PREDICTING CREDIT RISK USING FEATURE ENGINEERING AND STATE OF THE ART CLASSIFICATION ALGORITHMS

SUDHIR TELKAR

Student ID: 975174

Under the supervision of

SUVAJIT MUKHOPADHYAY

Final Thesis Report

Master in Data Science - Liverpool John Moores University

DECEMBER 2021

ACKNOWLEDGMENT

This dissertation is based on predicting credit risk for Home Credit using feature engineering and data balancing techniques and a comparison of classifiers and ensembles to identify the best performant one. It also provides model interpretability techniques for credit scoring studies.

I am grateful to my thesis mentor, Mr Suvajit Mukhopadhyay, who has helped me developed my ideas and continued to inspire me with his insights and subject matter expertise to complete this thesis.
A special word of gratitude to Mr Manoj Jayabalan of John Moore's University for his weekly guidance and support.

I thank the team at Upgrad for their continued and timely support, mentorship and guidance through the program.

I would like to acknowledge with gratitude, the support and love of my family - my wife, Swapna, and my sons Soham and Swaraj. Your unwavering support and encouragement was pivotal in achieving my academic goal.

Finally, a sincere thanks to my family and friends for their continued support.

ABSTRACT

Credit default score risk assessment is a relevant and critical topic to banking and financial institutions who are motivated to reduce the cost of error in approving a loan and also want to be aggressive in their expansion plans. Home Credit is a non-banking financial institution that focusses on the demographic that has little to no credit history. Data about applicants, their loan history, previous loans and repayment history is available. The objective is to develop a machine learning model to assess credit default risk and use the same to identify clients who may have poor credit bureau information but who demonstrate a strong probability of loan repayment. There is much literature and studies on credit scoring and credit default risk prediction using statistical and machine learning models. Machine learning models are replacing statistical approaches so as to avoid data assumption related risks. While sophisticated models are being constructed, there is no one approach that addresses credit scoring domain. Thus, studies have used feature engineering as well as models to present findings that can be used by future research. The credit scoring domain datasets often have class imbalance on target variable and sophisticated models suffer from lack of interpretability. This study uses the Home Credit Loan Default dataset and uses methodologies to treat the class imbalance by using multiple sampling techniques. Feature engineering and selection approaches are implemented which identify the best performing and significant features that can be used during model training. The results have shown that SMOTE+Tomek is the best sampling technique and step_forward_feature_selection wrapper method is the best feature selection technique on the Home Credit Dataset. The study compares performance of thirteen classifier models and three heterogeneous ensembles to find out the best performance achieved for this use case. The results identify LightGBM as the best base classifier and its performance is comparable to the ensembles generated in this paper. Finally model interpretability techniques are applied on the best-in-class model to showcase the use of explainable AI techniques on the model and dataset so that business and research teams can benefit from the study and findings. The feature importance and significance from the model interpretability results show that synthetic features generated from manual feature extraction appear in top echelons of feature importance signifying success of the manual feature generation exercise.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADABOOST...	Adaptive Boosting
ANN...	Artificial Neural Network
AUC...	Area Under Curve
BPNN...	Back Propagation Neural Networks
EDA...	Exploratory Data Analysis
EMIS...	Emerging Markets Information Service
FSVM...	Fuzzy Support Vector Machine
GPU...	Graphics Processing Unit
JRNN...	Jordan Recurrent Neural Networks
MLP...	Multi-Layer Perceptron
RBF...	Radial Basis Function
ROC...	Receiver Operating Characteristic Curve
SOM...	Self-Organizing Maps
SPSO...	Switching Particle Swarm Optimization
SVM...	Support Vector Machine
UCI...	University of California Irvine
CBR...	Case Based Reasoning
RST–CBR ...	Rough Set Theory with CBR
RST–GRA–CBR...	RST, Grey Relational Analysis, and CBR
CART–CBR...	Classification and Regression Tree with CBR
MDA...	Multi Discriminant Analysis
GAM...	Generalized Additive Model

LDA...	Linear Discriminant Analysis
PLS-DA...	Partial Least Square Discriminant Analysis
ENN...	Edited Nearest Neighbour
SMOTE...	Synthetic Minority Oversampling Technique
ADASYN...	Adaptive Synthetic
MARS...	Multivariate Adaptive Regression Splines
EDDM...	Early Drift Detection Mechanism
DDM...	Drift Detection Mechanism
LSTM...	Long Short-Term memory
CNN...	Convolution Neural Networks
GRU...	Gated Recurrent Unit
RNN...	Recurrent Neural Networks
SHAP...	Shapley Additive Explanations
LIME...	Local Interpretable Model-agnostic Explanations
GBM...	Gradient Boosting Machine
XGBoost...	Extreme Gradient Boosting
CatBoost...	Category Boosting
HEOM...	Heterogeneous Euclidean- Overlap Metric

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Background

The need for Credit score analysis can be traced to the beginning of commerce probably around the same time as money lending and borrowing started. There have been numerous studies on credit scoring models and statistical and machine learning models exist which can measure the credit worthiness of the consumer. Many such studies exist but are limited to the particular dataset or the features therein. Since the Basel Committee on Banking Supervision released the Basel Accords, specially the second accord from 2004, the use of credit scoring has grown considerably, not only for credit granting decisions but also for risk management purposes. Basel III, released in 2013, render more accurate calculations of default risk, especially in the consideration of external rating agencies, which should have periodic, rigorous and formal comments that are independent of the business lines under review and that re-evaluates its methodologies and models and any significant changes made to them. Home Credit is a non-banking financial institution, founded in 1997 in the Czech Republic. The company operates in 14 countries (including United States, Russia, Kazakhstan, Belarus, China, India) and has over 29 million customers, total assets of 21 billion Euro, over 160 million loans, with the majority in Asia and almost half of them in China (as of 19-05-2018). Home credit group focusses on the demographic that has little to no credit history with the objective of leveraging credit scoring models that reduce the cost of error of approving a bad loan but at the same time allow for Home Credit group to expand their consumer base by approving loans to credit worthy consumers as correctly identified by the model. Towards this end, they have gathered a number of datapoints from their consumer base and the credit patterns (past and present) which help in generation of credit scoring models.

The study applies best in class data preparation and feature engineering techniques on the dataset to generate synthetic features and identify significant features using feature selection.

The research develops a state-of-the-art model which calculates the credit risk score for the Home Credit dataset and compares the results of the new models with other classifiers that have been run on the dataset and attempt to identify best in class solution for the same. The study applies model interpretability techniques on the best-in-class model so that loan managers and regulators can interpret the decisions taken by the model.

1.2 Problem Statement

Credit scoring domain is a relevant topic and of peak interest to banking and financial institutions. While the risk of credit default is a high risk for the banks, it is also equally important to not pass on a client / applicant who is able to repay the loan with an erroneous decision as that would be losing business to competition.

A literature review on the credit scoring domain reveals that methodologies to predict credit risk continue to evolve but there is yet no one approach that is well suited to the domain. Also, the number of defaulters is very few compared to the overall volume of loans so the credit scoring dataset domain is often fraught with class imbalance that needs treatment.

Research papers on this domain conclude that to arrive at the right methodology to predict credit risk, information from multiple sources is important. Client details, personal information, loan history, bureau data, historical loan data are information domains that impact and influence the loan approval decision. With so much information, feature engineering and feature selection is proving to be very important to be able to arrive at the right model.

The use case used by this study is based on a client demographic who do not necessarily have a lot of bureau information or loan history. Home Credit use case's objective is to construct a model that predicts capability of applicants to repay the loan so that loans are sanctioned only for those loan applicants.

To address the above, this study makes effective use of sampling techniques or class imbalance treatment approaches and perform feature engineering and feature selection on the datasets to identify features that are significant and important to predicting credit risk. These featured datasets are inputted to multiple machine learning models and their performances compared. The best performing classifier is then subjected to model interpretability techniques to explain the significant features that influenced the model's decision.

1.3 Aims & Objective.

Overall, summary of literature review determines that credit scoring model evolution is still taking place and the current trend is to develop new models using ensemble approaches. The Home Credit dataset consists of data imbalance which when managed is ripe for the application of feature selection/extraction techniques. The objective is to use these features and train models using state of the art classification/ensemble approaches and showcase the results and compare those with previous studies. This furthers the research carried thus far on the dataset and at the same time provide more insights on the Home Credit default dataset credit scoring best in class model by using model interpretability techniques. This is useful to loan management and regulatory personnel in their ability to understand the decisions taken by the model and provides insights to future researchers in this area.

Following are the research objectives that have been put together on the basis of the aim of this study:

- To suggest a suitable class imbalance handling technique which can be applied on the imbalanced Home Credit Risk Default dataset.
- To improve the performance of classification/ensemble models for credit risk scoring using feature generation and selection techniques.
- To generate ensemble models and compare their performance vis-à-vis known better performing classification models to highlight the performance of the models when applied to the Home Credit Default Risk data set.
- To apply explainable artificial intelligence methodologies to the best performing model and to be able to explain the black box model better to business stakeholders

1.4 Research Question

The credit risk default prediction domain comprises of huge datasets and multiple data points or features related to the application and customer. Also, the actual default cases are disproportionately low as compared to the overall volume of applications. While there are trends to identify more sophisticated models, there is no clear one model or algorithm that solves the credit risk prediction problem. The above facts about the domain bring to light the following research questions for this study:

Question 1: What is the impact of class imbalance on credit scoring datasets. Which sampling technique works best for this domain?

Question 2: Which feature selection techniques work best on credit scoring datasets? What is the impact of feature selection techniques on credit scoring datasets?

Question 3: What are the performance gains and falls of using base classifiers and ensembles on credit scoring datasets? What are the best performing models?

Question 4: How to explain the credit scoring models to various business stakeholders like loan officers?

1.5 Scope of study

1.5.1 Scope

The study includes a comprehensive exploratory data analysis, data balancing, feature engineering and identification of best performing supervised classification algorithm for credit risk prediction on the Home Credit dataset.

The study will apply oversampling, under sampling, hybrid sampling to the dataset and compare the performances of each sampling technique. The expected outcome is identification of the best sampling technique for Home Credit dataset as well as observations on why other sampling techniques did not perform well.

The study will use manual feature generation to generate synthetic features from the dataset to identify if these synthetic features influence the dependent variable better than the individual base features. The large number of features associated with credit scoring datasets will be reduced using feature selection methods. This study will employ wrapper as well as embedded methods for feature selection. The expected outcome is to identify which feature selection method results in higher model performance as well as to ascertain if the synthetic features generated as part of manual feature generation are in the higher echelons of feature importance.

The study will compare the performance of thirteen supervised classifiers on the data balanced and feature engineered dataset to identify the top three models. Hyper-parameter optimization on the top three models is performed to arrive at top three base classifiers. Heterogenous ensembles using bagging, stacking and blending techniques will be created. The expected outcome is to identify the top base classifier from thirteen classifiers used in this study and compare the performance of top performing base classifiers to ensembles to arrive at the best model.

Finally, the study will apply model interpretability techniques on the best performing model to arrive at explanations that can be used by loan officers to explain how the model arrived at credit worthiness for the applicant. The expected outcome is global and local interpretations using Tree SHAP and local explanations for default as well as non-default cases using LIME.

1.5.2 Out of Scope

The study uses the Home Credit Loan Default dataset provided by Home Credit and apply feature selection and class imbalance techniques on the same. Also, the model behaviour and performance comparison are limited to data from just this dataset. An opportunity exists to apply the same techniques on other globally available credit scoring datasets (Australian Credit scoring dataset, Taiwanese dataset, Austrian dataset) but the same is out of scope of this study. The processing is done in python. Any equivalent implementations of the code in any other language are out of scope.

This study is limited to identifying the best classifier amongst supervised base classifiers and heterogenous ensembles. Implementation of Un-supervised learning techniques is out of scope of this study.

1.5.3 Limitation of the Study

The three feature selection techniques used in this study take as input 74 features which is a subset of total number of features from the sampling to Home Credit dataset which comprises of 7 datasets. The total number of features are in excess of 300 features and the total number of loans in training dataset is in excess of 0.5 million. Application of feature selection techniques on this large dataset with comprehensive features requires stronger hardware which is not available to this study. This study reduces the features by using a RandomForestRegressor model output and the top 74 features are chosen by this study as input to feature selection.

This study is limited to using manual feature engineering techniques like polynomial and domain-based feature generation. Implementation of automated feature engineering using libraries like featuretools to generate more synthetic features and identification of whether those synthetic features appear as top significant features is a limitation of this study.

Application of dimensionality reduction techniques and using that data for feature reduction and comparison of the same is a limitation of this study. The study only uses manual feature

generation and feature selection techniques to reduce features and identify top significant features.

