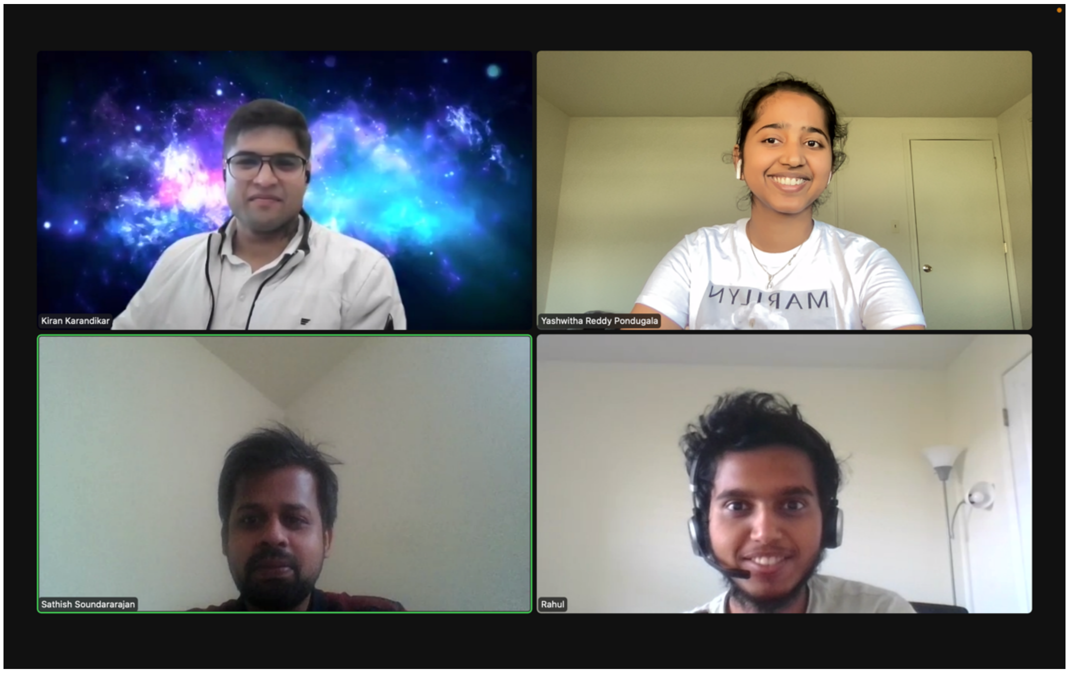
Home Credit Default Risk

Abstract

Finding the Loan applicants who are very likely to repay the loan is an existential dilemma for any Loan provider nowadays. Companies can prevent losses and make large profits this way. Home Credit provides straightforward, cheap, and quick loans for a variety of items such as home appliances, cell phones, laptops, two-wheelers, and a variety of personal necessities. So, an ML model can be used to predict who is capable of repaying a loan, given the applicant data, all credits data from Credit Bureau, previous applications data from Home Credit and some more data. Data used in building this ML model is downloaded from Kaggle. Data engineering is done using exploratory data analysis and feature engineering data. Building a model using various machine learning techniques from linear regression to non-gradient-based models like decision tress and random forest are used to estimate how competent each applicant is of repaying a loan, with the goal of only approving loans for those who are likely to repay them can be an effective method. So, using this model it can be determined if a person is likely to payback or not.

Group Members:
Kiran Karandikar(kikarand@iu.edu)
Sathish Soundararajan(satsoun@iu.edu)
Yashwitha Reddy Pondugala(ypondug@iu.edu)
Rahul Gomathi Sankarakrishnan(rgomathi@iu.edu)



