1.0.0.0. Phase 1: Literature survey & Data Acquisition

1.1.0.0. Problem Description

1.1.1.0. The problem at hand is an interesting and real-life business problem which is faced daily by financial lenders. The problem is about automating the loan approval process which in turn will impact the KPIs (Key Performance Index) of profitable lending teams.

A lending company - Home Credit - wishes to expand its lending business and provide loans to people with insufficient or non-existent credit histories. For taking the decision of whether to give loan to an individual or not, machine learning model(s) shall be trained and deployed to measure a clients' repayment abilities. For training these models 2.68 GB of data is provided in form of various csv files. These are a variety of in-house data and data collected from other sources.

1.1.2.0. The importance and impact of this problem are as under:

1.1.2.1. Profitability of lending company - If the process of decision making is automated it will positively affect KPIs like velocity, pull through, cost-to-close [a] which in turn enhance the profits of the lending company. It is quite evident that the time saved by automating this process will result in fast and efficient working of lending companies which translate to increase in profit. The trained model (for automating the process of loan approval) will be evaluated against various metrics (to be discussed subsequently). This evaluation is necessary so that the credit worthy people are not deprived from loans and the credit unworthy are not given loans.

1.1.2.2. Social and financial inclusion of people with insufficient or non-existent credit histories - The presence of organised and trustworthy lenders for the unbanked population is important else this population will be exploited by the unorganised and untrustworthy lenders. Loans from organised lenders will also ensure that the credit worthiness of this population is documented which in turn makes it easy for the same and other lenders to give loans to this populsation.

1.2.0.0. Dataset

1.2.1.0. Data (2.68 GB) sourced from kaggle [b] is in the form of csv files. There are 8 csv files containing data tables and 1 csv file containing the columns description of the other 8 files. To understand the data set it is important to go through all the tables represented by these csv files with the understanding of each column from HomeCredit_columns_description.csv. The data set is huge in size and needs considerable time to be spent to correlate various tables. The data files can be broadly classified into three categories which are listed below with brief description of each data table:

1.2.1.1. Train and test data - These are data provided for training and evaluating the model:
application_train.csv - This data table consists of 122 columns. Each row indicates an entry corresponding to 1 loan application. The column titled SK_ID_CURR is the

main identifier for each entry. Column titled TARGET has two values - 1 for clients with payment difficulties or default and 0 for clients without payment difficulties or no default. There are several other columns e.g., columns indicating whether the client has a car or not, gender of client, loan annuity, income of client, age of client, the day on which application was made, detail of client's home etc. We have a lot of features/parameters.

application_test.csv - The columns of this data table are the same as columns of application_train.csv except for the absence of the TARGET column. This data set shall be used to evaluate the model trained using application_train.csv.

1.2.1.2. Credit data from other sources - It consists of data tables which contain information of applicants collected from sources other than Home Credit and which were reported to Credit Bureau (Credit Bureau is a centralised agency which documents the data of all financial lenders). Following are the data sets in this category:

bureau.csv - This consists of data of the client's previous credits with each row indicating 1 credit. Thus there can be multiple rows pertaining to 1 client. The identifier for a credit is given in SK_BUREAU_ID and it is related to the client identified by SK_ID_CURR column.

bureau_balance.csv - This data table consists of monthly data against the credits specified in bureau.csv. The identifier for credit is SK_BUREAU_ID.

1.2.1.3. Credit data from client's previous application with Home Credit - It consists of data tables which contain information of applicants who are already availing loans from Home Credit. Following are the data tables in this category:

previous_application.csv - This data table consists of data of a client's previous loan application with Home Credit. Each row indicates a previous loan by a client where a previous loan is identified by SK_ID_PREV column and client is identified by SK_ID_CURR column.

POS_CASH_balance.csv - The rows in this data table indicate each month of previous cash loan with Home Credit.

credit_card_balance.csv - The rows of this data table capture the details of each month of each previous credit card the client has with Home Credit.

installments_payment.csv - This data table consists of 1 row of every payment that was made and one row for every missed payment pertaining to previous credits of clients from Home Credit.

1.2.2.0. Challenges with this data set - The data set is huge and hence requires considerable time to be spent for finding the correlation among different data tables. At the same time the pre-processing will involve a lot of effort for cleaning involving deduplication, missing value imputation, removal of data with values which are out of bounds etc. While so many tables provide opportunity for feature engineering, it also comes with challenges like creating and trying different features and finding which are really impacting the result.

1.2.3.0. The data is furnished in the form of csv files. Pandas and Numpy shall be used to read the data and do various processing to bring the data in the form which can be fed in the machine learning model. Libraries like Matplotlib, Seaborn etc. shall be used for visualisation. The choice of library shall be dynamic in nature and will depend upon the need as perceived during different phases of this project.

1.2.4.0. Data acquisition from other sources can only be possible if the lending companies are ready to share their data. It is also possible for someone to have access to other company's data if they are working for that company and have necessary authorization or are working with the regulatory board. Such data is sensitive for any lender and not publicly available and hence not possible to get easily.