The previous loan application history data points provide loan application / revolving loan / point of sales history data for the application for past years. This data can be split into sub-datasets by each previous year and the model can be applied on each of the datasets to identify if there exists a model drift or data drift use case on the Home Credit use case.

1.6 Significance of Study

Credit scoring default risk model research is established as a clear and present requirement for banking and financial institutions in the world. The need for the same is to ensure that the cost of error is minimized. Another important factor is also the competitive banking space of these times where banks and financial institutions want to grab every opportunity to improve customer base and ensure all demographic that are capable of loan repayment have loans available, Though there have been many statistical models developed in the past and more machine learning models and now hybrid models, there still exists opportunity to improve on the model creation and comparison of its performance to classifiers, ensemble and statistical methods. Literature review over the years showcase that current trend in model identification is using hybrid /ensemble models though the comparison of performance of hybrid vs classifiers needs more study and results. This study uses the Home Credit Default Risk dataset which comprises of 7 datasets having class imbalance and on which there have been some studies where base classifiers have been compared. The significance of the study is the ability to further the existing research papers by applying class imbalance techniques, apply feature selection algorithms and train classifiers and ensembles and compare the results of the same. This result is useful to future research in the domain of credit scoring models as it adds cadence on performance of these classifier and ensemble models for further analysis. Another significant feature of the study is to apply model interpretability techniques on the best classifier/ensemble model. This is an area of study hitherto not applied to the Home Credit dataset and the results are useful not only to the machine learning community but also the business groups (loan management and regulators) who can use the study to apply and adopt more complex black box models and use model interpretability to make sense of the results. This bridges the gap as identified in the study carried out by (Lessmann et al., 2015).

1.7 Structure of the Study

This thesis is divided into multiple chapters. Chapter 1 describes the background of the studies and research done in the area of predicting credit default risk or credit scoring domain. It also outlines the pending problem statements based on the literature review. The section 1.3 explains the aim and objectives of the current study. Section 1.4 outlines the various research questions that this study aims to address. The scope and the significance of the study are covered in section 1.5 and 1.6 respectively.

The chapter 2 unveils the journey of machine learning for credit risk prediction. Section 2.2, provides insights into the domain of credit scoring, its consequences, influence of banking sector and geographies and different types of risks associated with the domain. Section 2.3 elaborates on the evolution of methodologies and techniques for of predicting credit risk. A review of papers and studies conducted through the years give insights into model evolutions and how other areas like feature engineering, class imbalance and model interpretability are influencing future work. Section 2.4 focusses on research conducted on credit scoring domain which has attempted to solve the class imbalance issue. A review of studies analyses the sampling methodologies implemented and impact of the same. Section 2.5 focusses on feature selection and the importance of identifying the right features. The paper analyses the different approaches and gaps in studies conducted in this domain. Section 2.6 focusses on studies that are using the ensemble methodologies to predict credit risk. Section 2.7 focusses on studies that have used the same dataset as this paper. Section 2.8 highlights that as models become more sophisticated, the interpretability of the models is adversely affected. A review of multiple explainable AI techniques is made through a review of research papers.

The summary of the literature review is discussed and concluded in section 2.9.

The Chapter 3 is focused around the research design and the proposed framework. Section 3.2 outlines the framework pipeline which is evaluated. Section 3.3 and 3.4 describe the dataset, the data pre-processing and class imbalance handling that is done as the first step. Section 3.5 and 3.6 present the details of the feature section and feature scaling methods which is used prior to the model development. All the various classification models which are in the scope of this study are detailed in Section 3.7 along with the reasons for picking these models. Section 3.8 elaborates the model tuning and evaluation steps for the shortlisted models. Section 3.12 covers the model interpretation techniques while section 3.13 lays down the expected outcome from this study. The whole chapter is also summarized in section 3.14.

The Chapter 4 is focussed on the design and execution of the study. Section 4.2 describes the datasets and salient features which are of interest. Section 4.3 describes the data preparation steps which are cleaning anomalies in the data, transformation of features by fixing outliers. Section 4.4 explains how Exploratory Data analysis is conducted using graphs and visualization techniques. The class imbalance is explored and correlation amongst features are analysed. Section 4.5 involves feature encoding. The dataset is then transformed with the implementation of various categories of class sampling methods to smooth out the unbalanced target class, along with the application of feature extraction methods for choosing only the significant features to make them suitable for modelling.

Section 4.6 involves implementation of thirteen base classifiers and evaluating the top three models which are subject to model tuning. The top three models thus are used to generate heterogeneous ensembles and the performance of top performing base classifier is compared the ensembles. Section 4.7 covers the generation of explanations for model prediction using global as well as local interpretability techniques. Tree SHAP and LIME are used to generate global as well as local explanations of the feature importance and the impact of the feature on determining the risk of default.

Chapter 5 presents the results from the research methodology applied to the dataset and discusses the findings and possible conclusions. Section 5.2 presents the result of the feature selection techniques applied as part of this study and the combination approach used by using the intersection of features of two feature selection techniques. Section 5.3 showcases the model performance by comparing the result of thirteen classifiers using the evaluation criteria. It then goes to show the results from hyper parameter techniques applied to the top three models. The study further goes to create heterogenous ensembles and presents the results. The discussion on whether heterogenous ensembles perform better than base classifiers is presented. Section 5.4 presents the results of model interpretation techniques applied as part of the study. The various graphs and visualizations from LIME and SHAP are shown in this section and a discussion on the findings is presented. In section 5.5, the research questions put forth by this study are discussed using the results from the methodology used by this study on the dataset.

Chapter 6 summarizes the conclusions from the various techniques and methodology applied by this study.

Section 6.2 discusses the conclusions and findings applied by this study. It concludes how data pre-processing is applied on Home Credit dataset. It discusses the feature generation approach used in this study and its success factor. Credit scoring datasets comprise of a large number of features and how feature selection techniques compared and which one performed best. It goes

on to discuss how the top performing base classifier was identified and how and why it outperforms the results of the heterogenous ensembles. Section 6.3 summarizes how this this study's findings and results contribute to credit scoring literature. Section 6.4 maps the limitations of this study to opportunities for future work in this domain to aid/inspire future work.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The “Literature Review” chapter presents a detailed study of various reviews/surveys and individual studies done in the area of credit scoring. It explains what credit scoring is all about and then moves onto section where it describes studies and bibliography reviews of research papers explaining how credit scoring models have evolved. It then gives an overview of each of these modelling methods and then explains various studies which have been done for them and emerging patterns that are useful and can be applied to this domain. It summarizes the research methodology and results of the studies and highlights gaps and achievements on the same. It the summarises all the research studies in the discussion section and at the very end a summary section for the literature review. This section focuses to draw insights from all the research studies and understand the gaps highlighted by these studies and identify opportunities for future implementations.

2.2 What is Credit Scoring

Banks and Financial institutions employ risk management strategies to ascertain probability of default when it comes to loan application processing and loan approval processes. The cost of a wrong decision during loan approval would result in a non performing asset for the institution. However, commerce expansion drives the banks to find newer and faster ways to ensure the right decisions are made so as not to lose out on potential business opportunities.
Credit scoring is a methodology used by Financial Institutions for estimating the risk associated with granting a loan or for determining the probability of default for determining the probability of default / non repayment.

The study by (LC, 2000) explains that the early models for credit ratings were made by teams of credit analysts. The analysts would follow a judgemental approach on the basis of the information on the following areas: the capital sought by the applicant entity, the market or economic conditions, the collateral that the entity is able to put up, the repayment capacity or history and the character of the applicant entity. Over time, these judgemental approaches became business rules which could be used to train more credit analyst teams and apply the approach. The volume of loan applications increase manifold when you look at banks

expanding their reach or credit card penetration or even the small businesses applying for loans. This business demand resulted in employing statistical and operational research-based models for credit scoring.

The study by (Agarwal et al., 2008) explains about the importance of predicting loan defaults and how the business domain is set to grow in the future. The paper highlights efficacy of statistical and operational models in credit scoring. The study concludes that the best models are built using not only loan applicant data but also firm specific information and historical payment information provide important features that can influence credit scoring decisions.

The study by (Mester, 1997) explains that credit rating is to give score of credit status and dept-paying ability of the rating target. Credit scoring greatly reduces the time needed in the loan approval process. The study highlights that the loan approval process at banks that used traditional approaches without credit scoring took three times the time as compared to banks that adopted credit scoring as part of their loan approval process. Credit scoring also brings improved objectivity (applying the same criteria to all loan applicants regardless of race, gender or other factors prohibited by law) to credit decisions. A credit scoring model makes it easier for lender or financial institution to explain the factor/weightage in the model and explain the outcome of the loan application better. The study goes on to explain how credit scoring was adopted by major banks and the application of the same allowed the banks to increase their businesses as they could process loan applications faster and improved credit scoring models improved risk determinations. The study concludes that with the securitization of small business loans there is a lot more business volume that can be influenced through credit scoring methodology.

Having understood what is credit scoring now, this study now explores what are the different types of models and understand how effective they are and how the models evolved over time.

2.3 Evolution of Credit Scoring Models

Early credit scoring models followed statistical tools. The study by (Hand and Henley, 1997) establishes that banks and financial institutions are using enhanced methods to control credit risk. Prediction of credit risk can be performed through procedures of credit scoring. The study

defines three generations of statistical models which included discriminant analyses, binary response models, and hazard models. Discriminant analysis was the first-generation studies. These studies provided rankings of risk failure and generated credit score. As part of this a threshold is defined and based on this the corporate is classified as non-default if it is below the threshold else it is classified as being default. Discriminant is based on the assumption that multivariate normal distributions are followed by the independent variables. The covariance matrix is defined as normal for the 2 groups and defaults are identical.

The second-generation involved binary response models. A binary response model applies a logistic or probity function and using explanatory variables estimates the failure probability. Binary response models did not require any assumptions around distributions of the predictor variables or probability default and so they had an advantage over the first-generation discriminant analysis models. These models could also test the importance of individual independent variables and also find the probability default in the next time period.

The third generation involved studies using hazard models also known as survival analysis. The conclusion from these studies is that it had better predictability as compared to the traditional single-period techniques and additionally it is useful to calculate the probability default over a period of time.

Studies indicated that Logistic Regression is quite a successful model for credit scoring datasets. Yet another variant is classification trees sometimes referred to as recursive partitioning algorithms.

(Baesens et al., 2003) is a benchmarking study on credit scoring datasets and machine learning models to explain the performance of various classifiers. The study uses eight credit scoring datasets and compares the performance of seventeen techniques such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and others. The Operational Research techniques or models comprise of variants of linear programming as well as non-parametric statistical and AI modelling techniques. The study concludes that the modern and sophisticated machine learning techniques such as SVM, are able to perform better especially in the field of data mining when compared with other models built using classical statistical approaches.

The study by (Lessmann et al., 2015) furthers the study by (Baesens et al., 2003) and explores the effectiveness of new classification and ensemble algorithms and compares their ability to effectively predict credit scoring models that can be used by the financial industry. Towards

this end, the study compared 41 new and alternative classification algorithms on eight real life credit scoring datasets. The study also includes heterogeneous ensembles and determines that when compared to individual classifiers, heterogenous ensembles perform better. The study also quells the concern that more advanced models may need strong human expertise by evaluating that advanced classifier do not need any more human intervention than simple classifiers. The direction thus set by these papers is that newer and complex models be built or identified and their performance compared to earlier models to attempt to identify a ubiquitous model that can be a gold standard for credit scoring.

The study by (Louzada et al., 2016) performs a bibliographic review of relating theory and application of binary classification techniques for credit scoring. This work evaluates 187 papers, from 1992 to 2015, and groups the general methodologies applied in the context of credit scoring and summarize the research findings. The study also performs an experimental evaluation of three very used credit scoring datasets, Australian, German, and Japanese. They evaluate two versions of each of these three datasets, one balanced, with IR =1, and other imbalanced, with IR =9. The study concludes the trend that studies are evolving to apply newer models over statistical models for credit scoring and based on comparison of general classifier deems SVM and FUZZY models as having the best performance. The gap that the study identifies is that newer research focusses on new models but the comparison to statistical models is not clearly identifying a single ubiquitous model that fits all flavour of credit scoring. Thus, research should continue and try and apply a combination of models which might yield better results.

In the paper by (Hsu and Hung, 2009) multiple classifiers (multiple discriminate analysis (MDA), canonical discriminant analysis (CANDISC) ,and SVM) are analysed and compared in their classifying performance on the credit scoring domain using the Taiwanese dataset. The study determines that SVM performs better than the MDA and the CANDISC classifiers. The authors however, have not performed any feature selection or class imbalance operations on the datasets which otherwise could have impacted the MDA and CANDISC classifiers. This study gives insight in the importance of class imbalance and feature selection techniques which could influence or impact the behaviour of models applied on the same.

