Project Proposal

**AML SPR-19**

Naimesh,Nishad , Vishal

Indiana University, Bloomington

**Abstract(Naimesh):**

Assessing risk when providing credit to an individual is a major problem all lenders face. It is crucial to know how likely a person is to repay the borrowed money in an agreed upon time. There are many factors in a person’s life lenders can use to assess this risk. A person’s previous credit history, their occupation, their age, their location, their credit card usage, etc. We will be studying these factors into when trying to assess an applicant risk. The dataset we will be using is provided from Home Credit which contains samples of over 600K credit application. We will use all the factors provided in the csv files to understand which factors are the most important in predicting a person’s default risk. We have done a baseline run on the provided data and calculated an accuracy score of xx% via the Kaggle service. Our goal is going to be to create a machine learning model that is both quick and well optimized to asses this risk for leaders and provide them with the decision-making guidance they need to maximize their profits.

**Project Title**

Assessing approval risk for home credit applications

**Dataset Description(Vishal) 200-250 words:**

**Research Design and Methods (Naimesh) 200-250 words:**

Understanding that this is a classification problem, we will use algorithms that are best suited for this. Some of the algorithms we are considering are Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosted Trees. Logistic regression will be out baseline estimator as it fairly easy to understand and performs quick. All algorithms will be tested with multiple hyperparameters during the grid search process. Due to the raw amount of data and one to many relationships present in them we will be aggregating many of the dataset to an individual application. Along with this we will focus on creating many different features via transformations in hopes that they will be good estimators for our dataset. We will be testing multiple different components during PCA analysis to lower the dimensionality of our data. Our goal will be to take these steps into pipelines so that we avoid any data leakages. Once we have run the models through the pipelines, we will focus on assessing their accuracy on a test and a validation set. The accuracy metrics we will focus to help choose and optimal model are area under curve, accuracy score, log-loss, recall, precision, and F-score. We want to focus on assessing all these metrics together because there are many times that scores like accuracy might mislead us. In our dataset we have an imbalanced distribution and because of this we will be looking at all metrics together to give us a better grasp of the data and the validity of the models we have created.

**Block diagram / Pipelines (Nishad) – Optional**

**Work Contributions (Nishad)**