1. **Overview of the Credit Risk Analysis:** Explain the purpose of this analysis.

Predictions of outcome and analysis of testing methods are used more often that you think when it comes to everyday activities. How does your email know which emails are classified as spam and are there any instances where emails are filtered to the wrong folder? When testing for illness or disease, are diagnosis’s correct and what method has a higher sensitivity rating so patients don’t go undiagnosed? There are many statistical methods that can be taken into account and finding the most suitable method could be more detrimental than you would think. In this case, we are assessing multiple ways to determine whether a person is high or low risk for a loan.

1. **Results:**

Once we have been provided with a data set, we can begin to explore various methods of machine learning. In the first three methods we will use imbalanced-learn and scikit-learn libraries to resample the data. The last two models will use imblearn.ensemble, which includes BalancedRandomForestClassifier and EasyEnsembleClassifier. Our analysis will provide accuracy scores and classification reports that will determine the best method of prediction for this situation.

Our first step is to read the file and perform basic data cleaning such as dropping null values, converting data types and formats, and selecting the columns that you would like incorporated into the analysis. Our completed set of data contains 68,470 low risk loans, and 347 high risk loans. Once we run the train\_test\_split function we are ready to start our different methods of analysis.

Oversampling:

* + - Naïve Random Oversampling is a method that takes the group of less likely data, in this case the high-risk loans, and randomly samples from the set until there is an equal amount of low and high-risk samples. The accuracy of this sampling comes to approximately 66%. The confusion matrix shows that this testing method produced 73 actual high-risk loans, 28 false low-risk loans, 6,840 false high-risk loans and 10,264 actual low-risk loans. The imbalanced report then shows high-risk predictions had 1% precision, 72% recall, and a F1 score of 2%. Low-risk data came in with 100% precision, 60% recall and a F1 score of 72%.
    - SMOTE (Synthetic Minority Oversampling Technique) is a method that generates data based on minority values and the data that is closest to these points. The accuracy of this sampling comes to approximately 66%, similarly to the naïve method. The confusion matrix shows that this testing method produced 63 actual high-risk loans, 38 false low-risk loans, 5,260 false high-risk loans and 11,844 actual low-risk loans. The imbalanced report then shows high-risk predictions had 1% precision, 62% recall, and a F1 score of 2%. Low-risk data came in with 100% precision, 69% recall and a F1 score of 82%.

Undersampling:

* + - On the opposite spectrum of sampling would be undersampling. This method will take random samples of the more frequent data the cut the size down to be similar to that of the smaller cluster of data. The number of samples per cluster were 246 in this algorithm as opposed to 51,366 in the previous oversampling methods. Once again, the accuracy of this sampling comes to approximately 66%. The confusion matrix shows that this testing method produced 70 actual high-risk loans, 31 false low-risk loans, 10,324 false high-risk loans and 6,780 actual low-risk loans. The imbalanced report then shows high-risk predictions had 1% precision, 69% recall, and a F1 score of 1%. Low-risk data came in with 100% precision, 40% recall and a F1 score of 57%.

Combination (Over and Under) Sampling

* + - Combining aspects from both over and under sampling is called the SMOTEEN(Synthetic Minority Oversampling Technique and Edited Nearest Neighbors) algorithm. Ultimately, the sampling comes to 68,458 high-risk points and 62,022 low-risk points. the accuracy of this sampling comes to a lower 54%. The confusion matrix shows that this testing method produced 78 actual high-risk loans, 23 false low-risk loans, 7,357 false high-risk loans and 9,747 actual low-risk loans. The imbalanced report then shows high-risk predictions had 1% precision, 77% recall, and a F1 score of 2%. Low-risk data came in with 100% precision, 57% recall and a F1 score of 73%.

Ensemble Learners

* + - Balanced Random Forest Classifier
    - Easy Ensemble AdaBoost Classifier

1. **Summary:**