The study by (Taha and Malebary, 2020) uses credit datasets and focus on model comparisons on the same. The paper introduces Light Gradient Boosting modelling and creates a hybrid

model with optimized features and applies the same on the credit dataset. The study goes on to apply a number of base classifiers like LR, RF, DT on the credit dataset to compare performances. The study concludes with the assertion that the hybrid sophisticated optimized light gradient boosting model achieves higher accuracy in predicting defaults as compared to other base classifiers.

The study by (Wang et al., 2019) uses a credit scoring dataset for China bank. The paper applies a combination of natural language processing (word2vec) to create vectors and the attention mechanism and applies the same to a deep learning model (LSTM). The study shows that this hybrid deep learning model that uses nlp vectors does not need artificial feature selection. The study concludes that the hybrid-deep learning model outperforms the standard feature extraction and modelling techniques used for credit scoring. The gap in the paper is that while this new approach using deep learning is only feasible with large datasets and significant features and may not be applicable to small datasets.

The study by (Dastile et al., 2020) presents a systematic literature review of papers on credit scoring including journals and conference papers between 2010 and 2018. The authors evaluate the most commonly used statistical and machine learning techniques applied at credit scoring. The performances (on German and Australian credit datasets) of statistical, classical machine learning and deep learning models that were reported in literatures were compared and the results surmise that an ensemble of classifiers generally outperform single classifiers. The authors also surmise that deep learning models (convolutional neural networks) in credit scoring literature showed better results compared to statistical and classical machine learning models. The limitations identified by the authors is better exploratory data analysis inclusion of macroeconomic variables lead racing correlation and collinearity between target variable and independent features

GAP

This section summarizes the evolution and trend of credit scoring models. The classical statistical models focussed on constructing improved discriminating rules but are dependent on distribution of features and that the target variable in credit scoring studies is small in nature causing problems like overfitting. As models evolved using machine learning, Logistic Regression, Support Vector Machine, Decision Trees, Random Forest based models are efficient to train and make no assumptions of the distribution of features. The gap identified

from these studies is that though the evolution has been to find a sophisticated model which addresses credit scoring scenarios, there is no one model that stands out. Based on the credit scoring dataset and features, different models behave differently. The summary from this evolution is that while model evolution/sophistication is the direction forward, bettering the class imbalance or improving features on the dataset is the way to improve credit default prediction.

2.4 Class imbalance learning models in Credit Scoring

An imbalanced dataset has volume of instances from one particular class much higher as compared to the volume of instances from other class. The minority class sample volume is very low in datasets and adversely impacts the prediction task in imbalanced datasets. In a credit scoring context, imbalanced data sets frequently occur as the number of defaulting loans in a portfolio is usually much lower than the number of observations that do not default.

The study by (Crone and Finlay, 2012) indicates that the low volume of minority class samples in imbalanced datasets adversely impacts the model and reduces the prediction capability of models. The model gravitates to sample classes with high volume and is biased to predict classes with majority data. The minority samples are ignored and treated as noise and cannot be classified correctly.

To handle class imbalance on datasets, oversampling is a technique that can be applied on the datasets. The paper by (Chawla et al., 2002) develops a oversampling method using k nearest neighbours called Synthetic Minority Oversampling technique also referred to as SMOTE. The methodology uses k nearest neighbours and generates synthetic samples. This technique addresses the shortcoming of the under-sampling technique where important information on datasets is eliminated in that process.

The study by (Bahnsen et al., 2014) uses credit scoring datasets and partitions the datasets into three bins and applies SMOTE and under sampling to the training datasets whereby creating two different training datasets. The datasets are put through LR, DT, RF classifiers and the results compared. The study concludes that under sampling performed better on their datasets as compared to smote treated dataset.

The study by (Fithria Siti Hanifah et al., 2015) uses the credit scoring dataset from a Indonesian bank to highlight the importance of class imbalance in credit scoring datasets and compares two classification models viz : one where class imbalance technique (SMOTEBagging) is applied on top of LOGR and the other being Logistic Regression (LOGR). The study successfully concludes that applying class imbalance sensitive algorithms to imbalanced datasets result in improved accuracy of machine learning models.

The study by (Brown and Mues, 2012) highlights the class imbalance problem on credit scoring datasets and conducts a study of various classification techniques on the five datasets comprising of 2 UCI Machine learning datasets(Australian and German) and three other datasets from the Benelux institution. The authors have further increased the class imbalance in the datasets by randomly under sampling the minority class of defaulters. The performance factors used have been Area under the receiver operating characteristic curve (AUC), Friedman's statistic and Nemenyi post hoc tests. The results from the study indicate that Random Forest and gradient boosting classifiers perform very well on pronounced class imbalance credit scoring datasets whereas kNN, QDA and c4.5 DT classifiers performed the worst. The study ends with future opportunities for trying ensembles and class imbalance learning strategies on credit scoring datasets.

The study by (Soares De Melo Junior et al., 2019) performs a benchmark comparison of 11 base classifiers vis-à-vis 3 ensembles vis-à-vis 5 imbalanced learning strategies on the Australian, German and Japan credit scoring datasets from the UCI Machine Learning repository. This study builds on the studies by (Brown and Mues, 2012) and determines that a more profound grid search can provide better results on the base classifier performance as compared to (Brown and Mues, 2012). The two hybrid ensembles of Random Forest Decision Tree ensemble (RNDF) and Extreme Gradient Boost (XGB) are the best performing classifiers on imbalanced Ratios. The study inspires to use the combination of ensembles and class imbalanced learning strategies on Home Credit dataset so as to compare the performance of the models on another different imbalanced dataset for credit scoring.

GAP

The previous studies in this domain have articulated the problem of class imbalance but are mostly using oversampling technique of SMOTE or oversampling techniques only. A study and

comparison of various data balancing techniques on credit scoring dataset and an explanation of their results and conclusion is missing.

To summarize this section, most credit scoring datasets suffer from class imbalance and the studies show that there exist a number of techniques that can be used to remedy the class imbalance. Class Imbalanced treated datasets improve model performances for some classifiers. A review of above studies provide inspiration to apply oversampling and under sampling-based imbalance techniques on the Home Credit dataset and to see the difference in model accuracy on that dataset.

2.5 Feature Engineering based studies

While innovation in creating new and sophisticated models to predict credit defaults has merit, applying feature engineering and selection techniques on datasets greatly improve the results. This study analysis studies made in the feature engineering and selection domain to identify best practices, gaps and get inspiration for future work.

The study by (Al-qerem, 2019) uses the credit scoring dataset from lending club and applies a combination of three feature selection techniques to the dataset. Information Gain (IG), Generic algorithm-based feature selection and Particle Swarm based optimization. A combination of classifiers (Naïve Bayes, Random Forest and Decision Trees) is used to identify the best performing model. The study concludes that application of feature selection technique to the datasets was very influential to improve accuracy of the models as opposed to using the dataset without feature selection.

The study by (Bellotti and Crook, 2009) compares the performance of the Support Vector Machine, Linear Discriminant Analysis and the Logistic Regression models on credit scoring datasets. The paper focusses on SVM as means to use in feature selection by using the magnitude of weights on features as a feature selection criterion. The study concludes that SVM performed better than other classifiers on their chosen dataset.

The study by (Chen and Li, 2010) uses SVM as the base classifier and applies four different feature selection algorithms on two credit scoring datasets imported from UCI. Decision Tree, statistical LDA, rough sets and F1 score are used to optimize the feature space on the datasets

and to remove redundant features. The study concludes that their implementation of a hybrid model is robust and optimal for credit scoring sets.

The study by (Maldonado et al., 2017) uses two credit scoring datasets related to Chilean bank's loans. The paper uses binary variables to identify relevant attributes and applies a budget constraint to identify low-cost accurate solutions. The feature selection approach and classification are achieved using two strategies using the support vector machine classifier. The study concludes that the SVM based strategy resulted in robust feature selection and achieved good accuracy when applied on the dataset.

The study by (Huang et al., 2007) uses the Australian and German credit scoring datasets from UCI. It applies three feature selection processes on the datasets using the following methodologies viz: Grid search, grid search plus F-score and a hybrid genetic algorithm-based approach for selecting the feature subset and optimizing the parameters. The paper uses SVM with the above-mentioned feature selection processes and compares the results to other models which do not use feature selection features. The study concludes that SVM with genetic algorithm-based feature selection is the optimal algorithm. The gap or shortcoming on this approach is the execution time and training time required by this approach.

GAP

The previous studies identify rightly that credit scoring datasets have large number of features and use tree classifiers and other models to identify significant features. There is very little work on feature generation using manual feature engineering or automated feature engineering to generate synthetic features and compare their performance. Also, wrapper-based feature selection approaches and their impact on credit scoring datasets is gap in existing literature.

2.5 Ensemble Learning studies

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Bagging, Boosting and Random Forest are different types of ensemble methods.

The study by (Wang et al., 2011) conducts a comparative assessment of the performance of three popular ensemble methods, i.e., Bagging, Boosting, and Stacking, based on four base learners, i.e., Logistic Regression Analysis (LRA), Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine (SVM). The study is made on three datasets out of which the first two are from UCI Machine learning repository (Australian and German) and the third dataset is about banks from China. The study results show that the following three ensemble methods greatly improve the results when compared to those of individual base learners. Taking credit scoring datasets as a premise, it is observed that the bagging ensemble approach performs much better than boosting ensembles. The study concludes that when evaluated under type I error, type II error and accuracy as evaluation criteria, Bagging and Stacking proved the best results. This study provides inspiration to apply ensemble methods to the Home Credit dataset and see the performance of ensembles on the same. The only limitation for the ensemble approach is the lack of interpretability of results.

The study by (Wang and Ma, 2012) starts with the premise that ensemble methods improve base learners and presents a new hybrid ensemble approach, called RSB-SVM, which is based on two popular ensemble strategies, i.e., bagging and random subspace and use SVM as base learner for enterprise credit risk assessment. Experiments based on the enterprise credit risk dataset, which collected by the Industrial and Commercial Bank of China, demonstrate that RSB-SVM gets the best performance among the eight methods, i.e., SVM, Bagging SVM, Random Subspace SVM, Boosting SVM, LRA, DT and ANN. And in practice, the non-linear kernel of SVM is more feasible than the linear kernel for credit scoring datasets

The study by (Abellán and Castellano, 2017) extends a previous work about the selection of the best base classifier used in ensembles on credit data sets. It is shown that a very simple base classifier, based on imprecise probabilities and uncertainty measures, attains a better trade-off among some aspects of interest for this type of studies such as accuracy and area under ROC curve (AUC). Recent works show that ensembles of classifiers achieve the better results for this kind of tasks. This study shows that a simple classifier based on imprecise probabilities, called CDT, improves to other more complex ones when it is used as base classifier, in an ensemble scheme, for credit risk assessment. By the results obtained from both measures, the CDT method can be considered giving better results when compared using the AUC and accuracy characteristics.

The study by (Ala'raj and Abbod, 2016) presents a new hybrid ensemble credit scoring model through the combination of two data pre-processing methods based on Gabriel Neighbourhood Graph editing (GNG) and Multivariate Adaptive Regression Splines (MARS) in the hybrid modelling phase. In addition, a new classifier combination rule based on the consensus approach (ConsA) of different classification algorithms during the ensemble modelling phase is proposed. Five of the well-known base classifiers are used, namely, neural networks (NN), support vector machines (SVM), random forests (RF), decision trees (DT), and naïve Bayes (NB). Several comparisons are carried out in this paper, as follows: 1) Comparison of individual base classifiers with the GNG and MARS methods applied separately and combined in order to choose the best results for the ensemble modelling phase; 2) Comparison of the proposed approach with all the base classifiers and ensemble classifiers with the traditional combination methods; and 3) Comparison of the proposed approach with recent related studies in the literature. The experimental results, analysis and statistical tests prove the ability of the proposed approach to improve prediction performance against all the base classifiers, hybrid and the traditional combination methods in terms of average accuracy, the area under the curve (AUC) H-measure and the Brier Score. The model was validated over seven real world credit datasets.

