1.0.0.0. **INTRODUCTION**

1.1.0.0. Motivation

1.1.1.0. Over the course of several years of evolution, humankind has devised various tools and systems to ensure their continued survival and to enhance quality of life. One such system that was conceived, widely adopted and has stood the test of time is the concept of money. Money helps people achieve a better quality of education, larger chance of business success, access to medical facilities and higher work output. In anyone's life, a situation may come when all of sudden you require cash. In such times of crisis, not everyone can readily arrange for funds from friends, relatives or such means.

1.1.2.0. To address this need, the concept of loan originated. Loans are an important means to tide over difficult times, aim for upward mobility and in the development of individuals and industries alike. With such profound socio-economic considerations, judicious accessibility to loans as well as mutual benefit to lenders as well as recipients are aspects that need to be ensured with a high degree of integrity. Majorly, the onus for ensuring an objective and mutually beneficial lending process lies with the lender. Any mechanism to aid the lender in this process will help them sustain and become or stay profitable and also enable greater disbursal of loans to applicants deemed eligible.

1.2.0.0. Problem At Hand

1.1.2.1. Given a loan application of a potential or existing client at Home Credit, the ML model needs to predict whether the client will be able to repay the loan or not. Access to past data for a sample of Home Credit applicants aid in the building of this prediction model.

1.3.0.0. What is the problem all about?

1.3.1.0. Increasing population coupled with modernization and consequently, human's quest for lifestyle enhancements have resulted in a huge demand for credit or loans. There is intense competition among traditional banks and numerous credit-lending start-ups to grab their share of this business and attract people by providing different attractive schemes. Start-ups especially, are targeting and catering to unbanked population or first-time credit seekers who have insufficient or non-existent credit histories due to which they are at disadvantage with traditional financial institutions. This unprecedented accessibility in credit availability, market competition and consumption has led to an increase in losses resulting from bad loans. Instead of making money from loan interest, lenders are suffering a huge capital loss. In order to prevent the loss, it is very important to have a system in place which will accurately predict the loan defaulters even before approving the loan. This is especially important for institutions with lending as their primary source of business; such as Home Credit, in order to sustain and grow in the market.

1.4.0.0. Why is this an important problem to solve?

1.4.1.0. Post-pandemic world has disrupted many aspects of life including financial requirements. Many are resorting to loans to ensure basic subsistence. Some struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, such a population is often taken advantage of by unscrupulous lenders. Secondly, lending institutions need to ensure very low credit delinquency rates to stay profitable and provide loans to worthy applicants. An objective model helps the lending agencies

disburse prudent loans. This system helps them process and disburse loans faster, increase profit and client base as well as possibly protect genuine debtors from predatory lenders.

1.5.0.0. Business/Real-world impact of solving this problem.

1.5.1.0. Lender's context - Data-driven, objective decision regarding credit-delinquency ensures a lending process which is swift, has a rational basis and safeguards the agencies' interests while maximising profits. Correlating the metrics of model evaluation to the KPIs of the organisation helps in understanding their standing in the market, potential client base and formulation of future strategies. Going beyond mere binary classification, a model predicting a loan amount threshold which can be offered to the applicant ensuring low default score can help lending-only agencies, especially new companies to gain new clients and grants them the opportunity for greater engagement and interpersonal interactions, a base for future business.

1.5.2.0. Debtor's context - A model that does not rely extensively on past credit histories allows for financial inclusion of people with insufficient or non-existent credit histories. Such groups, which are vulnerable to exploitative lending practises can be catered to by the organised sectors. Loans from organised lenders will in turn improve credit worthiness of this population helping them gain loans in the future, empowering them, financially and socially.

2.0.0.0. **ABOUT THE DATASET**

2.1.0.0. Source of data

2.1.1.0. The dataset is sourced from Home credit Default Risk Kaggle competition.[3]

2.2.0.0. Understanding the dataset

2.2.1.0. The dataset totaling to approximately 2.68GB is in csv format split over 9 files.

2.2.2.0. For ease of understanding, the entire dataset can be grouped under three major heads.

2.2.3.0. The main dataset

2.2.3.1. This is the primary data for training the model and testing its performance. It has two files, '*application_train.csv*' and '*application_test.csv*'.

2.2.3.2. The training set contains 307,511 observations of 122 variables and provides static data for all applicants of Home Credit services. The target variable resides in this dataset and indicates whether clients have had difficulties in meeting payment with two values - 1 for clients who defaulted and 0 for those that did not default. We may consider this as the default dataset. Each observation is a loan application and includes the target value, demographic variables and some other information.

