Home Credit Default Risk (HCDR)

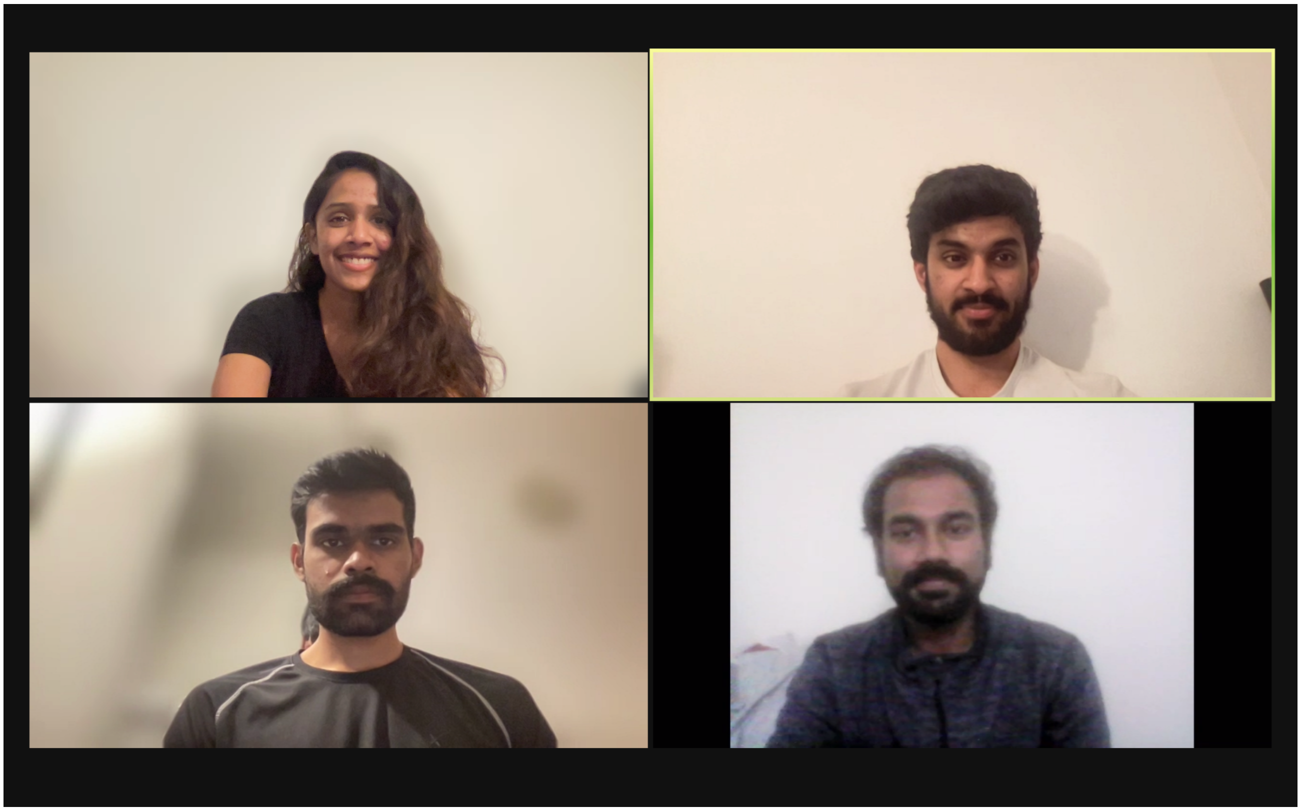
Group Members:

Chandra Sagar Bhogadi : cbhogadi@iu.edu

Hanish Chidipothu : hachid@iu.edu

Nidhi Vraj Sadhuvala : nsadhuva@iu.edu

Srinivas Vaddi : svaddi@iu.edu



Abstract

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. Home credit strives to broaden inclusion for the unbanked population by providing a positive and safe borrowing experience. The objective is to predict whether a client will repay a loan or not . In order to make sure that people who struggle to get loans due to insufficient or non-existent credit histories have a positive loan experience, Home Credit makes use of a variety of alternative data, including telco and transactional information to predict their clients' repayment abilities. Credit history is a metric that explains a user's credibility, and it's calculated using variables like the user's average/minimum/maximum balance, Bureau scores reported, salary, and repayment patterns, which may be analyzed using the user's previous timely defaults/repayments. Other criteria, such as location data, social media data, calling/SMS data, and so on, are included in alternative data. As a part of this project, we will be using dataset provided by Kaggle platform to perform EDA, design ML pipelines using several machine learning methods to assess models against a variety of metrics before deploying them, and test our models using various evaluation metrics such as ROC/AUC, Accuracy, and F1 Score and assess models against a variety of metrics before deploying them