The study by (Marqués et al., 2012a) affirms the modelling trend that ensemble approaches offer better results when it comes to AUC and accuracy than base classifier performance. This study is a benchmark study to identify which base classifiers should be employed in each ensemble in order to achieve the highest performance. Towards that goal, the present paper evaluates the performance of seven individual prediction techniques when used as members of five different ensemble methods. The ultimate aim of this study is to suggest appropriate classifiers for each ensemble approach in the context of credit scoring. The statistical tests and experiment results show that the C4.5 decision tree constitutes the best solution for most ensemble methods, closely followed by the multilayer perceptron neural network and logistic regression, whereas the nearest neighbour and the naive Bayes classifiers appear to be significantly

The study by (Marqués et al., 2012b) affirms the modelling trend that ensemble approaches offer better results when it comes to AUC and accuracy than base classifier performance. In continuation with the study trends that show that classifier ensembles generally result in more accurate models than single prediction models, this paper introduces a integration of ensembles

thereby creating composite ensembles that leverage strategies for diversity induction. The attribute selection methos of rotation forest and random subspace are leveraged in tandem with AdaBoost and bagging ensembles to generate algorithms that can resample data and result in composite ensembles which can be used to improve prediction performance. The study concludes with the results that this multi-layer composite ensemble approach results in better results than a traditional ensemble. This study inspires further work to focus on having multiple iterations or combinations of classifier to see if models can be improved. At the same time, it does not address the fact that composite ensembles result in poor interpretability of model.

The study by (He et al., 2018) chooses a credit scoring dataset and applies data pre-processing, and class imbalance techniques which treat the class imbalance on the target variable and allow for better performance by machine learning models.. The study employs Balance Cascade approach. For model generation, the paper uses two tree-based classifiers viz: XGBoost or extreme gradient boosting and Random Forest. These classifiers are used as base classifiers for a three-stage ensemble machine learning model construction. The paper uses particle swarm optimization algorithm for parameter optimization of the base classifiers and includes the use of stacking to generate results for prediction. The study concludes by stating that the performance of the ensemble method is much better than the other base classifiers.

The study by (Moscato et al., 2021) focuses on credit risk associated in the data domain of Peer to Peer lending and compares their results to state of the art papers on the same domain.
The author's approach is to design a benchmark for credit risk prediction for social lending platforms, also be able to manage unbalanced data-sets. The study compares behaviour of classifiers (Logistic Regression, Random Forest, Multilayer Perceptron) in combination with class imbalance techniques (Random Oversampling, Random Under sampling, SMOTE). The three best models identified from above are subjected to multiple explainable AI techniques (LIME, Anchors, SHAP, BEEF and LORE) and the performance of the Explainable AI techniques is compared. This study provides good inspiration to apply a combination of class imbalance and classifier performance comparison on the Home Lending dataset to assess the best performing combinations. The comparison of explainable AI techniques is another direction that has hitherto not been explored.

The study by (Xia et al., 2017) chooses a credit scoring dataset and sets out to compare the performance of base classifiers vs ensemble techniques on the dataset. The study performs

exploratory data analysis and data pre-processing to cleanse the data and then subjects the cleaned dataset to embedded feature engineering using a model to assess the feature importance and select features on the dataset that can be submitted to the model. The study chooses XGBoost or extreme gradient boosting as the sequential ensemble machine learning model on the dataset. The paper uses Bayesian hyper-parameter optimization for tuning the hyper parameters of the model. The paper also uses random search, grid search and manual search hyper parameter tuning methods. The model evaluation measures employed by the study are: accuracy, error rate, the area under the curve (AUC) H measure (AUC-H measure), and Brier score. The study concludes by showcasing that the ensemble method built using XGBoost outperforms base classifiers. The study also presents a comparison of various hyper-parameter tuning approaches on credit scoring datasets and determines that Bayesian hyper-parameter optimization outperformed all the other approaches.

The study by (Zhang et al., 2010), employs two credit scoring datasets from the UI Machine Learning Repository and builds a new credit scoring model which is quite novel by using the decision trees classifiers in a bagging ensemble with vertical distribution. In a traditional bagging ensemble, classifiers are trained with sample subsets and every classifier has the same attributes. The novelty of the bagging method used in this study is the usage of classifier groups and subjecting them to train samples but only a subset or combination of predictive attributes. The study concludes that this method ably avoids overfitting of models and the analysis results show that the prediction accuracy obtain through the usage of this method is much better than performance of individual classifiers. One of the gaps in the study is lack of usage of any class imbalance techniques which would have further given insights into how low samples of target variable affects study output

GAP

While previous studies show a trend in the adoption of ensembles for better predictability and for improving the performance of weak learners, this comes at a cost of interpretability of the model. How does a strong base classifier learner generated on a data balanced, feature engineered dataset compare to ensembles is a gap identified from studies?

In summary, ensemble machine learning models are a resilient method and avoid over-fitting easily on the datasets and also perform well on high dimensionality in datasets. The studies indicate that ensembles model-based prediction provides better scores than individual

classifiers. However, the gap identified is that the improvement in scores should be offset against the cost of using ensemble models. Some of the limitations from ensemble models are that complexity of classification increases thereby driving up cost and time for model execution. Some ensembles are sensitive to outliers and they introduce a loss in interpretability of the model.

2.6 Model Interpretability based Studies

As models evolve and become more sophisticated, there is a noticeable trend that the model interpretability suffers. Also, government and regulatory demands in some countries mandate banks and financial institutions to explain the rationale for decisions and predictions so as to prove there was no bias in their process. This prioritises the need to explain decisions generated by non-transparent models. Model interpretability or Explainable AI methods exist are more and more are being adopted by banks and financial institutions. This literature review analyses some studies in this domain and how it impacts credit scoring domain.

The paper by (Lundberg and Lee, 2019) acknowledges the trade-off between model interpretability and accuracy in credit scoring datasets and introduces the explainable AI methodologies that can help interpret predictions made by the model. The study describes LIME and SHAP methodologies in detail and goes on to describe multiple estimation methods using SHAP. This paper can be a good reference point for implementing SHAP for future studies.

In the study by (Demajo et al., 2020), the paper uses the HELOC and Lending club datasets for model building and analysis. The baseline is the study by methods using Dash et al. [25] where the BRCG model was put forth as the benchmark model. The paper by Demajo builds a model using XGBoost algorithm and is able to keep a good balance between Type-1 and Type-II errors. The study goes on to compare the performance of three model interpretability techniques on this model. Global feature explanations are performed using a combination of SHAP and GIRP explainable AI methodology. Local Feature explanations are put forth using Anchors and local instance-based explanations are rendered using the ProtoDash methodology. The study concludes with the comparison of the three model interpretability techniques and influences future work by showcasing the comparison of three model interpretability techniques.

The study by (Dastile and Celik, 2021) converts tabular datasets into images and employs models built using 2D CNNs on said credit scoring dataset. The model accuracy is good and a state-of-the-art deep learning model performed better than other models on credit scoring datasets that are publicly available. Model interpretability techniques like Saliency Map, SHAP, Grad-GAM and LIME were applied on the best performing deep learning model. The comparisons of the model interpretability techniques were performed using quantitative measures and also sanity checks. The study concludes by showcasing SHAP as the most efficient model interpretability techniques used on the dataset. The purpose of the XGBoost model was to evaluate quantitatively the performances of the explanation methods.

The study by (Qi et al., 2021) chooses the Lending Club credit scoring dataset and attempts to apply a classifier model and showcase the efficacy of model interpretability on the credit scoring dataset. The dataset provides comprehensive features that can be presented to the model and provide an opportunity for the model interpretability technique to review the feature importance's. The machine learning model employed by the study is built using CatBoost. The study showcases that the model proves to be accurate on the dataset. SHAP as a model interpretability method is applied here and the resultant global feature explanations showcase that from the comprehensive features submitted to the model, some of the features have no influence on the outcome. This study showcases the efficacy of the model interpretability method on credit scoring dataset and also highlights the importance of feature engineering and selection on credit scoring datasets. This inspires future work to perform a combination of feature engineering and selection in combination with sophisticated models and model interpretability and define how that impacts credit scoring domain-based datasets.

GAP

There are no model interpretability results available on the Home Credit dataset which can better explain the feature significance on the dataset.

In summary, a review of studies in this section determine that model interpretability is a very important concept driven by government and regulations in the domain of credit scoring and credit default prediction. Several methodologies of explainable AI such as local feature based or global feature-based explanations are applied in the studies and impact explained. The studies indicate that the credit scoring dataset used in combination with models and explainable

AI determine best outcomes. The Home Credit dataset chosen in this study does not have any existing literature where the impact of model interpretability can be explained.

.

2.7 Home Credit dataset-based studies

The study by (Qiu et al., 2019) uses the Home Credit dataset from the credit scoring domain and compares classifiers to arrive at a conclusion on which classifier fits best on this dataset. The study applies cleanses the data and applies feature engineering techniques to the dataset and compares the performance of three classifiers viz. Logistic Regressions (LOGR), Random Forest (RF) and Light Gradient Boosting Machine (Light GBM). The study concludes that based on accuracy and AUC score the Light GBM model scores much higher than any of the other classifiers used in the case study. Further direction on this study invites the approach of applying class imbalance techniques on the Home Credit dataset and to see how the other base classifiers perform on the same.

The study by (Tounsi et al., 2020) uses the Home Credit dataset from the credit scoring domain and applies three classifiers to the same and compares the performance of the same to arrive at the best performing model on the same. The study analysis the dataset and applies data cleansing and then subjects the dataset to three classifiers viz: (XGBoost, CatBoost and LightGBM). The study concludes that LightGBM seems to be significantly faster than the other gradient boosting methods when compared using the ROC and the accuracy characteristics of other models.

GAP

There is no comprehensive comparison of classifiers and ensembles on the Home Credit dataset to identify top performing model. Model interpretability results are missing in studies performed on this dataset. While class imbalance has been rectified, a comparison of different class imbalance techniques is missing.

In summary, for the Home Credit dataset, there are some studies which have performed feature engineering and compared classifier performances thus far. The gap identified is that there are no class imbalance remedied studies which can compare performance of models or any explainable AI based studies. That serves as inspiration for future work that can be done on the dataset.

2.8 Discussion

The literature review highlights the fact that banks and financial institutions are very competitive to extend their reach and are investing heavily for better ways in which the probability of default on loans can be minimized and calculated in as fast a manner as possible. Towards this end, credit scoring as an area is very relevant and much sought after for impartially influencing the loan approval process.

Credit scoring methodology matured from being a judgemental process to being calculated using statistical models. These models reliably calculated the credit risk process. However, while statistical analysis is fast and easy to implement, it is fraught with issues of overfitting because the target samples in credit scoring datasets are small and the normality assumptions associated with statistical methods especially around multivariate normality for independent variables are often invalid.

Research papers evolved and focussed on Operational Research models and newer Artificial Intelligence techniques which built more sophisticated models that are able to manage the complexities on data distributions in credit scoring datasets and automatically extract knowledge from training data. While statistical models do well on linear datasets, research papers conclude that machine learning and artificial intelligence base models perform much better when dealing with nonlinear datasets. However, the newer models take a lot of time during the training stage and require large volumes of data which the statistical models avoid. While both methodologies have their pros and cons, there is no one single ubiquitous model or technique which is the gold standard for credit scoring domain.

Credit scoring datasets used in the studies focussed on the literature reviews have mostly been based on the UCI Machine learning datasets or other such synthetic datasets. This is not uncommon as Credit scoring involves confidential and private data about the clients of the financial institutions and the data cannot be made public. The dataset chosen for this paper's study is credit dataset from Home Credit which is one of the newer datasets available. This literature review highlights two studies on the Home Credit dataset wherein the studies employ research methodology of comparing classifiers to determine which is the best performing classifier.

A common problem with all credit scoring datasets is of class imbalance. Newer research papers identify that application of credit imbalance techniques on the datasets significantly influences the accuracy and efficacy of the models applied on the dataset. Also, the datasets are comprised

of many features and selection of the right features for model training impacts the performance of the model. Thus, the trend towards credit scoring model evolution is to focus on class imbalance techniques and feature selection techniques in addition to comparison of base classifiers and models. This literature review acknowledges this trend and determines that hitherto, the same has not yet been applied on the dataset chosen for this study.

The literature review determines that newer intelligent models are more accurately able to determine credit scores than statistical models, but do not require a higher human effort for the same. This is positive news for the adoption and implementation of newer models but brings to fore gaps with the newer models. In addition to higher training time and large volume of data required by newer models, the interpretability of newer models is greatly reduced. This problem is further exacerbated by newer trends in the banking and financial business where it has become a regulatory requirement to be able to explain why a model decided on an outcome and to interpret its results. Model Interpretability or Explainable AI techniques significantly help towards improvement of the interpretability of the model and this literature review highlights a number of studies where multiple Explainable AI techniques have been applied on credit scoring datasets and their performance compared.

This literature review highlights the current pattern of applying ensemble modelling approaches towards credit scoring. A number of studies have concluded that an ensemble approach through layering approach can improve on the shortcomings of one base classifier and positively influence the performance of the models.

