**Module 17 Challenge: Credit Risk Analysis: (Supervised Machine Learning)**

**Overview:**

This challenge requested that we use “Supervised Machine Learning” techniques to determine the best model to evaluate “Credit Risk” on new loans. Using a credit card credit dataset from LendingClub, a peer-to-peer lending services company we created several models using both Over & Under sampling algorithms as well as a combined Over/Under algorithm. After training the models we tested the models and measured them for accuracy, created “confusion matrix and imbalanced classification reports to evaluate each model and finalize our conclusion.

We decided on the following popular Machine Learning methods to create our models:

Random\_Over\_Sampler

SMOTE

Cluster\_Centroids

SMOTEEN

Blanced\_Random\_Forest\_Classifier

Easy\_Ensemble\_AdaBoost\_Classifier

**Analysis Results:**

### Deliverable 1

### **Random Over Sampling - Naive RandomOverSampler**

Text

Description automatically generated

* An accuracy score of the **RandomOversampler** model is **0.6438627638**
* The Confusion Matrix generated from this model

Table

Description automatically generated with low confidence

* The Imbalanced Classification Report from this model:

Table

Description automatically generated

#### **Over sampling – SMOTE**

Graphical user interface, text

Description automatically generated

* An accuracy score of the **SMOT** model is **0.6438327638**
* The Confusion Matrix generated from this model:

A picture containing chat or text message

Description automatically generated

* The Imbalanced Classification Report from this model:

Table

Description automatically generated

#### **Under Sampling - ClusterCentroids**

Text

Description automatically generated

An accuracy score of the **ClusterCentroids** model is **0.6438627638**

* The Confusion Matrix generated from this model

Table

Description automatically generated

* The Imbalanced Classification Report from this model:

Table

Description automatically generated

### Deliverable 2

### **Combination of (Over and Under) Sampling – SMOTEEN**

Text

Description automatically generated

* An accuracy score of the **SMOTEEN** model is **0.5441781794**
* The Confusion Matrix generated from this model:

Table

Description automatically generated with medium confidence

* The Imbalanced Classification Report from this model

Graphical user interface

Description automatically generated with medium confidence

### Deliverable 3

### **Ensemble Learners - Balanced Random Forest Classifier**

Graphical user interface, text, application

Description automatically generated

* An accuracy score of the **Balanced Random Forest Classifier** model is **0.7885466545**
* The Confusion Matrix generated from this model:

Text

Description automatically generated

* The Imbalanced Classification Report from this model:

Table

Description automatically generated

**Ensemble Learners - Easy Ensemble AdaBoost Classifier**

**Graphical user interface, text, application

Description automatically generated**

* An accuracy score of the **Easy Ensemble AdaBoost Classifier** model is **0.9316600714**
* The Confusion Matrix generated from this model:

A picture containing chart

Description automatically generated

* The Imbalanced Classification Report from this model:

Table

Description automatically generated

Another request for the challenge was for us to sort the features by importance. This could allow us to determine if we want to handle features differently based on their importance. As you can see the “total\_rec\_prncp” is the most important with an importance score of .078768, which is about 2% points higher than the next most important item. The next few features have minimal differences and several at the bottom have no importance. We could probably remove this data from our analysis. Below show the top and bottom five items out of a total of ninety-four features.

Table

Description automatically generated

### **Summary:**

Here is a summary table of our six models with their accuracy scores and the Precision, Recall and F1 results from their imbalanced Classification reports:

Table

Description automatically generated

Scoring definitions:

**Accuracy**: How many of our predicted scores are correct.

**Precision**: How close our predicted scores are together in a group

**Recall:** Also known as **sensitivity** and the closer to one the better the score.

**F1**: Is a harmonic mean of the precision and recall scores.

With these definitions the F1 score is probably the best score to evaluate our models. The better the model is at making correct prediction the closer the F1 score is to 1.0. Given the results above the “Easy Ensemble AdaBoost Classifier” method is significantly better than all the other in evaluating high-risk loans and on average is very close to the 1.0 score across the board, so that would the model we would recommend.