Data Description

We intend to use datasets provided by Kaggle platform:

Link: Home Credit Default Risk | Kaggle

application_{train|test}.csv

- This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
- Static data for all applications. One row represents one loan in our data sample.

bureau.csv

- All client's previous credits provided by other financial institutions were reported to the Credit Bureau (for clients who have a loan in our sample).

- For every loan in our sample, there are as many rows as the number of credits the client had in the Credit Bureau before the application date.

bureau_balance.csv

- Monthly balances of previous credits in the Credit Bureau.
- This table has one row for each month of history of every previous credit reported to the Credit Bureau

POS_CASH_balance.csv

- Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.

credit_card_balance.csv

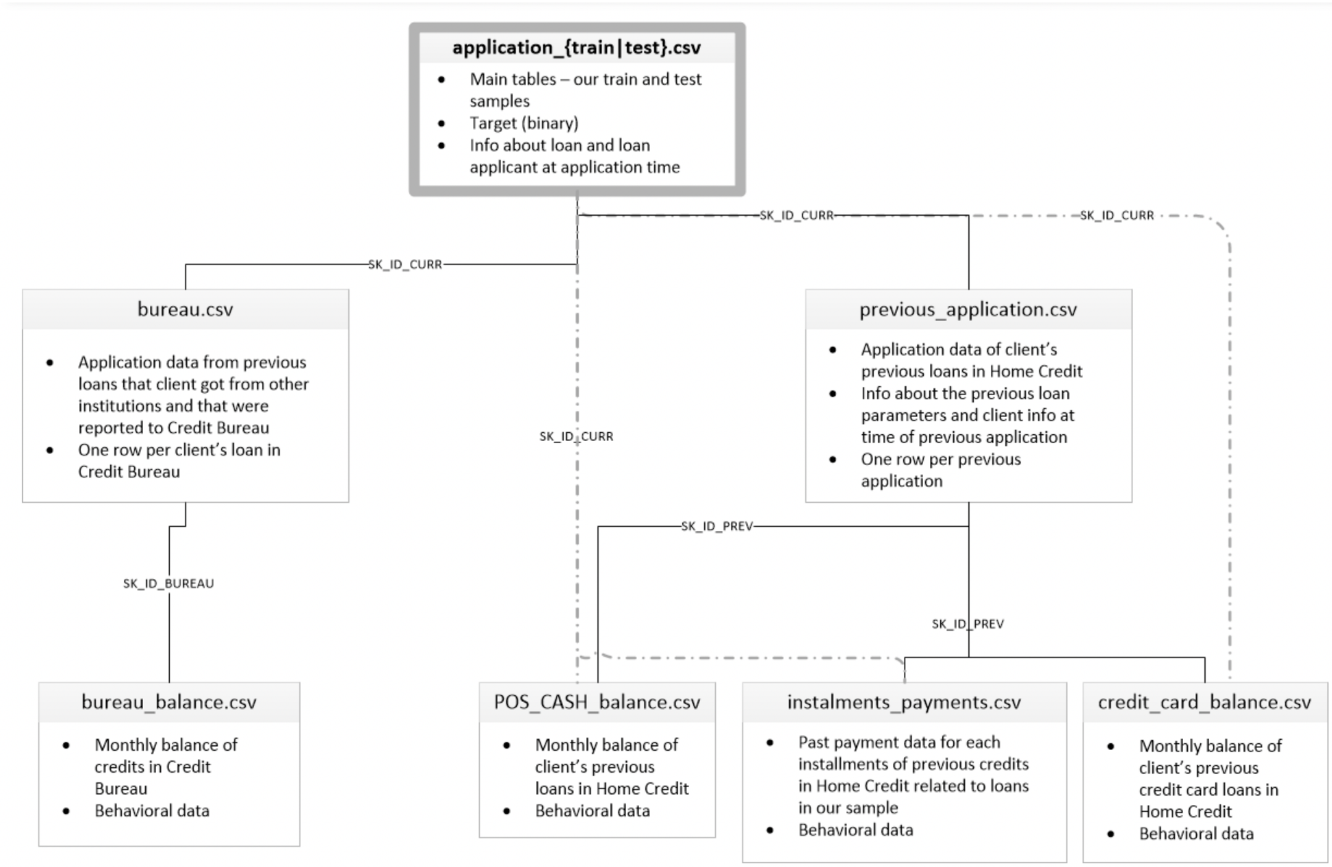
- Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.

