Phase 1

The topic that I chose to undertake for my machine learning project is credit risk. I am using loan stats from 2019 to predict who the bank should give loans to and how much money the bank should loan. The model that I build will predict who will be able to pay their loans back and who will not.

I decided to take on credit risk for my final project because I have a friend who recently took a job in banking and has worked in loan management. I have talked to him about his work experience in this field and it has intrigued me. Learning about how he decides who should receive loans and who should not receive them gave me the idea that I could create a machine learning model on my own to predict giving out loans. I decided to take the opportunity to use my final project to learn and predict who the bank should give loans to, who they should not loan money to, and how much money the banks should loan to their applicants.

I will be using numerous variables to predict who will be capable of paying back their loans to the back. A few examples of variables that I will use to predict who will repay their loans are loan amount, interest rate, installment, income, and installment. Taking all the variables that are considered on a loan application and creating a machine learning model will be able to show which variables are most important on a loan application. Having this information will allow the bank to loan money to people who will pay the bank back.

Phase 2

I found my data for this project on [Kaggle](https://www.kaggle.com/datasets/gova202138/loanstats?resource=download). The data consists of loan applicants from the first quarter of 2019. I chose this data set because it has many different features for applicants, and it has a very large sample size. The data consists of over 100,000 applications and almost 150 features.

I created a variable called columns and made a list of all the different columns I planned on using in order to create my model. I then loaded in the file and modified the file to only produce the features listed in the columns variable by using the loc method. I dropped all null columns and rows and changed the dtypes where necessary.

From sklearn.metrics, I imported balanced accuracy score, confusion matrix, and classificiation report imbalanced. I imported train test split in order to create a training set and a test set to run the data on. From sklearn.linear model, I imported logistic regression in order to get a logisitic regression of the data. Imblearn.over smnapling allowed me to import SMOTE and Cluster Centroids. Imblearn nombine allowed me to import SMOTEENN. And imblearn.ensemble imported BlaancedRandomForestClassifier and EasyEnsembleClassifier.

Phase 3

To complete the project I used several models. I used logistic regression, random oversampler, SMOTE, Cluster Centroids, SMOTEENN, BalancedRandomForestClassifier, and EasyEnsembleClassifier. I am expecting a high accuracy rate because I expect that a few variables will be key indicators of who will be able to pay back their loans.

I chose logistic regression because I have many different variables in my data that are independent from one another. Logistic regression is the model that I hypothesize will be the most effective for this project. I am predicting this because of its ability to use variables that are independent of one another.

Random oversampler selects examples from smaller proportion of the data and adds them to the training set. I used random oversampler because I felt that it was important that the minority class of the data was not overlooked.

SMOTE is used to create manufactured samples for the minority class and it helps prevent the possibility of overfitting that comes from over sampling. I used SMOTE in order to prevent the training set from getting overfitted.

Cluster Centroids represents the center of a cluster and it organizes clusters into non-hierarchal categories. BalancedRandomForestClassifier works just like a random forest classifier, but it balances the prevalence class by under sampling.

I will be using each of these models to find out how accurate they all are and then find the one that is best to predict credit risk. I expect them all to be pretty accurate because I am predicting that a few key features will play a major role in who the bank should give loans to.