# Overview

This analysis was conducted to understand how to utilize Machine Learning statistical algorithms for making predictions based on patterns in the dataset provided from **LendingClub**, a lending service company.

To complete this analysis, we use different Supervised Machine Learning procedures to train and evaluate the data with unbalanced classes. The dataset has an unbalanced classification problem due to many good loans over the lower amount of risky loans. To balance out the classifications, allow for more insightful predictions and increase the accuracy score, various Machine Learning algorithms to resample the data were employed. The following algorithms were used:

* RandomOverSampler
* SMOT
* ClusterCentroids
* SMOTEENN
* BalancedRandomForestClassifier
* EasyEnsembleClassifier.

# Results

The **LendingClub** dataset contained 115,675 loan applications in the first quarter 2019 (1Q19). Loan Status was to determine if the application was considered a low or high risk. Applications that had current as the loan status were classified as low risk and the remaining applications as high risk. The dataset was condensed to to 68,817 applications with 99% of the applications classified low risk.

Image

The data was split using the 75/25% method for training vs. testing. 51,366 applications were listed as low risk and 246 were categorized as high risk in to the training set.

Image

## ##Deliverable 1: Use Resampling Models to Predict Credit Risk

### ###Oversampling

The RandomOverSampler Model results classified 51,366 records each as High Risk and Low Risk.

Image

* Calculated balanced accuracy score: 79%.

Image

* The precision rate for High Risk was 1% with a recall of 70% giving this model an F1 score of 2%.
* The precision rate for Low Risk was of 100% and recall of 61%.

Image

Image

## Synthetic Minority Oversampling Technique (SMOTE)

* Calculated balanced accuracy score: 66% only slightly higher.

Image

The precision rates were only slightly different than the first model.

* The precision rate for High Risk was 1% with a recall of 63% giving this model an F1 score of 2%.
* The precision rate for Low Risk was of 100% and recall of 69%.

Image

Image

### ### Undersampling

The undersampling model classified High Risk and Low Risk records as 246 each.

Image

* Calculated balanced accuracy score: 54.5% lower than the above models

Image

* The precision rate for High Risk was 1% with a recall of 69% giving this model an F1 score of 1%.
* The precision rate for Low Risk was of 100% and recall of 40%.

Image

Image

## ## Deliverable 2: Use the SMOTEENN algorithm to Predict Credit Risk

### ###Combination (Over and Under) Sampling

The SMOTEENN model classified 68,460 records as High Risk and 62,011 as Low Risk.

Image

* Calculated balanced accuracy score: 65%

Image

* The precision rate for High Risk was 1% with a recall of 72% giving this model an F1 score of 2%.
* The precision rate for Low Risk was of 100% and recall of 57%.

Image

Image

## ### Deliverable 3: Use Ensemble Classifiers to Predict Credit Risk

#### BalancedRandomForestClassifier Model

The BalancedRandomForestClassifier classified 51,366 as High Risk and 246 as Low Risk.

Image

* The calculated balanced accuracy score: 79%, which is higher than the above models.

Image

* The precision rate for High Risk was 3% with a recall of 70% giving this model an F1 score of 6%.
* The precision rate for Low Risk was of 100% and recall of 87%.
* The total\_rec\_prncp feature was at 7.9% of the total.

Image

Image

Image

#### EasyEnsembleClassifier Model

* The calculated balanced accuracy score: 93%, which is higher than all the models.
* The precision rate for High Risk was 9% with a recall of 92% giving this model an F1 score of 16%.
* The precision rate for Low Risk was of 100% and recall of 94%.

Image

Image

# ### Summary

Of the six models, the EasyEnsembleClassifer model provided the highest accuracy rate of ?? with a rate of precision of predicting High Risk and low risk candidates at ?? and ??, respectively. The high risk and low risk recall or sensitivity rate was ?? and ?? respectively. These were also the highest rated of all of the six models.

Based on the results I would recommend the EasyEnsembleClassifer model to perform these types of analysis.

For future analysis would be use a different data set. The **LendingClub** dataset had 99% of the applications classified as low risk with only 1% of the data classified in the high risk category. Machine Learning algorithms using a dataset like this one may cause a skew in the results by creating groups from a dataset that is has a very small number of High Risk applications.