accuracy**
Out of the Box:	0.43496	0.32245	0.66808	0.81756
Multicollinearity:	0.43496	0.32245	0.66808	0.81756
RFE:	N/A	N/A	N/A	N/A
Outliers:	0.44748	0.33292	0.68222	0.79037
Oversampling	0.51784	0.60714	0.45144	0.75378
Undersampling	0.52087	0.61122	0.45379	0.75511
Outliers, Undersampling	0.54437	0.61723	0.48689	0.73654

Random Forest

Random Forest				
	F1	**Recall**	**Precision**	**Accuracy**
Out of the Box:	0.31292	0.20204	0.69352	0.80678
Multicollinearity:	0.30616	0.19643	0.69369	0.80611
RFE:	0.28443	0.17806	0.70648	0.80489
Outliers:	0.37562	0.256	0.70508	0.78299
Oversampling	0.52106	0.57449	0.47671	0.77000
Undersampling	0.52316	0.57347	0.48096	0.77233
Outliers,	0.54190	0.59692	0.49616	0.74266

Oversampling				
Outliers, Undersampling	0.54196	0.59015	0.50104	0.74565

Random Forest parameter tuning:

{'n_estimators': 200,
'min_samples_split': 35,
'min_samples_leaf': 4,
'max_features': 'auto',
'max_depth': 20,
'bootstrap': False}

K-Nearest Neighbor

K-Nearest Neighbor				
	F1	**Recall**	**Precision**	**Accuracy**
Out of the Box:	0.27176	0.22704	0.3384	0.735
Multicollinearity	0.26833	0.21939	0.34538	0.73944
RFE:	N/A	N/A	N/A	N/A
Outliers:	0.31157	0.26277	0.38262	0.70391
Oversampling	0.34831	0.46071	0.28000	0.62456
Undersampling	0.35044	0.46684	0.28050	0.62311
Outliers, Undersampling	0.40617	0.58338	0.31153	0.56504

Guassian Naive Bayes

Guassian Naive Bayes				
	F1	**Recall**	**Precision**	**Accuracy**
Out of the Box	0.38326	0.86429	0.24622	0.39422
Multicollinearity	0.38323	0.89541	0.24378	0.37233
RFE	N/A	N/A	N/A	N/A
Outliers	0.42385	0.54031	0.34869	0.62545
Oversampling	0.37503	0.93010	0.23486	0.32489
Undersampling	0.37488	0.93214	0.23462	0.32300
Outliers, Undersampling	0.42746	0.76246	0.29698	0.47921

Least Squares Support Vector Machine

Least Squares Support Vector Machine				
	F1	**Recall**	**Precision**	**Accuracy**
Out of the Box	0.08544	0.05	0.29341	0.76689

All Models, after Outliers and Undersampling				
	F1	**Recall**	**Precision**	**Accuracy**
KNN	0.40617	0.58338	0.31153	0.56504
Random Forest	0.54196	0.59015	0.50104	0.74565
Ada Boost	0.54437	0.61723	0.48689	0.73654
Gradient TB	0.55867	0.63877	0.49641	0.74266
Gaussian NB	0.42746	0.76246	0.29698	0.47921
stacked_models	0.55122	0.62585	0.49249	0.74015
stacked_models_tunedRF	0.55921	0.65969	0.48529	0.73482
stacked_rf_tuned_gtb_bagged	0.56237	0.65600	0.49282	0.74015