previous_application.csv

- All previous applications for Home Credit loans of clients who have loans in our sample.
- There is one row for each previous application related to loans in our data sample.

installments_payments.csv

- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
- There is a) one row for every payment that was made plus b) one row each for missed payment.
- One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Machine Learning Algorithms and Metrics:

The goal here is to predict whether the customer who has reached out to Home Credit for a loan is a defaulter or not. Therefore, this is a supervised classification task, and the output of the target variable is either 1 or 0 where 1 means non-defaulter and 0 means defaulter.

We are planning to use the following algorithms in our machine learning pipeline.

1) Logistic Regression
2) Decision Trees
3) Random Forests
4) Support Vector Machines (with different kernels like- linear, polynomial, radial basis function)
5) Neural Networks (RNN, LSTM'S)

- Logistic Regression can be used as a baseline model along with feature selection techniques like RFE,PCA, SelectKbest.

- Support vector machine is a non-probabilistic binary classifier . It maps training examples to points in space so as to maximize the width of the gap between the two categories. It can also use the kernel trick to perform non linear classification, implicitly mapping the inputs into high dimensional feature space.

- Random forests or random decision forests is an ensemble learning method that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

- LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms which can be used for classification.

- Deep learning neural networks may be used to improve the prediction model's accuracy, but we wouldn't be able to give the attributes that determine whether or not a client is a defaulter. This would lead to compliance concerns, as we would need to provide the specific features that would cause the loan to be rejected for the likely non-defaulters.

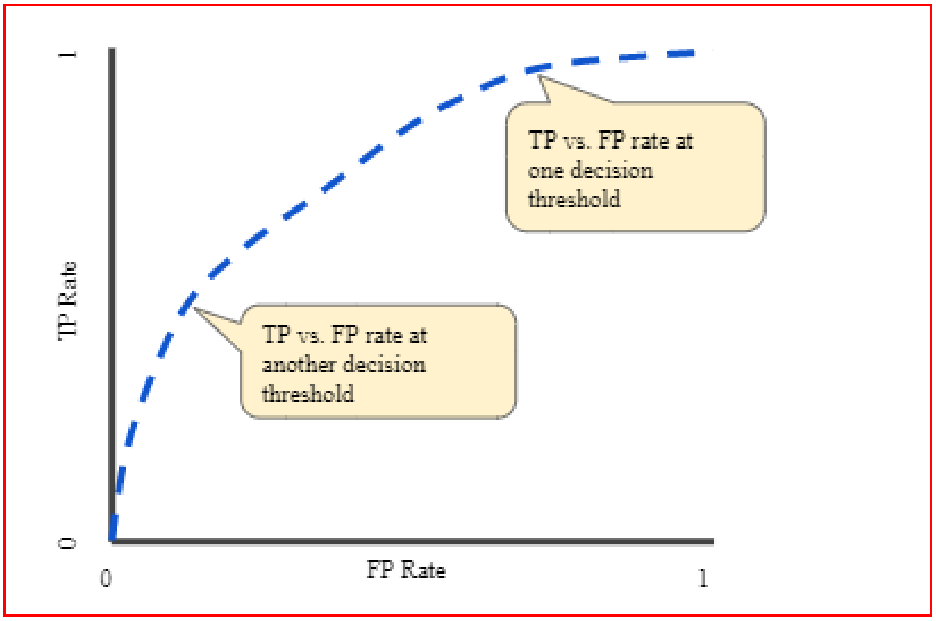
Loss Functions:

The loss function for logistic regression is log loss (or cross entropy), the loss function for support vector machines is hinge loss, information gain (KL divergence—mutual information) is used for splitting variables in a decision tree as the output variable is categorical, and algorithms like ID3, CART, c4.5, and c5 can be used to build the decision tree, and the loss function for neural networks is MSE (difference between the expected output and the outcome produced by the model).