1.3.0.0. Key metric (KPI) to optimize

There are certain business metrics to evaluate for financial lenders. Some examples of these are velocity, pull through, cost-to-close [a]. These are financial lending domain specific metrics and can be understood in easy language through references made in this document. However, the metrics used for the machine learning task at hand shall be measurements like accuracy, deductions from confusion matrix (Precision, Recall, F1 score etc.), AUC (area under curve) in ROC (receiver operating characteristic) curve etc.

Data is expected to be highly imbalanced as for any profitable lending business the default rate has to be very low. In such a situation, accuracy will not be helpful as accuracy can be high just by using a simple yes or no model in case of imbalanced data. Hence, metrics related to confusion matrix and AUC in ROC curve will be useful in this case.

In this case the requirement is that the genuine clients should get a loan and expected defaulters should not get a loan. The cost of false positive and false negative both are high, especially false positive (where loan is approved for lenders who will default). Similar situations arise where the cost of false negatives is very high e.g., medical; and in case of detection of fraudulent transactions. In all these situations the metrics discussed above are used for evaluating the model.

An implementation of these metrics from scratch for a toy data set is annexed with this document. The implementations also contain comparison of results with those obtained from standard libraries. Moreover, explanations of codes are also given in the form of comments. Complete code is split into logical sections. Implementation of these metrics through standard libraries shall be done in relevant phases of this project.

1.4.0.0. Real world challenges and constraints

1.4.1.0. The various challenges and constraints are discussed as under:

1.4.1.1. Size and structure of data - 2.68 GB of data with 7 data tables for training, 1 data table for evaluation and hundreds of columns is a lot to process. While handling is

not difficult with so many optimised libraries, data preprocessing/cleaning becomes a challenge with this size of data. As there are a lot of columns spread across 8 data tables, a lot of correlation has to be understood across different tables. This needs considerable time and effort.

1.4.1.2. Imbalance in data - The data is expected to be imbalanced. So accuracy cannot be a reliable metric for evaluation of a model. This calls for use of other metrics for evaluation e.g., confusion matrix, AUC in ROC curve.

1.4.1.3. Domain knowledge - This is financial domain data. An understanding of the basic concepts and sometimes applied concepts is required for a machine learning engineer for feature engineering. The lending business involves understanding various ratios [c] which are key metrics while deciding credit-worthiness of a person (i.e., whether a person shall be given a loan or not).

1.4.1.4. Data cleaning - A cursory view of the tabular data reveals that a lot of data is missing. Missing value imputation shall be done in this situation. However, the missing value imputation methodology will vary from column to column. It may happen that some columns have to be dropped as there is a very high percentage of missing value. Data needs to be checked for values which are out of bound and such data need to be removed or replaced. Data shall be checked for deduplication and removal of duplicate rows. Categorical data shall be converted into numerical data. For this one hot encoding shall be used. All these shall be done along with EDA (exploratory data analysis).

1.4.2.0. While implementing machine learning, more than 1 model shall be tested and the best shall be deployed. This model is expected to give low false positives and low false negatives.

1.5.0.0. <u>Approach to solve similar problems:</u>

1.5.1.0. After reading blogs and literature available on the internet [d], it is found that the approach to solve such problems involves EDA and data cleansing as first steps. EDA helps us get insight into data and gives an idea of the various types of data cleansing to be employed. After the data is cleaned, more than 1 model is tried. The most common models that are applied are logistic regression, tree based models and neural networks.

1.5.2.0. My approach towards solving this problem will include the following. The details shall be documented during the relevant phases. The approach and methodology may evolve over time as we get more insight during different phases of this project:

1.5.2.1. Studying the data and finding correlation among different data tables is the first step. During this step various domain specific insights shall also be collected which will help in feature engineering if required. This step is the phase 1 (current phase) of the project thesis.

1.5.2.2. EDA and data cleansing shall be done simultaneously. During EDA, various insights of data shall be obtained. This will give a bird's eye view of the correlation that exists among various parameters.

1.5.2.3. Machine learning models shall be evaluated. As it is a classification problem, the first choice is logistic regression. Tree based models and neural networks will also be tried. Based on the performance of these models, a final model will be finalised for deployment.

1.5.3.0. During the course of the project, this document's previous phases may be updated owing to learning and evolution of ideas & approach.

1.6.0.0. References:

[a] https://himaxwell.com/resources/blog/5-kpis-profitable-lending-teams-measure/, https://smeloan.sg/blog/5-commercial-loan-ratios/

[b] https://www.kaggle.com/c/home-credit-default-risk/data

[c] https://corporatefinanceinstitute.com/resources/knowledge/finance/lending-ratios/, https://www.investopedia.com/terms/c/credit-worthiness.asp

[d] https://towardsdatascience.com/machine-learning-predicting-bank-loan-defaults-d48bffb9aee2, https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012042/pdf