For this assignment we looked at loan data to evaluate 6 models and their ability to assess credit risk. We cleaned the data set down to 68,817 individual loans. We used a total of 86 variables and divided all the scrubbed loans into 2 risk categories. The source data set had 68,470 loans in the “low-risk” target and 347 loans in the “high-risk” target. Due to the imbalanced nature of the dataset, with 200 times more “low-risk” loans than “high-risk” loans, particular attention was made to models that either extrapolate or synthesize target data points.

Assumptions:

Credit Risk is never an exact science because large pools of people are affected by life circumstances differently. The prediction goal is to identify “high-risk” targets. There will always be some low-risk targets that, over time, become problem loans. Such is the nature of the business. Our goal in this exercise is to identify loans that are presently at a higher risk of default. Precision is the one metric we are most interested in because we want to avoid classifying good loans as high-risk. Because the random likelihood of a loan being “high-risk” is 0.5%, all things being equal, we would rather accept the additional risk of a small number than misclassify a much larger number of performing loans.

1: Balanced Random Forest Classifier

True Positives: 68

True Negatives: 15,355

False Positives: 1,749

False Negatives: 33

Balanced Accuracy Score: 0.78

The false positives are unacceptably high in this model and the Classification Report (Imbalanced) shows this with a 0.037 Precision score.

Recall (0.67) is acceptable, but not great. 33 loans returned as false negatives, and given the low number of true positives (68) failing to identify one third of the high-risk loans makes the model less than ideal.

The Balanced Accuracy Score was enhanced by the sheer number of true negatives, which somewhat muted the large number of 1,749 false positives.

2: Easy Ensemble AdaBoost Classifier

True Positives: 93

True Negatives: 16,109

False Positives: 995

False Negatives: 8

Balanced Accuracy Score: 0.93

Clearly the winner in this competition. While 995 false positives is still less than ideal, it is by far the best result of the 6 models tested.  
False positives are high but better than the previous model. The Classification Report (Imbalanced) shows this with a 0.09 Precision score.

Recall (0.92) is commendable, with only 8 false negatives.

The Balanced Accuracy Score reflects a significant reduction of true negatives and a consistency across specificity and recall.

3: Naïve Random Oversampling

True Positives: 70

True Negatives: 10,393

False Positives: 6,711

False Negatives: 31

Balanced Accuracy Score: 0.65

False positives are unreasonably high and the Classification Report (Imbalanced) shows this with a 0.01 Precision score.

Recall (0.69) is acceptable, but not great, with 31 false negatives.

The Balanced Accuracy Score reflects some consistency across specificity and recall, but this is disappointing given the excessive volume of false positives.

4: SMOTE Oversampling

True Positives: 64

True Negatives: 11,813

False Positives: 5,291

False Negatives: 37

Balanced Accuracy Score: 0.66

The false positives are unacceptably high in this model and the Classification Report (Imbalanced) shows this with a 0.012 Precision score.

Recall (0.63) is acceptable, but not great. 37 loans returned as false negatives, and given the low number of true positives (64) failing to identify over one third of the high-risk loans makes the model less than ideal.

The Balanced Accuracy Score reflects some consistency across specificity and recall, but this is disappointing given the excessive volume of false positives. SMOTE was marginally better than Naïve. While their BA scores were nearly identical, and SMOTE identified fewer false positives, however it did have a 20% better Precision (0.010 vs 0.012), and Precision is more important than other metrics.

5: Undersampling

True Positives: 68

True Negatives: 7,102

False Positives: 10,002

False Negatives: 33

Balanced Accuracy Score: 0.54

False positives are “ridiculously” high and the Classification Report (Imbalanced) shows this with a 0.006 Precision score.

Recall (0.67) is acceptable, but not great, with 33 false negatives.

The Balanced Accuracy Score reflects the worst specificity yet (0.42).

6: SMOTEENN

True Positives: 79

True Negatives: 9,795

False Positives: 7,309

False Negatives: 22

Balanced Accuracy Score: 0.67

False positives are unreasonably high and the Classification Report (Imbalanced) shows this with a 0.01 Precision score.

Recall (0.78) is actually second best of the 6 models, with only 22 false negatives.

The Balanced Accuracy Score is saved by the higher recall and falls in line with 4 of the other models.

You can see the advantage a combined over and under-sampling can have, however for this data set we simply cannot live with such a high false positive volume.

Summary:

The EEC model was the clear winner. However, it did misidentify 995 false positives. That is 995 loans that could, in theory, be rejected in order to avoid 93 legitimately bad loans. Is that a sound business decision? That is above my pay grade. However, this model has clear advantages over the other 5 models, with 4 of them being particularly horrible at predicting datasets where the goal is avoiding false positives.