Data Description

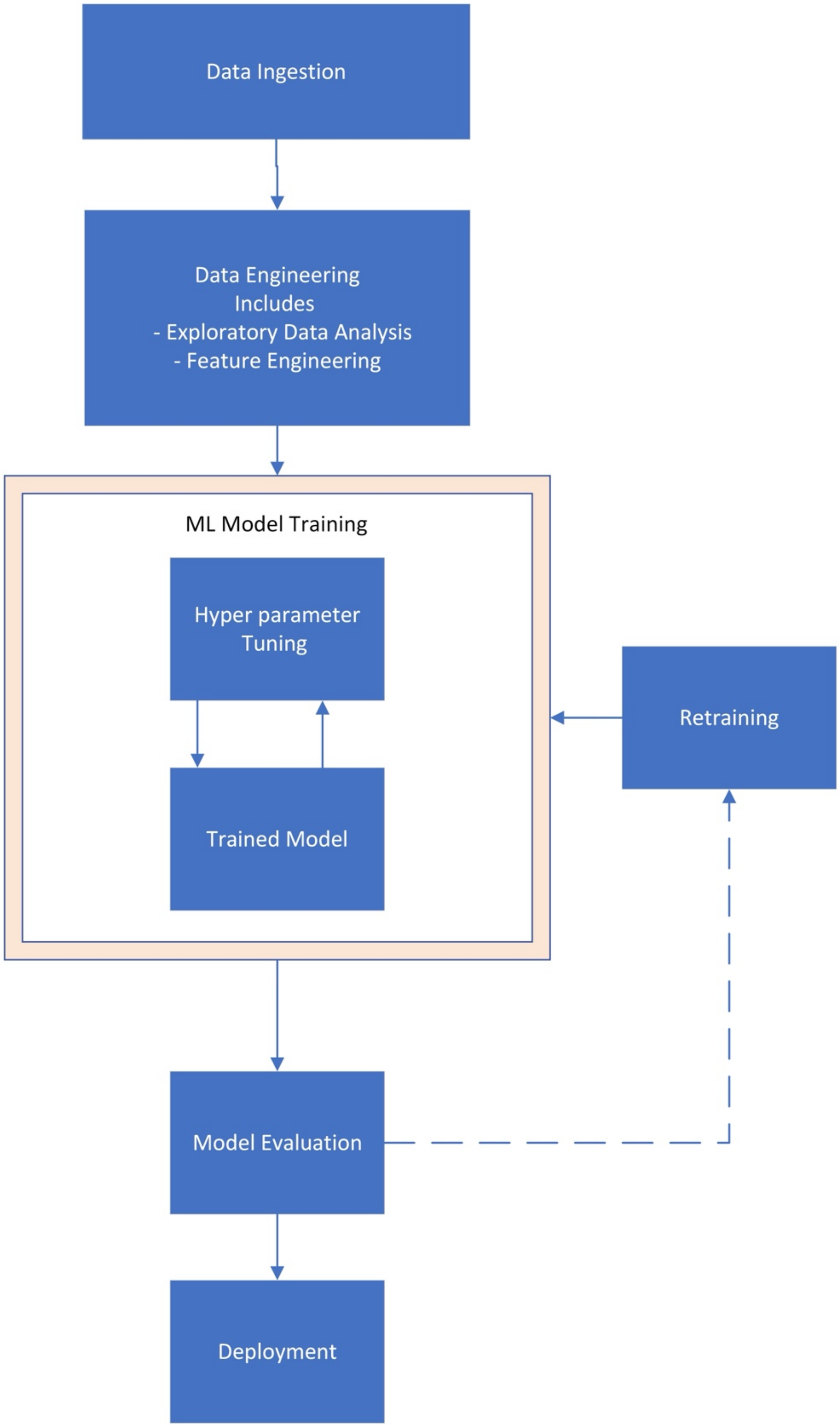
The data is divided into train and test data. A unique loan ID corresponds to a single loan in application.csv. Though a person might have several loans each loan will be treated uniquely. The bureau_balance.csv consists of the information regarding the balance of the previous credit card bills owned by the home credit takers. Each row has each client's loan in credit Bureau. The Previous_application.csv consists of the data regarding the applicant if there are ay previous loans present. It also contains information regarding the applicant during the time of previous application. Each row in this file represents one previous application.

This Previous_application.csv further can be connected to Pos_Cash_Balance.csv and installments and credit card balance. The Pos_cash_balance.csv contains info that contains monthly balance of client's previous loans regarding home credit and some behavioral data. The csv file labelled instalments_payments.csv contains data regarding past payments of each installment. The credit_card_balance.csv contains information regarding monthly balance of credit card loans in home credit.