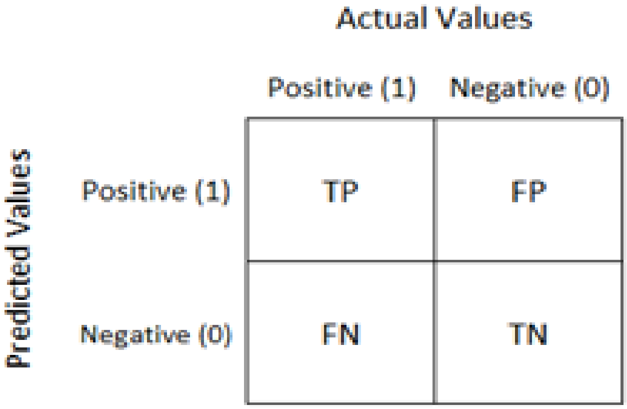
Evaluation Metrics:

1)ROC/AUC-

As this is a classification problem AUC-ROC curve can be used as a performance measurement for classification problems at various threshold settings. It is a probability curve and tells us how much the model is capable of distinguishing between the classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. The curve is plotted with TPR(True positive rate) against the FPR(False positive rate).



2)Confusion matrix - With a confusion matrix, we can see if our model generates the right predictions with respect to the actual value from the test dataset.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

3)Recall

Equation - (True positives)/(True Positive + False Negative)

The above equation can be explained by saying, from all the positive classes, how many we predicted correctly. Recall should be as high as possible.

4)Precision

Equation - (True positive)/(True Positive + False Positive)

The above equation can be explained by saying, from all the classes we have predicted as positive, how many are actually positive. Precision should be as high as possible.

5)Accuracy

From all the classes (positive and negative), accuracy tells how many of them we have predicted correctly.

6)F1 Score

Equation - 2*(precision*recall) / (precision + recall)

It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time.

Machine Learning Pipeline Steps:

1. Data Preprocessing

a. Gather Kaggle's raw data.
b. Perform exploratory data analysis on the dataset.
c. Feature engineering for improving performance of machine learning model.