2.9 Summary

This literature review on credit scoring research papers has started out by explaining what is credit scoring as a domain and why is it relevant today. It has appreciated studies that help explain how the credit scoring model evolved from a judgemental approach to more statistical and artificial intelligence-based models. The constant driving factor for banks and financial institutions has been the ability to arrive at a generic model that can improve the accuracy of predicting scores and at the same time be fast in its execution. The literature review highlights gaps in studies where though newer and intelligent models where better than classic statistical models in their prediction scores, other studies have proved that applying a much broader approach to class imbalance rectification strategies and feature selection techniques resulted in much better results on base classifiers than just new model generation without application of those techniques. The study also acknowledges the business domain need for model

interpretability and showcases the methodologies used and the current trend of applying explainable AI on credit scoring model creation. This literature review highlights the Home Credit dataset based studies and identifies gaps where the current trend of class imbalance and explainable AI techniques are not yet applied on this dataset. This study appreciates the research into ensemble model building techniques and the efficacy of the same on other datasets.

Overall, the literature review highlights the trends and patterns towards credit scoring model creation and the newer techniques that are emerging for the credit scoring domain. Given that a generic gold standard model or technique still does not exist for credit scoring domain, this literature study identifies gaps and approaches that can be applied on the Home Credit dataset which only help further the research done this far and helps/aids future research in this domain.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The various approaches and procedures utilised to present the study in an attractive manner are referred to as research methodology.

In this section, the paper elaborates on the methodologies and best practices employed to meet the aims and objectives of the study. The aim is also to improve the outcomes of models using class imbalance and feature selection techniques. The final outcome is expected to elaborate the best performing model with the adequate explanation techniques. Overall, the study advises significant features along with best fit and interpretable classification model which helps the Home Credit use case. This chapter gives an overview of the complete research methodology for usage of machine learning to predict credit risk.

Research techniques are classified into various categories based on the goal, objective and the type of information pursued. Among the goal-based classification types, an ***Applied Research*** involves making use of a scientific study to solve a practical problem. The goal in such research methodology is to find a solution to the issue at hand by first defining a hypothesis and then testing it using trials. Since the aim of this study is to ascertain a consistent classification model for classifying credit default cases by comparing different models, this study is classified as *applied research*. Among the objective-based classification types, ***Explanatory Research*** explains the phenomena of the possible relationships among things and also explains the cause-and-effect relationship among them. Since the aim of this study is to help interpreting the best performing model with the adequate explanation techniques, this study is classified as *explanatory research*. Among the information-based classification types, ***Quantitative Research*** is an analysis of the phenomenon using quantifiable data generated or gathered from online surveys/polls, questionnaires, etc., and then implementing statistical analysis on the data**.** This study makes use of the Home Credit datasets, apply sampling and feature engineering, apply multiple classification techniques and perform a comparison study of the result to conclude the best performing model. Hence this study is classified as q*uantitative research*.

3.2 Research Approach

Research has been carried out in the 6 phases. Each of the phase and steps are explained here

1. The following datasets for the Home Credit domain are loaded:
    a. Current loan data (application train/test)
    b. Previous loan data (previous application)
    c. Bureau information (bureau and bureau balance)
    d. Loan and instalment history (pos_bal, credit_card_bal and instalment history)
2. The above datasets are reviewed
3. Application train and test data is split. This is done at this step to avoid overfitting or data leakage at a future step
4. Data Pre-processing is applied on all datasets except the test dataset. Missing values, outliers, data columns, formats are treated. Collinearity of datasets is analysed
5. Data is analysed using univariate, bivariate analysis. Features are encoded and normalized
6. Class Imbalance is treated by comparing five data sampling techniques. The balanced dataset is sent for further processing
7. As part of feature engineering, the datasets are merged in three combinations and derived features are constructed using domain knowledge and polynomial feature generation approaches. Three feature selection techniques are applied and three feature engineered datasets are created. The best performing dataset is used for model creation and tuning.
8. For the above dataset, thirteen base classifier models are constructed and trained using stratified k-fold validation. The top three performing classifiers are used to create heterogenous ensembles. The base classifier performance is compared to ensembles.
9. Trained Models are ready for evaluation
10. The test dataset is processed and prepped for use
11. Feature engineering and selection are applied on test dataset
12. The top three models applied to each of the datasets are evaluated using the following metrics viz: accuracy, AUC, Recall, Precision, F1, Kappa
13. The best performance received for the model and featured dataset is selected
14. The features of the best performing model are put through explainable AI techniques like SHAP and LIME
15. The Research observations and conclusions are summarized.

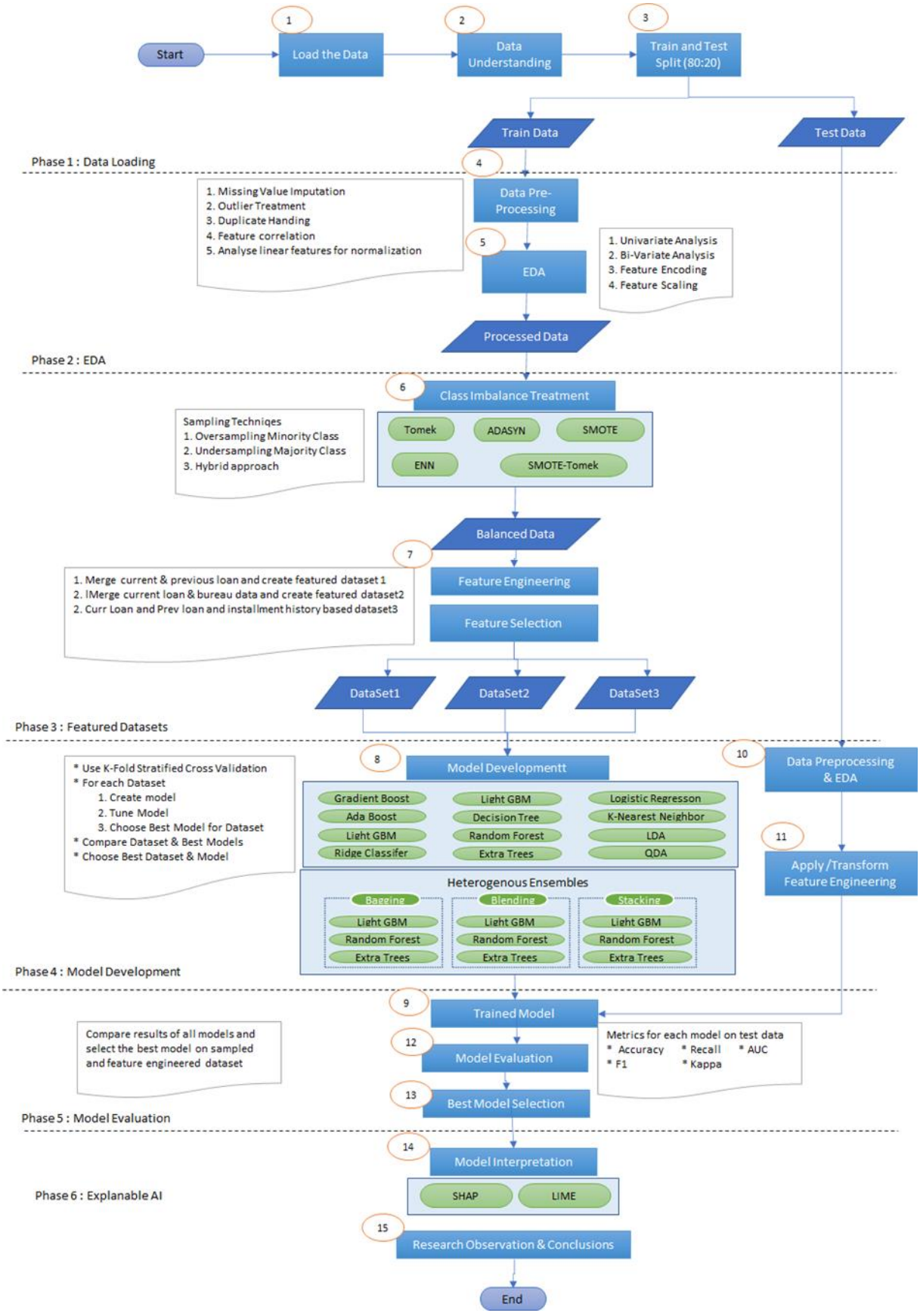


Figure 1.1: Flow chart of research methodology

3.3 Dataset Description and Analysis

Home Credit uses a variety of data like loan application information, client demographics, type of loan availed, credit bureau information, transactional to predict their client's repayment abilities.

The dataset provided by Home Credit comprises of seven sources of data.

- Application train/test: Information about each credit application at Home Credit. This becomes the main training and testing dataset. Each row in the dataset represents one loan application and is identified by the feature sk_id_curr. There is a total of 307,511 rows of loan information and sidecar each row has 122 variables. The target variable defines if the client had payment difficulties meaning he/she had late payment more than X days on at least one of the first Y instalments of the loan. Such case is marked as 1 while other all other cases as 0.
- Bureau: This dataset contains information on previous loans that the customer has availed from different banks & institutions. Every previous loan is represented by a row in bureau and one loan from application data for a client may have multiple rows in previous credits. The sk_id_curr feature from the application train dataset has a relationship with the
- Bureau balance: This dataset contains monthly balances of loans in Credit Bureau. Every row represents 1 month of every previous balance submission and for a single previous loan, there may be multiple rows representing months of credits for previous loans.
- Previous or historical loan application: historical applications related to credits applied to Home Credit by customers having entries in the current loan application data. Every current loan in the application train dataset can have multiple previous loans. The previous application is represented by one row and is identified by the feature sk_id_prev.
- Point of Sale cash balance: Month wise information on previous loans availed by customers for point of sale or cash credits.
- Revolving Loans or Credit card balance: Month wise information on previous loans availed by customers for revolving loans or credit card. This type of financial transactions can be referred to as revolving loans and this dataset contains three million rows which is significantly lower than the volume for cash credit transactions. This

behavioural trend indicates that Home Credit client demographic operate in a market where cash is preferred over credit cards.

- Instalment's payment: historical payment lineage for previous credits at Home Credit. The domain insight here is that the first few instalments are more important than last instalments. The behavioural trend of the applicants indicates that if a loan applicant is late once, she/he might be late from then. Another domain and data insight are the late days period for instalment payment as treated by Home Credit's collection team. As the late days period increases the level of follow up and chase by the Collection team changes.

The relationship between different datasets is explained below:

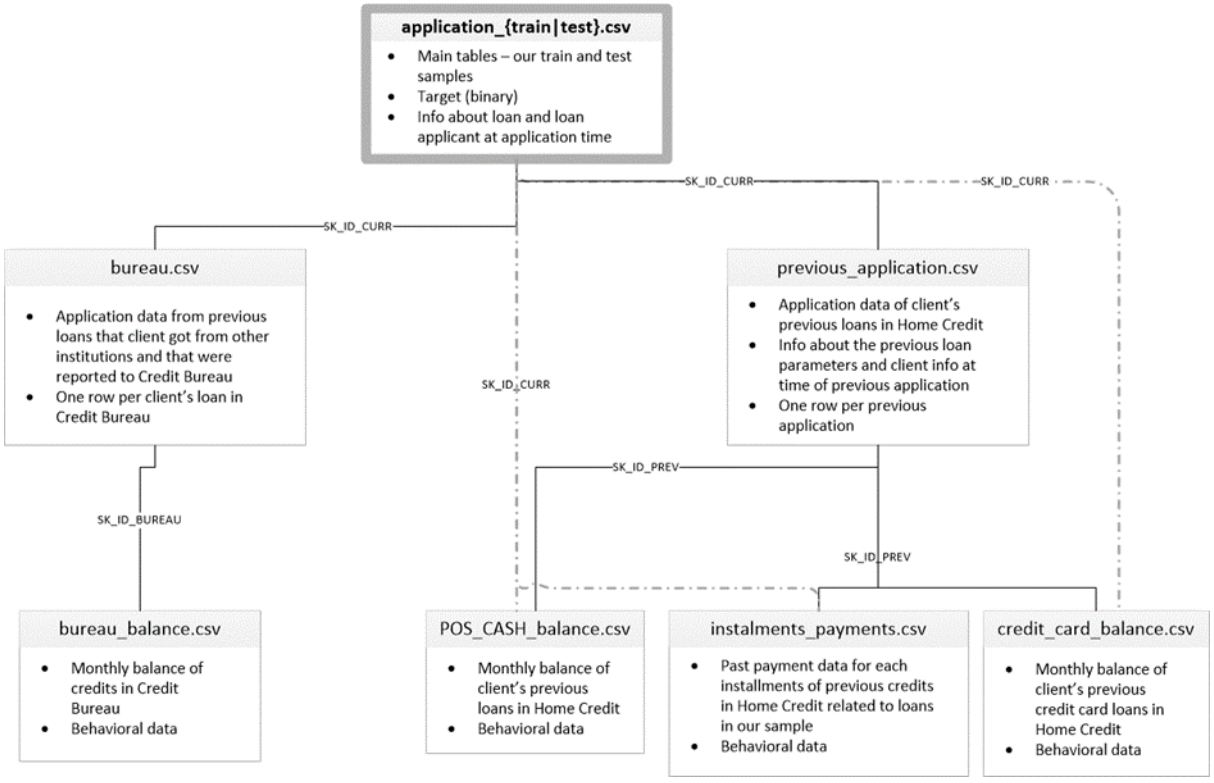


Figure 2.2: The relational data model - Home Credit Default Risk

3.4 Data Pre-processing

Each of the seven datasets related to Home Credit are loaded and subject to the following data processing steps. The dataset is introspected to check the column sizes, types, mean, average values etc to get some understanding of the dataset before further pre-processing. Post the dataset inspection, the following pre-processing steps are performed equally on each of the datasets.