2.2.3.3. The test set contains 48745 entries with 121 variables as it being the test data, target values column shall not be present.

2.2.4.0. 'Applicants having existing history with Home Credit' dataset

2.2.4.1. This data is pertaining to applicants who are already existing clients of Home Credit

2.2.4.2. The file '*previous_application.csv'* contains records of all previous loan parameters and client information for Home Credit loans of clients who have loans in the sample.

2.2.4.3. The file '*POS_CASH_balance.csv'* details monthly balance snapshots of previous PoS (point of sales) and cash loans that the applicant has had with Home Credit.
2.2.4.4. *The file 'installments_payments.csv'* has repayment history for the previously disbursed loans in Home Credit related to the loans in the sample.
2.2.4.5. The file '*credit_card_balance.csv'* contains monthly balance records of previous credit card loans that the applicant has had with Home Credit.

2.2.5.0. 'Applicants having no history with Home Credit' dataset
2.2.5.1. This data is pertaining to applicants who have no prior history with Home Credit.
2.2.5.2. The file '*bureau.csv'* contains data pertaining to previous loans a client had secured from other financial institutions, the details of which were reported to the Credit Bureau (for clients who have a loan in the sample).
2.2.5.3. The file '*bureau_balance.csv'* details the monthly balances of the client's previous credits reported to the Credit Bureau.

2.3.0.0. Quirks and highlights of this Dataset
2.3.1.0. The Home Credit dataset is fairly large sized [~2.6GB] with 121 columns or features in the primary (train) dataset to tinker with. Moreover, the secondary data provides scope for creation of new features, feature interactions and combinations to help the model.
2.3.2.0. As is the case with any profitable lending agency, the Home Credit training dataset has a very small number [approximately 8% of total samples] of credit defaulters. This is the case of an imbalanced dataset which needs to be addressed by either sampling techniques or using appropriate models and metrics.
2.3.3.0. Standard preprocessing and data cleaning procedures like missing value imputation, removal of duplicate entries, detection and handling of glaring anomalies/outliers are especially important in this dataset in order to have relevant features as inputs and produce interpretable results.

2.4.0.0. Tools to get started on the dataset
2.4.1.0. As the dataset is primarily a csv file, using Pandas is preferred owing to its capabilities in handling tabular data such as this dataset; and phenomenal documentation and support in case of possible bugs. In the case of encountering memory problems or sluggishness, Vaex or Dask may be used as an alternative.
2.4.2.0. For EDA and visualizations, Seaborn shall be used for mapping the data on the informative and interactive plots and deriving visual insights.
2.4.3.0. Sklearn shall be used for building machine learning and statistical models such as clustering, classification, regression, etc. Depending on the outcome of EDA and feature engineering, more specific libraries may be tried out.
2.4.4.0. For working out initial strategies regarding metrics, features, popular and powerful libraries such as Numpy and SciPy shall be used.
2.4.5.0. More specific libraries may be used depending on the outcome of EDA and feature-specific cases and the same shall be highlighted at the appropriate stages.

2.5.0.0. Possibility of Dataset augmentation
2.5.1.0. The Home Credit dataset from Kaggle competition is used in accordance with kaggle's Data Access and Use policy. The dataset is created with Home Credit's existing credit delinquency model parameters in context.

2.5.2.0. Open-source data from other sources may not be used directly as both the datasets may have been created with different philosophies and hence, different parameters. However, after feature engineering there is a possibility of using other datasets with similar features, employing suitable imputation or encodings.

3.0.0.0. **METRICS FOR MODEL VALIDATION**

3.1.0.0. The prominent KPIs in the credit lending sector

3.1.1.0. There are some common KPIs frequently followed in lending agencies to assess their healthiness, processes and growth and formulation of business strategy. A brief regarding these is as follows:[1]

3.1.2.0. Pull Through Rate - This KPI measures process efficiency by dividing total funded loans by the number of applications submitted during a defined period.

3.1.3.0. Abandoned loan rate - This KPI measures the percentage of loan applications that are abandoned by a borrower after they have been approved by the lender.

3.1.4.0. Application approval rate - This KPI is calculated by dividing the amount of approved applications by the amount of submitted applications.

3.1.5.0. Customer Acquisition Cost - This key financial measurement is the ratio of a borrower's lifetime value to a borrower's acquisition cost. This KPI is used by lenders to help determine how much of its resources can be profitably spent on a particular customer. The costs include but aren't limited to research, marketing and advertising. Ideally, the customer acquisition cost should be greater than one since a borrower isn't profitable if the cost to acquire is greater than the profit they will bring to a lender.