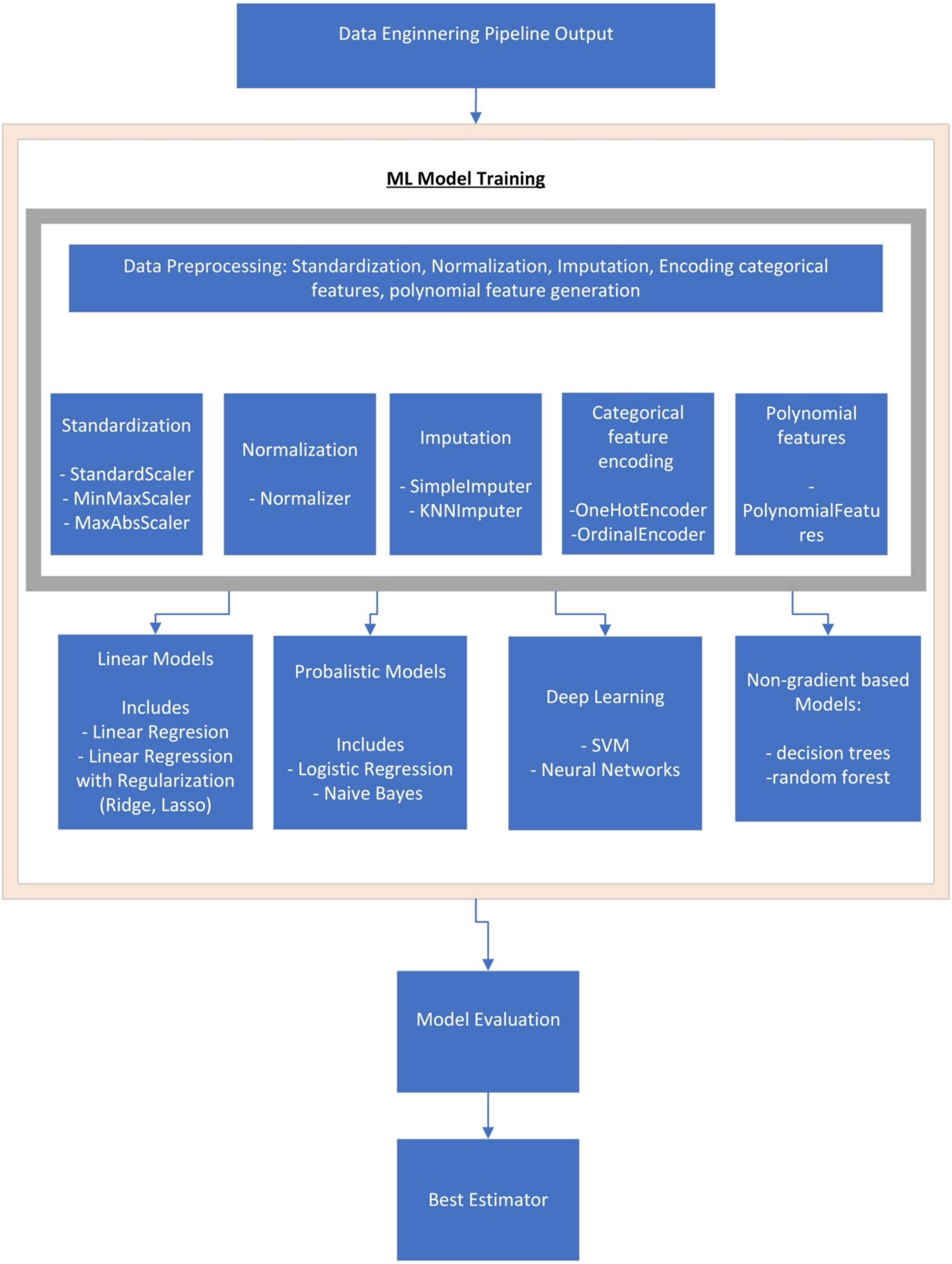
Machine Learning Algorithms and Metrics

The data obtained from Kaggle will be ingested. Then we plan on data engineering using Exploratory Data analysis and Feature engineering. Then we are going to train the model. The machine learning model training includes data preprocessing and various machine learning models. The data preprocessing contains standardization, Normalization, Imputation, Encoding categorical features, polynomial feature generation. We plan on using StandardScaler, MinMaxScaler, MaxAbsScaler for standardization. Normalization is done using a normalizer. Whereas imputation is done with SimpleImputer and KNNImputer. OneHotEncoder and OrdinalEncoder are used for encoding categorical features.