- Column nomenclature and data format checks: After data loading, the study as a best practice refers to the dataset feature description data and check for inconsistencies in the dataset for improper column names or for erroneous data formats. Such errors could be introduced through either human error during data entry or creation and also through issues during data loading or transmission. Such issues if left untreated hinder the model building at later stages.
- Duplicate Data removal: Duplicate data in the dataset if left untreated results in models suffering from overfitting. The datasets that the study operates on uses sk_id_curr as the identifier. Using that identifier id, a check is made on the application and test data to see that duplicate data does not exist on the dataset. Duplicates, if found are removed.
- Missing Values Imputation: A data science best practice is to analyse for missing values and to treat them so as to render the dataset optimal for machine learning model execution. Skewed distributions and erroneous outcomes are the result of untreated missing values on datasets. If the column contains a low percentage of missing values, the same can be left untreated and model building can still continue. For columns with high percentages of missing values, a best practice is to impute them with either the mean/mode or median value of the column depending on the type of column and business domain understanding and requirements. If imputation is not possible, the columns with high missing values should be dropped from model building.
- Outliers Treatment: Outliers, in statistics, are datapoints that differ significantly from other observations. Causality of outliers could be due to variability in measurement or due to experimental errors. Outliers can cause serious problems in analysis and the results of the models. Techniques such as interquartile range distributions, binning and Gaussian distributions can be applied to treat the date and count of days fields in the datasets.

- Analyse date and count of days related attributes on the datasets and the distribution of data and normalize the data so that it can be used during feature engineering and extraction. Normalize any erroneous dates and treat negative values of count of days.
- Correlation: Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It provides an idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. Model inference improves significantly by elimination of correlated variables. This practice improves performance and storage of the model and features.
- Multi-Collinearity: Multicollinearity occurs when two or more independent variables (also known as predictors) are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model. Multi-collinearity reduces models' predictive power as it reduces the precision of predicted coefficients. As a best practice, elimination of multi-collinear data is highly recommended.

This study implements the above data cleaning strategies to ensure that the datasets are in peak optimal state for future processing and to derive inferences.

3.5 Exploratory Data Analysis (EDA)

The process of identifying patterns, trends anomalies and relationships within the data using statistics is called exploratory data analysis (EDA).

Each of the datasets that have gone through data preparation are analysed to get insights and inspiration for feature engineering later.

As part of EDA, analysis to determine relationship between dependent and independent variables, this study employs statistical techniques such as univariate, bivariate and multivariate analysis using histograms and box plots. Bi-variate analysis involves two variables and what relationship exist between them. Bivariate analysis can be performed by using plots like scatter plots to chart one variable against another and determine the relationship between them. Multivariate analysis involves more than two variables and relationship which exists between each variable and multiple variables. For this analysis, Heat Map can be used to show the relationship between target variable and the other independent variables This analysis helps identify the patterns of data within the dataset which affects the credit default prediction.

Additionally, the below treatment is applied to the datasets to render them optimal for feature selection and model creation.

- Feature Encoding: Machine learning algorithms prima facie operate on numerical values. Datasets have a number of categorical values which are either ordinal or nominal. The process of converting the categorical values into numerical values is feature encoding. For ordinal categorical feature conversions one-hot encoding is the best method whereas for nominal categorical features label encoding can be used.
- Feature scaling is a method used to normalize the range of independent variables or features of data. This process is used to effectively manage those features that suffer from hugely changing values, magnitudes or units. Leveraging scaling ensures that the model does not create bias towards features with extremities. This study presents two methodologies for applying scaling on the datasets.

  Standardization: Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

Equation 1.1: Standardization

μ = is the mean of the feature values

σ = is the standard deviation of the feature values

Normalization:

Normalization, also known as Min-Max scaling, is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Equation 2.2: Normalization

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator is 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1

- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

3.6 Class Imbalance Techniques

Most credit scoring datasets suffer from class imbalance and there are various types of sampling methods which can be used to handle this class imbalance. Under-sampling methods reduce the instances of the majority cases while over-sampling methods either duplicate or introduce new instances of the minority cases to balance the data. There are also hybrid techniques that are actually amalgamation of over and under sampling methods. Certain algorithm-based sampling techniques are also available to balance the imbalanced class dataset. SMOTE is a Synthetic Minority Oversampling Technique which is used to handle the class imbalance in the dataset. It is based on K nearest neighbour wherein new data points are created from existing minority data points along the line segments joining the minority class K nearest neighbours. Based on the amount of oversampling the k nearest neighbours can be randomly chosen. The algorithm takes the difference between the sample and its nearest neighbour and multiplies it by a number between 0 and 1 to create a new synthetic data point.

Section 2.4 of this study's literature review section evidences instances where SMOTE technique has been successfully tried for other credit scoring datasets. This study compares performances of 5 sampling techniques and how they perform on the credit scoring dataset

3.7 Feature engineering.

Feature engineering is the process of using domain knowledge to extract features from raw data. A feature is a property shared by independent units on which analysis or prediction is to be done. Features are used by predictive models and influence results. After data pre-processing and analysis, it is a best practice to leverage domain expertise and transform and translate features to a form that positively influences models and algorithms. As is the case with our chosen dataset group, there are multiple datasets available with current loan, previous loan history, bureau information, loan history and also instalment payment history of the client. This study employs feature transformation techniques where the datasets are merged and certain features are altered or aggregated or group to enhance the meaning and provide more cadence

and clarity during model execution. This aggregation or alteration ensures to maintain the integrity of the original dataset.

One approach towards feature engineering is through manual feature construction where the below mentioned two simple feature construction methods are used:

- Polynomial features: This involves generation of new features that are a combination of multiple individual variables. “Interaction items” are features that capture interaction between two or more variables. Another way to generate polynomial features is to generate datapoints that are powers of existing features.
- Domain knowledge features: As part of this feature generation approach, features from previous loan dataset as well as loan history datasets are merged with current loan application dataset and make those features available as loan repayment behaviour of the loan applicant. Also, features from bureau and historical repayment history are used. This approach is to construct features through merging of datasets which influences or impacts positively the abilities of the model to predict default.

A second approach to feature engineering would be a combination of manual and automated feature engineering techniques. Domain knowledge features are constructed and feature tools is leveraged to generate features on the dataset.

3.8 Feature Selection

Credit risk evaluation is a multidimensional and imbalanced problem, mainly based on a large volume of historical data such as: job status, credit history, personal account status and so on.

While using all relevant features increases the model performance and coverage, a large number features increase the risk of noise and ambiguous data points which decreases the model accuracy. To overcome this issue, it is advisable to apply feature selection techniques to identify relevant features from datasets. Feature selection techniques are broadly classified in the following two categories:

Unsupervised Feature Selection techniques: Applicable to datasets with unlabelled data.

Supervised Feature Selection: Applicable to datasets with labelled data. The Home Credit dataset comprises of labelled data and uses the following Supervised Feature Selection Techniques

<u>Wrapper Method:</u>

Wrappers require some method to search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset. The feature selection process is based on a specific machine learning algorithm that is fit on a given dataset. It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. The wrapper methods usually result in better predictive accuracy than filter methods but is costlier computationally.

The following two methods are used in the study.

i. Forward Selection: This approach starts with the best performing feature against the TARGET variable. Each successive iteration, a new feature is added and the model performance gains are compared. This process of feature addition continues till the pre-set criteria of number of features is met.

ii. Backward Elimination: This approach is similar to forward selection but starts with all the features and compares model performance. Every successive iteration removes a feature and compares model performance. The process continues till the pre-set criteria is met.

Embedded Method:

These methods encompass the benefits of both the wrapper and filter methods, by including interactions of features but also maintaining reasonable computational cost. Embedded methods are iterative in the sense that takes care of each iteration of the model training process and carefully extracts those features which contribute the most to the training for a particular iteration.

This study uses Recursive Feature Elimination embedded method built using Random Forest for feature selection

3.9 Models Development.

Decision Tree.

Decision Tree is a supervised classification algorithm. It has a tree structure having nodes which represent the features and the decision rule is represented by the branches. A decision tree consists of a set of sequential binary splits of the data. Decision Trees have been widely used in lot of studies related to credit scoring prediction. (Wang et al., 2011)(Xia et al., 2017)(Dastile et al., 2020)

Table 1.1: Decision Tree model comparison

Advantages	Disadvantages
A decision tree does not require normalization of data.	A small change in the data can cause a large change in the structure of the decision tree causing instability.
A decision tree does not require scaling of data as well.	For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.	Decision tree often involves higher time to train the model.
Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.	Decision tree training is relatively expensive as the complexity and time has taken are more.

Random Forest

Random Forest is a supervised classification algorithm which is based on ensemble learning. The advantage of random forests depends on strength of relationships between variables. In data sets with little interaction effects random forests may not outperform. On large credit data sets, behavioural models, application scoring random forests can improve existing credit models (Brown and Mues, 2012)(Ala'raj and Abbod, 2016)(Soares De Melo Junior et al., 2019)(Marqués et al., 2012a)

Table 2.2: Random Forest model comparison

Advantages	Disadvantages
Model can be used for classification as well as regresssion purposes and handles categorical as well as continuous features equally accurately	Random Forest creates a lot of trees (unlike only one tree in case of decision tree) and combines their outputs. To do so, this algorithm requires much more computational power and resources. Thus, complexity for the model increases
No feature scaling required: No feature scaling (standardization and normalization) required in case of Random Forest as it uses rule based approach instead of distance calculation.	Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes. Thus, the model has a longer training period
Handles non-linear parameters efficiently: Non linear parameters don't affect the performance of a Random Forest unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Random Forest may outperform as compared to other curve based algorithms	
Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.	

Light Gradient Boost

Gradient boosting uses the loss function which is the residual errors from earlier models in the iteration to create a new predictor. Light Gradient Boost divides the tree leaf-wise as compared to other boosting methods that build the tree model using level-wise and it does that by choosing the leaf which has the largest delta loss. This leaf-wise method has smaller loss when compared

to the level-wise method. But it has a downside that it increases complexity and lead to overfitting but this can be controlled by specifying the max depth parameter. Even Gradient boosting has been explored as part of ensemble techniques and comparative analysis by many studies. LGBM has been used successfully in a number of credit scoring studies (Taha and Malebary, 2020)(Tounsi et al., 2020)(Qiu et al., 2019). Some of the advantages of using LGBM are:

- Faster training speed and higher efficiency: Light GBM use histogram-based algorithm i.e., it buckets continuous feature values into discrete bins which fasten the training procedure.
- Better accuracy than any other boosting algorithm: It produces much more complex trees by following leaf wise split approach rather than a level-wise approach which is the main factor in achieving higher accuracy. However, it can sometimes lead to overfitting which can be avoided by setting the max_depth parameter.
- Replaces continuous values to discrete bins which result in lower memory usage.

Extra Tree Classifier

Extra Tree Classifier is also a supervised classification algorithm based on ensemble learning. It also uses several decision trees to decide the best-ft model. Extra Tree classifier uses a random way to select attribute and the cut-off point for node as compared to the Random Forest. This ensures more diverse trees and less splitting during Training. Also, it is little faster in terms of execution because of the randomness involved. Previous studies have not explicitly used Extra trees so this dissertation explores its usage.

Logistic Regression

Logistic regression can be used to predict default events and model the influence of different variables on a consumer's creditworthiness and has been used as the base model in a number of papers or studies for credit scoring (Hsu and Hung, 2009) (Soares De Melo Junior et al., 2019), (Chen and Li, 2010). Logistic regression is easy to implement, interpret and very efficient to train and makes no assumptions about the distribution of classes in feature space. It does construct linear boundaries and may result in overfitting if the number of observations is lesser than the number of features.