3.1.6.0. These business KPIs can be correlated with the model evaluation metrics to have a quantifiable, rational, data-based evaluation criteria. To elaborate, KPIs like abandoned loan rate as well as application approval rate can be directly arrived at by the input data for the model and the predictions. Other KPIs such as customer acquisition cost perhaps may need additional data to make sense.

3.2.0.0. Setting the context for deciding metrics appropriate for case at hand

3.2.1.0. The Home Credit dataset has the target equal to 0 for clients who repay the loan on time and target equal to 1 for those that default. So this is a two state or binary classification problem.

3.2.2.0. The data is quite imbalanced because there is a high number of clients who repay the loan compared to clients who default.

3.2.3.0. While translating the model prediction to business outcome for Home Credit, there are two cases which result in a situation of loss.

3.2.3.1. Case 1 - The model has predicted the client will repay the loan but actually he has defaulted. This is critical as it results in loss of capital equivalent to defaulted credit to Home Credit.

3.2.3.2. Case 2 - The model has predicted the client will default but he can actually repay the loan back. Here, Home Credit faces the loss of a potential client and potential loss in return interest or lost business opportunity cost. Apart from this, there exists the fact that a deserving client is not getting a loan on account of the model prediction.

3.2.4.0. As the primary intent of using the model is to protect the interests of the credit lending agency, case 1 shall be a focal point in deciding the performance metrics.

3.3.0.0. Listing and assessing the possible Metrics appropriate for the context

3.3.1.0. Since the problem at hand is a binary classification problem, the following metrics are insightful -

★ Accuracy ★ Precision ★ Recall ★ F1 Score ★ ROC-AUC score

3.3.2.0. Evaluation of the suitability of these metrics to the Home Credit model is as below:

3.3.2.1. Accuracy is the most intuitive performance metric and it is simply the ratio of correctly predicted observation to the total observations. However, considering the imbalanced dataset, even a dumb model predicting every client as non-defaulter can bag an accuracy score of ~0.92 considering the training sample distribution, which is quite erroneous. Unless the imbalance is resolved, accuracy is a poor metric to use.

3.3.2.2. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, by the model. Its interpretation is fairly simple and intuitive. In context to the Home Credit dataset, this translates to ratio of correctly identified defaulters to the sum total of correctly and incorrectly identified defaulters, by the model. Precision is a good metric when the cost of a false positive is high. This metric shall address the case 2 scenario. As case 2 is not the dominant loss case, this may not be the primary metric.

3.3.2.3. Recall is the ratio of correctly predicted positive observations to the total positive observations, by the model. Its interpretation is fairly simple and intuitive. In context to the Home Credit dataset, this translates to the ratio of correctly identified defaulters to the actual defaulters, by the model. Recall is a good metric when the cost of a false negative is high. This metric shall address the case 1 scenario. As case 1 is the dominant loss case, this may be the primary metric.

3.3.2.4. F1 Score is basically the geometric mean of Precision and Recall. This score takes both false positives and false negatives into account. Though F1 Score is usually more useful than accuracy, especially for imbalanced distributions, Intuitively it is not as easy to interpret and understand as accuracy. Though this metric addresses both the loss cases, to translate into business impact it has to be viewed along with the constituent Precision and Recall scores. This can be considered as a secondary metric.

3.3.2.5. ROC-AUC Score is basically the area under the curve for a plot of True Positive Rate [TPR] or Recall on Y-axis vs False Positive Rate [FPR] on X-axis. An excellent model scores closer to 1 implying it has a good measure of separability. A poor model scores near 0 which describes that it has the worst measure of separability. In fact, it means it is reciprocating the result and predicting 0s as 1s and 1s as 0s. When an AUC is 0.5, it means the model has no class separation capacity present whatsoever and is basically a random model. AUC score covers both loss cases and the graph has very good interpretability. Hence ROC-AUC Score can be the primary performance metric. Also, the original Kaggle Home Credit competition had this as the evaluation criterion.

3.3.2.6. PR Curve is basically the plot of the precision (y-axis) and the recall (x-axis) for different thresholds, much like the ROC curve. Similar to the AUC of ROC curves, AUC-PR is typically in the range [0.5,1]. If a classifier obtains an AUC-PR smaller than 0.5, the labels should be inverted. But importantly, the Precision-Recall Plot is more informative than the ROC Plot when evaluating Binary Classifiers on Imbalanced

Datasets, especially when one cares more about positive than negative class[2]. The Home Credit fits the criteria as the dataset is highly imbalanced and as explained in the cases, predicting the defaulter (positive case) is the priority. Hence, the PRC graph shall also be plotted.

