Ky Anh Nguyen

Capstone Three Project Proposal

For this capstone project, I will be working with the "Home Equity Line of Credit" dataset. The data consists of clients who have been given loans of a maximum amount within an agreed period (called a term), where the collateral is the borrower's equity in his/her house (akin to a second mortgage). The clients are separated into two classes: "good" and "bad." A client is placed into the "good" category if they repaid their HELOC account within 2 years, and are placed into the "bad" category if they did not. There are 23 predictors in the dataset. The ultimate goal of this project is to build a model which will predict which customers will repay their loans within 2 years and which will not. For this a variety of models will be considered, such as K-Means clustering, SVM, Random Forest, and XGBoost. Our criteria for success will be the accuracy scores and classification results of our models. The scope of solution space will consist of examining relationships between the predictors and the predicted variable (in this case the predicted variable is called "Risk Performance," which is where clients are categorized as either good or bad), relationships between the predictors, and preparing for model development by using techniques such as feature selection, normalization/standardization, etc. Some possible constraints for our project are relevant variables related to the client's financial situation that are not included in the dataset. Since we can only build our models based on the available features in our dataset, our models may not be as accurate as desired. The stakeholders or people who would be interested would be lenders and banks. The dataset comes from the kaggle website and is originally from the "Explainable Machine Learning Challenge" organized by a FICO company.