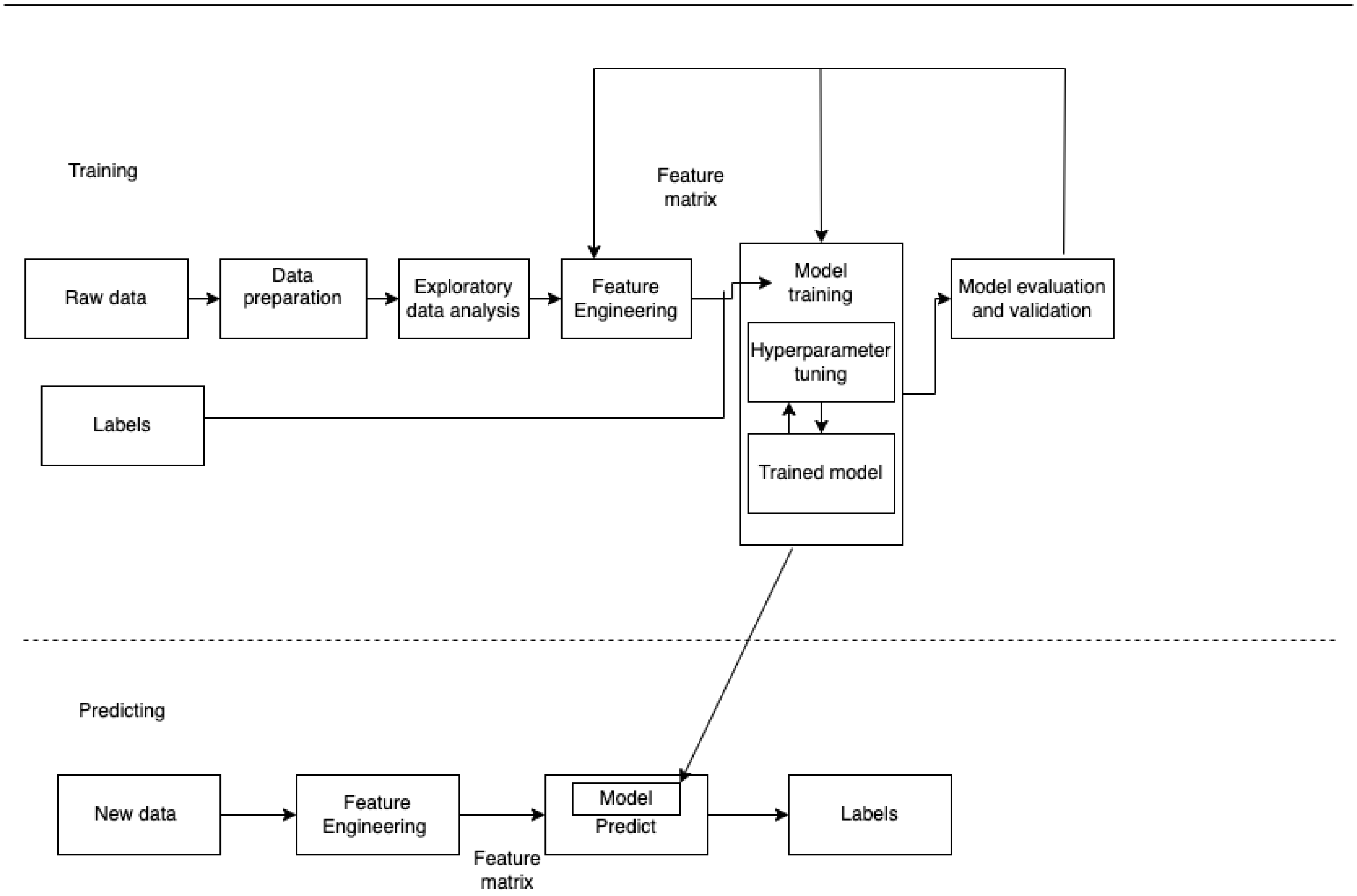
2. Model Selection

a. Develop and test various candidate models, such as "Logistic Regression," "Decision Making Trees", "Random Forest", and "SVMs.
b. Based on the evaluation measures, select the best model.
c. Use various evaluation metrics like "accuracy," "F1 Score," and "AUC."