Table 3.3: Logistic Regression model comparison

Advantages	Disadvantages
Logistic regression is easier to implement, interpret, and very efficient to train.	If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
It makes no assumptions about distributions of classes in feature space.	It constructs linear boundaries.
It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.	The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

K Nearest Neighbours

One of the most commonly used methods for credit scoring is k nearest neighbour (KNN) and has been used in a number of credit scoring studies (He et al., 2018)(Bellotti and Crook, 2009)(Crone and Finlay, 2012). This method belongs to the category of nonparametric classification method. It is known that the non- parametric classifier usually suffer from the existing outliers, especially in the situation of small training sample size.

Table 4.4: KNN model comparison

Advantages	Disadvantages
Comparatively faster than other algorithms. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression.	Sensitive to noisy data, missing values and outliers: KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers
KNN is very easy to implement. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)	Does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension
Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.	Does not work well with large dataset: In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.

SVM-Linear Kernel

Support Vector Machines and metaheuristic approaches have constantly received attention from researchers in establishing new credit models and has been used extensively in credit scoring papers (Hsu and Hung, 2009)(Chen and Li, 2010)(Wang et al., 2011)(Bellotti and Crook, 2009)

Table 5.5: SVM model comparison

Advantages	Disadvantages
SVM works relatively well when there is a clear margin of separation between classes.	SVM algorithm is not suitable for large data sets.
SVM is more effective in high dimensional spaces.	SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
SVM is effective in cases where the number of dimensions is greater than the number of samples.	In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
SVM is relatively memory efficient	As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

Naïve Bayes

Naïve Bayes is a simple supervised classification algorithm which is based on Bayes theorem. Naïve Bayesian algorithm has been able to predict credit request as good or bad respectively and has been used in a number of credit scoring studies(Marqués et al., 2012a)(Ala'raj and Abbod, 2016)(Abellán and Castellano, 2017)

Table 6.6: Naïve Bayes model comparison

Advantages	Disadvantages
Simple and easy to implement and does not need large training data. Algorithm is fast and can be used to make real time predictions	Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases.
Ability to work with continouous and discrete data points. Highly scalale with the number of predictors and datapoints. Model is not sensitive to irrelevant features	This algorithm faces the 'zero-frequency problem' where it assigns zero probability to a categorical variable whose category in the test data set wasn't available in the training dataset. It would be best if you used a smoothing technique to overcome this issue.

Adaboost

Adaboost is a type of boosting method which is used as an ensemble method. It combines different weak classifiers to finally built a string classifier. It assigns more weight to instances which are difficult to classify as compared to the ones which are easy to classify. It is less prone to overfitting as it decreases the bias and variance. Adaboost is less prone to overfitting as the input parameters are not jointly optimized. The accuracy of weak classifiers can be improved by using Adaboost. Nowadays, Adaboost is being used to classify text and images rather than binary classification problems. The main disadvantage of Adaboost is that it needs a quality dataset. Noisy data and outliers have to be avoided before adopting an Adaboost algorithm.

3.10 Models Tuning.

While building machine learning models there are a lot of options for defining and optimizing the model. This is called tuning the model which involves tweaking various options in such a way that you get the best performing model. Machine learning models are instructed to carry out this exploration and then help select the best-fit model architecture. The parameters which determine the above are called hyperparameters and tweaking these parameters is called hyperparameter tuning. One of the methods of doing hyperparameter tuning is using Grid Search

Grid search is the simplest method for hyper-parameter tuning. In Grid Search the model is created by tuning hyperparameter with all for all possible combinations available and then each model is evaluated to select the best performing model. But this method involves a lot of sampling across the hyper-parameter and this could lead to wastage of efforts.

3.11 Models Evaluation.

Model Evaluation is the process through which the quality of system's predictions is quantified. The newly trained model performance is evaluated on a new and independent dataset. This model compares labelled data with its own predictions.

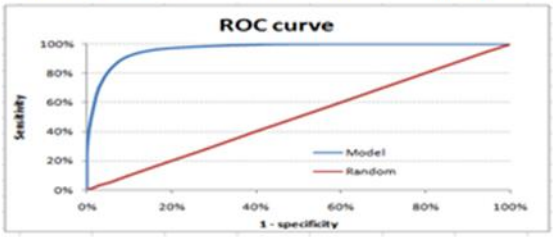
Cross Validation: This is a general framework to assess how a model performs in the future; it is also used for model selection. It consists of splitting your training set into test and control data sets, training your algorithm on the control data set, and testing it on the test data set.

There are four different outcomes that can occur when your model performs classification predictions:

- True positives (TP) occur when your system predicts that an observation belongs to a class and it actually does belong to that class.
- True negatives (TN) occur when your system predicts that an observation does not belong to a class and it does not belong to that class.
- False positives (FP) occur when you predict an observation belongs to a class when in reality it does not. Also known as a type 2 error.
- False negatives (FN) occur when you predict an observation does not belong to a class when in fact it does. Also known as a type 1 error.

Using the outcomes listed above, this study evaluates model performance using the following metrics.

Table 7.7: Model Evaluation Metrics

<table>
<tr><th>Evaluation Metric</th><th>Description</th></tr>
<tr><td>Accuracy</td><td>measures the proportion of true results to total cases. Aim for a high accuracy rate.
$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$</td></tr>
<tr><td>Precision</td><td>It is the proportion of true results over all positive results.
Precision - True Positives / (True Positives + False Positives)</td></tr>
<tr><td>Recall</td><td>It is the fraction of all correct results returned by the model
Recall - True Positives / (True Positives + False Negatives)</td></tr>
<tr><td>F1 Score</td><td>F1 score is determined using both Precision & Recall and is measured as below.
$$F_\beta = (1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+\text{recall}}.$$</td></tr>
<tr><td>ROC/AUC</td><td>ROC stands for Receiver operating characteristics curve. AUC measures the area under the curve plotted with true positives on the y axis and false positives on the x axis.
 
</td></tr>
<tr><td>Confusion Matrix</td><td>Confusion Matrix the correlation between the label and the model's classification. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label
<table>
<tr><td colspan="2" rowspan="2">Confusion Matrix</td><td colspan="2">Target</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Positive</td><td>Negative</td></tr>
<tr><td rowspan="2">Model</td><td>Positive</td><td>a</td><td>b</td><td>Positive Predictive Value</td><td>a/(a+b)</td></tr>
<tr><td>Negative</td><td>c</td><td>d</td><td>Negative Predictive Value</td><td>d/(c+d)</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>Sensitivity</td><td>Specificity</td><td colspan="2" rowspan="2">Accuracy = (a+d)/(a+b+c+d)</td></tr>
<tr><td>a/(a+c)</td><td>d/(b+d)</td></tr>
</table></td></tr>
<tr><td>Kappa</td><td>The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves</td></tr>
<tr><td>MCC</td><td>The Matthews correlation coefficient (MCC) or phi coefficient is used in machine learning as a measure of the quality of binary (two-class) classifications
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$</td></tr>
</table>

3.12 Interpretable Models

As attested in the literature review of this study, in the credit scoring domain, the research trend has been focussed to construct newer and more sophisticated hybrid models that can predict credit defaults with higher accuracy. As models become more complex, they tend to negatively impact the model interpretability. Given government and regulatory demand that highlights model interpretability, this area has a lot of focus from banks and financial institutions. Explainable AI methods are a process where model prediction can be understood. This paper explains the following two methodologies

LIME (Local Interpretable Model-Agnostic Explanations): is an algorithm that can explain the predictions of a model, by approximating it too locally with an interpretable model. A single data sample is modified by changing the values of a feature and the impact on the output prediction is calculated. It explains the predictions from every data sample.

SHAP (SHapley Additive exPlanations): it is a theoretic approach for explaining the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. Shapley values are approximating using Kernel SHAP, which uses a weighting kernel for the approximation, and Deep SHAP, which uses Deep Lift to approximate them.

Table 8.7: Explainable model comparison

SHAP	LIME
Explanations for local as well as global features are provided	Explanations are for local features and to specific instances
Computation speed is slow	Computation speed is fast
Result explains the variable contribution to the difference between actual and means prediction	Results are determined based on variable selection and its impact on prediction. Difference between keeping the variable and removing the same are calculated

3.13 Expected Outcome

The outcome of the study is expected to help predicting credit risk based on the current loan, previous loan, loan repayment history and bureau features of Home Credit datasets.

- The study compares the impact of applying data balancing techniques on credit scoring datasets. It presents a performance comparison of over sampling, under-sampling and hybrid sampling techniques on Home Credit dataset
- The study demonstrates implementation of manual feature engineering techniques and identify the optimal features for model creation by comparing performances of wrapper and embedded feature selection techniques
- The dissertation compares the performance of 13 classifiers and three heterogenous ensembles and presents a comparative analysis of their performance.
- The best performing base classifier is used to create interpretability and explain-ability for the identified model. This aids adoption of the model by Home Credit business groups.

3.14 Summary.

The Research methodology has the following sections. Section 3.2 outlines the framework pipeline which is evaluated. Section 3.3 and 3.4 describe the dataset, the data pre-processing and class imbalance handling that is done as the first step. Section 3.5 and 3.6 present the details of the feature section and feature scaling methods which are used prior to the model development. All the various classification models which are in the scope of this study are detailed in Section 3.7 along with the reasons for picking these models. Section 3.8 elaborates the model tuning and evaluation steps for the shortlisted models. Section 3.12 covers the model interpretation techniques while section 3.13 lays down the expected outcome from this study.

CHAPTER 4: EXPERIMENTS & ANALYSIS

4.1 Introduction

In this chapter, the data set features, data visualization and treatment performed as part of data processing to make the model optimal for machine learning models is explained. Also, the Exploratory Data Analysis performed on the dataset, the observations and inferences drawn on the data visualization are explained. It covers the feature engineering approach covered in the study and the features generated using multiple datasets available for the use case. It then covers the nuances of data sampling and feature extraction and creates multiple datasets against which model performance can be compared. The chapter describes which machine learning algorithms are used, the framework used in building the models, the hyper parameter tuning approach and model validation. Using the best performing models, the process of building ensembles is explained in detail. The top model features are explained using LIME and Tree SHAP and the features are compared with visualizations provided by each of the libraries.

4.2 Dataset Description

Home Credit uses a variety of data like loan application information, client demographics, type of loan availed, credit bureau information, transactional to predict their client's repayment abilities.

The dataset provided by Home Credit comprises of seven sources of data.

- Application train. Each current credit application made by the customer is represented by a single row in this dataset. Each row is uniquely identified by SK_ID_CURR feature. The dependent variable is represented by the TARGET feature. A preliminary analysis of the data shows that the dependent variable distribution is imbalanced in the dataset. Only 8% of the total volume has credit default or dependent/TARGET variable value set as 1. There are a total of 307511 rows and 122 features. There are 16 categorical features that explain various aspects of the customer, his/her home preferences/loan preferences and aspects of the customer. The other features are numerical in nature
- Bureau. This dataset contains information about the loan history of the customer from different banks and institutions as kept in bureau. Each row represents a loan row of the customer, uniquely identified by SK_ID_BUREAU and containing the customer key (SK_ID_CURR). The dataset contains features about credit bureau loan debt size,

date started, date ended, annuity paid, type of loan availed and the currency of the same. This dataset features offer an opportunity to generate aggregate features which can then be linked to customer using the SK_ID_CURR key and be used for analyse the impact of the features on dependent variable.

- Bureau Balance. This dataset contains monthly balances of loans in Credit Bureau. There is an opportunity to create aggregate features on monthly balance by SK_ID_BUREAU feature and net that up to the bureau table which can be used to compare the impact of the aggregated feature on dependent variable.
- Previous loan application: This dataset contains details about previous loan applications made by customers at Home Credit as each row in the dataset identified by feature SK_ID_PREV. The previous application's annuity amount, credit amount, goods price and decision are good features for feature engineering which has a bearing on the current application's dependent variable.
- Point Of Sale Cash Balance: This dataset contains features related to cash or point of sale credits. The monthly balance and instalment count features when aggregated by SK_ID_CURR can give insights on whether the current application customer is likely to default or not.
- Credit Card Balance: This dataset contains features related to the revolving loans availed by the customer. The annuity amount, credit amount, balances, and repayment history features from this dataset when aggregated on SK_ID_CURR gives insights on the customer's ability to repay loans.
- Instalment history: This dataset contains features related to instalment history of the customer. If the first few instalments are paid on time by the customer, that is a good domain indicator of the customer being credit worthy. The days past due and the payment percentage insights obtained from this dataset can be analysed to see how they impact the dependent variable on the current application.

4.3 Data Preparation & Cleaning

In this section, the datasets for Home Credit are analysed for any anomalies in data, missing data and outliers within the dataset. These errors in the dataset may adversely impact the performance of predictive model. Data Cleansing techniques are implemented in a systemic manner on the datasets. At every level, the data integrity is validated to ensure no information is lost. The steps followed are listed below:

4.3.1 Feature Elimination:

The application_train dataset has the following missing % of missing values on the dataset.