3.4.0.0. Bonus - Cases where the above mentioned Metrics shine

3.4.1.1. When the dataset is almost balanced, Accuracy is the one of the most widely used metrics owing to its high interpretability. A classic example of the balanced dataset is the Iris dataset.

3.4.1.2. Precision is best used when the cost of false positives is high as in the case of weather prediction for launching satellites. If the model predicts that it is a good day, but it is actually a bad day to launch the satellite (false positive) then the satellites may be destroyed and the cost of damages will be in the billions.

3.4.1.3. Recall is best used when the cost of false negatives is high as in the case of cancer detection. A false positive can be detected by further specialised tests but a false negative can be lethal by preventing timely diagnosis and thus, treatment.

4.0.0.0. **REAL WORLD CHALLENGES AND CONSTRAINTS**

4.1.0.0. Balance between Domain knowledge and Machine Learning expertise

4.1.1.0. Sound knowledge of the KPIs in the credit sector as well as the correlation between the ML model metrics and the KPIs is essential in order to not merely use the model as a binary segregator, but as a quantifiable means to monitor the organization's health. Moreover, the ML model should be interpretable and be able to achieve the intent.

4.1.2.0. As humans are filling the loan application form fields and there is possibility for error or falsification, a domain knowledge of financial industry and understanding of the demographics helps in addressing this aspect. Also, depending on the interactions with a potential client seeking a substantial sum, knowledge and experience in the financial domain might motivate investing time and efforts to convert the application which might not be fully implementable in a binary model.

4.1.3.0. ML expertise can help understand and tackle the quirks related to the problem at hand such as imbalanced dataset, missing values or additional data needed. Suitability of models with context to accuracy vs. interpretability also requires knowledge of ML. It always helps to have an understanding of the inner logic of the model and its parameters today to assess whether the same is valid tomorrow.

4.1.4.0. Finding the right balance between business utility, interpretability and fidelity, robustness of ML model is an important constraint which needs to be resolved before dwelling further.

5.0.0.0. **VARIOUS APPROACHES TO SIMILAR PROBLEMS AND REFERENCE**

5.1.0.0. Common approaches frequented for similar problems[7]

5.1.1.0. Upon reading blog posts, resource forums and research papers pertaining to modelling ML solutions to similar problems, the logic can generally be categorized in any of the three broad strategies -

5.1.1.1. Extensive feature engineering, creating new features, interactions and using these inputs on fairly standard classification models such as LR, DT and its variants and

ensemble models; with relatively lesser emphasis on model hyperparameter optimization.

5.1.1.2. Standard or minimal feature engineering and using many standard models with intensive hyperparameter tuning, usage of neural networks. Here, the majority of thought work is focused on optimising the model parameters.

5.1.1.3. A very few works have actually done both, extensive feature engineering as well as trying out a variety of models, each with hyperparameter tuning. Some have tried out DL models with varying results.

5.1.2.0. All of these approaches are in a way relevant to the problem at hand. Initial strategy shall be to understand the data with EDA and also try out the standard classification models. Eventually, after dwelling further, model-specific approaches can be formulated.

5.2.0.0. Citations, reference and further reading

[1] The 8 Most Important Loan and Mortgage Performance Metrics [Blog link - https://www.lightico.com/blog/lending-kpis-most-important/]

[2] The Precision-Recall Plot Is More Informative Than The Roc Plot When Evaluating Binary Classifiers On Imbalanced Datasets - Takaya Saito, Marc Rehmsmeier [Research paper link - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432]

[3] Home Credit Default Risk [datasource - https://www.kaggle.com/c/home-credit-default-risk/overview/description]

[4] Factors that affect loan giving decision of banks [paper link - https://deepnote.com/@rhishab-mukherjee/Loan-Prediction-Project-TermPaper-VPSOpiywSu6FZeN2fK8fug]

[5] Default Risk - Investopedia [Blog link - https://www.investopedia.com/terms/d/defaultrisk.asp]

[6] Home Credit Loan Default Risk by Winston Fernandes [Resource link - https://medium.com/analytics-vidhya/home-credit-loan-default-risk-7d660ce22942]

[7] Machine Learning in Banking Risk Management: A Literature Review by Martin Leo, Suneel Sharma and K. Maddulety [Resource link - https://www.mdpi.com/2227-9091/7/1/29]