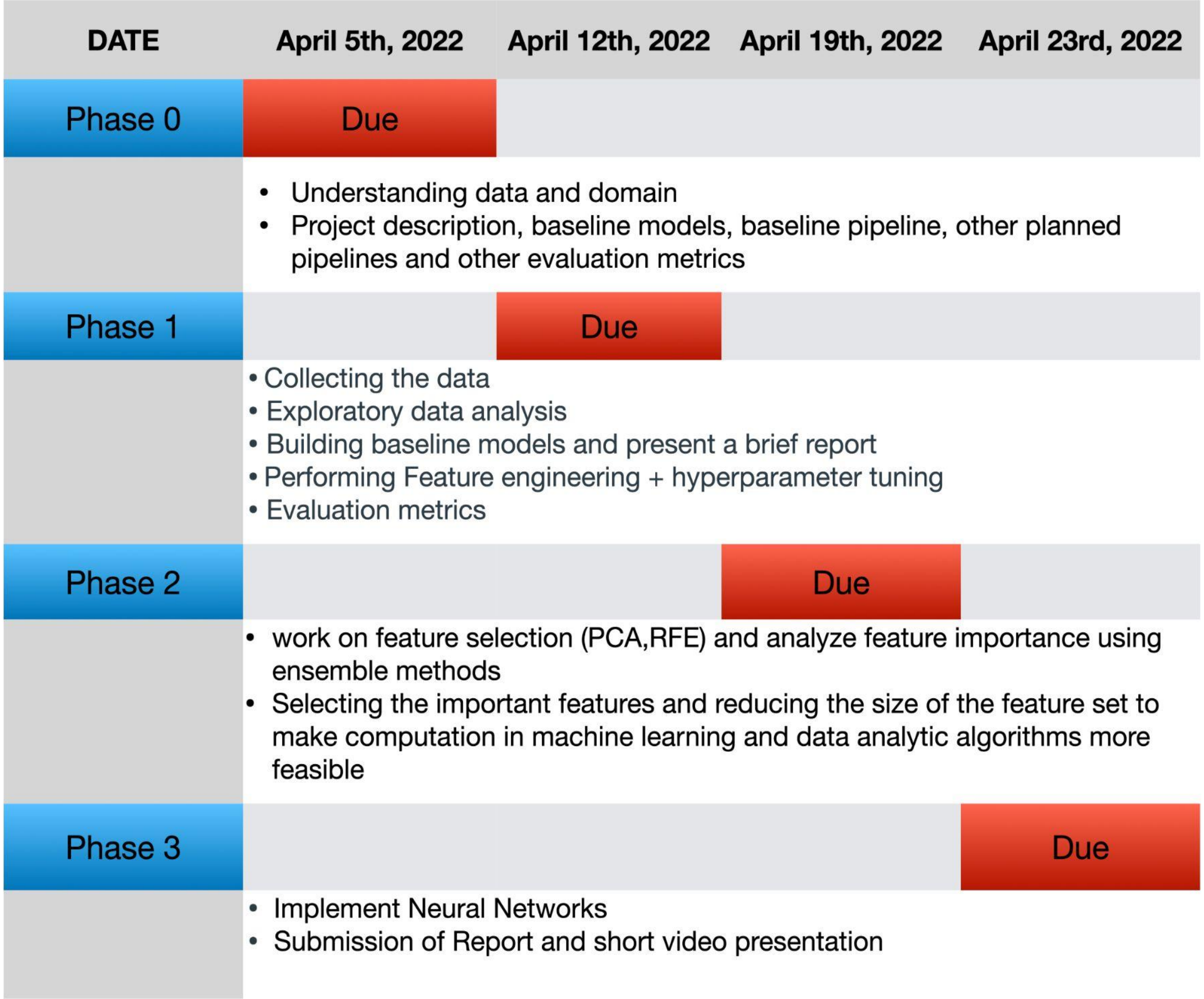
3. Prediction Generation

a. Prepare the new data and extract the features as before.
b. Once the winning model has been chosen, use it to make predictions on the new data.

Pipeline Block Diagram:



Project Timeline



DATE	April 5th, 2022	April 12th, 2022	April 19th, 2022	April 23rd, 2022
Phase 0	Due			
	• Understanding data and domain • Project description, baseline models, baseline pipeline, other planned pipelines and other evaluation metrics			
Phase 1		Due		
	• Collecting the data • Exploratory data analysis • Building baseline models and present a brief report • Performing Feature engineering + hyperparameter tuning • Evaluation metrics			
Phase 2			Due	
	• work on feature selection (PCA,RFE) and analyze feature importance using ensemble methods • Selecting the important features and reducing the size of the feature set to make computation in machine learning and data analytic algorithms more feasible			
Phase 3				Due
	• Implement Neural Networks • Submission of Report and short video presentation			

Member's Contribution:

We intend to collaborate on each segment, but each member will be responsible for managing few tasks to assure its completion:

- The group will be in charge of drafting and laying out the project proposal.
- Nidhi Vraj is in charge of exploratory data analysis and the creation of ML baseline models.

- To attain optimal accuracy and efficiency, Chandra Sagar Bhogadi will be in charge of the model's Feature Engineering and Hyperparameter Tuning.
- Hanish Chidipothu will focus on feature selection (PCA, RFE) and feature importance analysis with ensemble approaches.
- Srinivas Vaddi will be in charge of building the PyTorch model and implementing Neural Networks.
- We will collaborate on the presentation and report as a group.