Models we plan on using include linear models, Probabilistic models, deep learning models and non-gradient-based models. Models will be improved assessing appropriate metrics such as AUC_ROC curve, mean squared error or RMSE. AUC_ROC will be included as it is specified as metric in the Kaggle and also because it can measure the false positive rate of the model. The diagrams below show the basic flow of the model and also the algorithms used respectively.



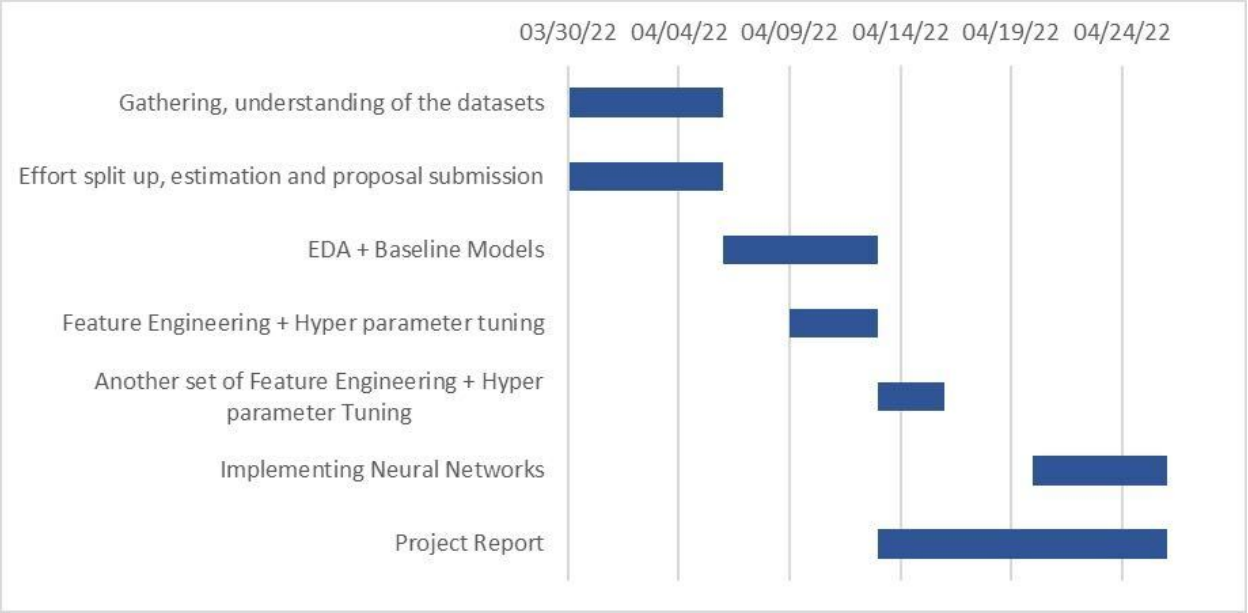
Basic Flow of the Model



Algorithms Used in the model

The linear models used are Linear Regression and Linear Regression with Regularization (Ridge, Lasso). The probabilistic models include Logistic regression and Naïve bayes. Deep Learning algorithms such as SVM and Neural Networks are also being used in our model. Lastly the non-gradient models such as decision trees and random forest are used. Later the model is evaluated, and the applications will be labelled "No Risk" if they have been assessed they have the capacity to repay it and "Default Risk" otherwise. Models will be improved assessing appropriate metrics such as AUC_ROC curve, mean squared error or RMSE.

Timeline



Roles and Responsibilities

We want to work on each segment cooperatively, but each member will be in responsible for "managing" the task to guarantee its completion. We have decided on working on project in the following way:

- Rahul will be responsible for EDA and Baseline Modelling
- Kiran is responsible for Feature engineering and Hyperparameter tuning to create a model that can be accurate, reliable, and efficient.
- Yashwitha will be responsible for the Documentation of the project at each stage and final report.
- Satish will be taking up the pyTorch model and deep learning model.
- We are planning on working as a group regarding the ensemble methods.