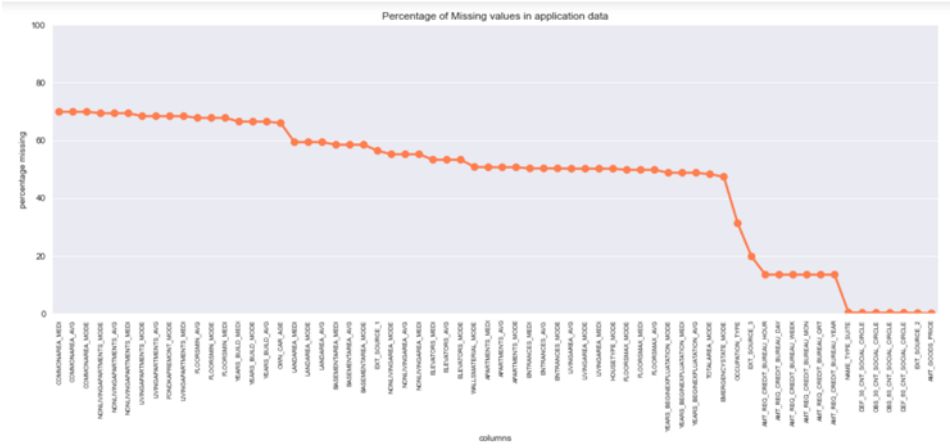


Figure 3.1: Application train dataset features with missing values

- The External Source columns demonstrate a high correlation to the target variable and thus their missing values are imputed to the median value of the column.
- The amount related fields which have less than 20% missing values are imputed such that the missing values are replaced by the median value for the column.
- Columns having more than 40% of missing values are dropped from the dataset above.
- The features with names starting with 'FLAG' have mostly zero values thereby having very skewed data and not providing enough insights for modelling. Thus, all except 'FLAG3' feature are dropped.
- The following columns are dropped from the dataset as their values are very skewed to zero. AMT_REQ_CREDIT_BUREAU_HOUR 'AMT_REQ_CREDIT_BUREAU_WEEK, , AMT_REQ_CREDIT_BUREAU_QRT', AMT_REQ_CREDIT_BUREAU_MON

Thus, the application train dataset features which were 122 at the start are now reduced to 53 features.

The below charts showcase the percentage of missing values in other datasets and the treatment performed on the same.

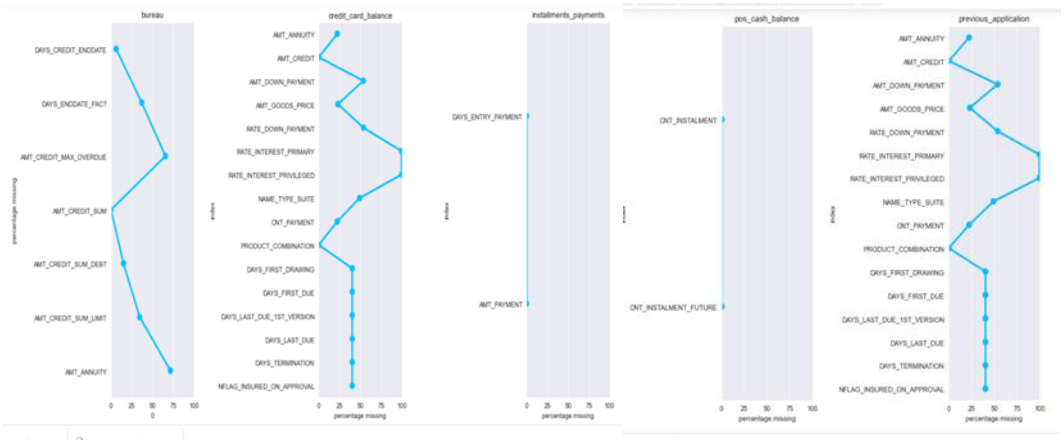


Figure 4.2: Home Credit dataset features with missing values

In the bureau dataset, the following amount columns have their missing values imputed to the median value of the column. (AMT_ANNUITY, AMT_CREDIT_MAX_OVERDUE, AMT_CREDIT_SUM, AMT_CREDIT_SUM_DEBT, AMT_CREDIT_SUM_LIMIT, AMT_CREDIT_SUM_OVERDUE)

In the previous application dataset, the following columns have 20% of the population as missing values and the same is imputed with the median value of the column (AMT_GOODS_PRICE, AMT_ANNUITY, CNT_PAYMENT).

The point-of-sale cash balance dataset has the following columns with 0.3% of data missing and the same is imputed with the median value of the column. (CNT_INSTALMENT and CNT_INSTALMENT_FUTURE)

4.3.2 Outlier Data Treatment using Univariate & Distribution graph analysis

For each of the datasets, the numerical column distributions are analysed and if the frequency of outliers is low, the boundaries calculated are set for those outlier caps. The boundaries are calculated as per below:

- For normally distributed numerical features, the upper and lower boundaries are calculated using gaussian rule.
- For non-normally distributed numerical features, the upper and lower boundaries are calculated according to interquartile proximity rules. IQR is first calculated using 75th quantile - 25th quantile and then the upper boundaries are calculated using 75th quantile + (IQR * 3)

Using the above approach, features from each of the datasets are analysed using univariate box plots as well as distribution/ KDE plots and outlier treatments are applied accordingly.

Application train dataset: Out of 52 features, 14 features are categorical. The numerical variables are analysed on whether they are normally distributed or have a non-normal distribution. A univariate analysis of the individual features shows the distribution as below.

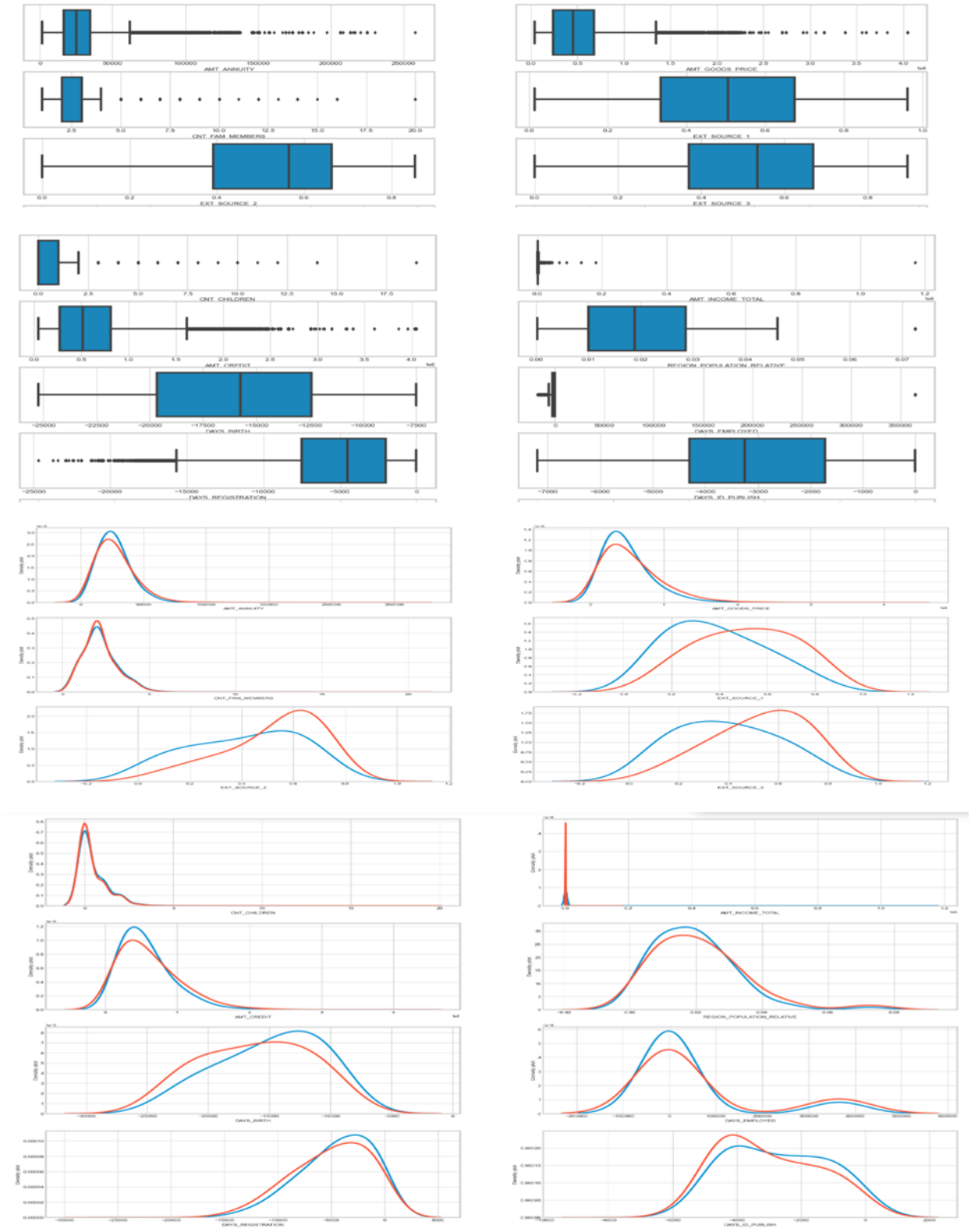
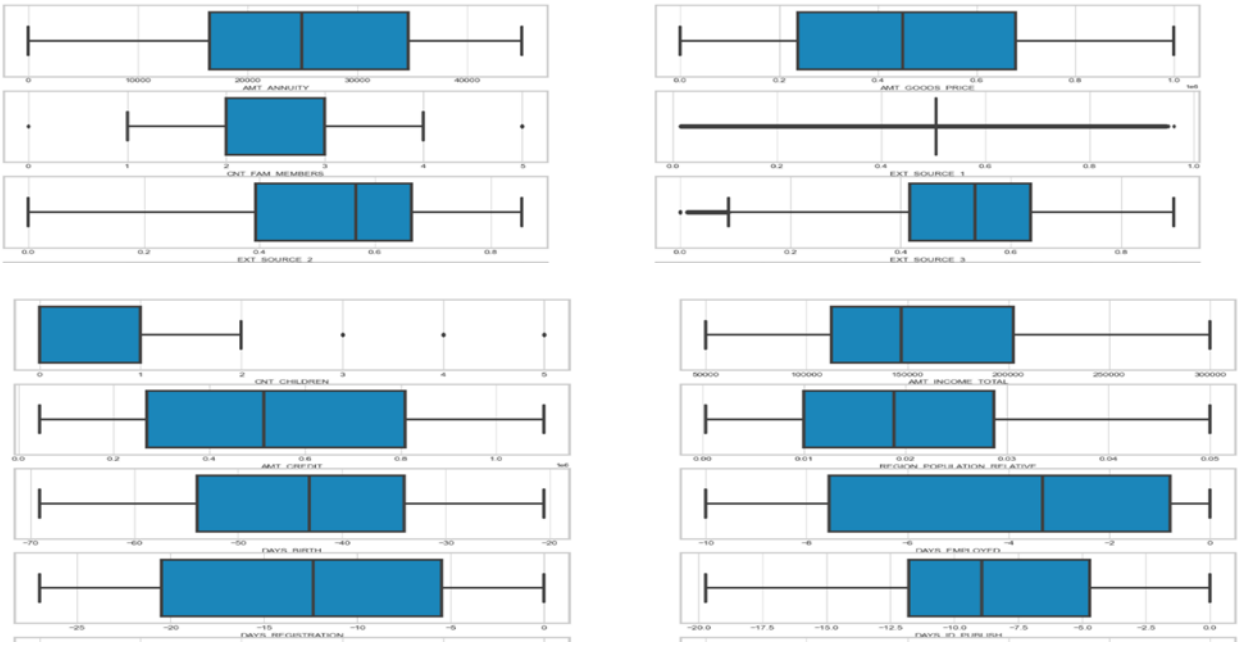


Figure 5.3: Application train dataset features before outlier treatment

- The amt_income_total feature is capped on the upper end at 0.3 million and at the lower end at 50 thousand. Values lower or above these thresholds are low in frequency
- The annuity amount feature values are low in frequency above 45K and thus are capped at that value.
- The goods price amount feature is capped at 1 million
- The credit amount is capped at 1.1 million as any values beyond that are low in frequency and capped to the maximum value.
- The family member count and children count features each are capped to value 5. Any values above that have strong credit default values.
- Region population relative feature is capped to 0.05.
- The four features related to social circle values are capped at 10 and 3 respectively to box in the values and prepare the feature to be used by models.
- The days employed has outliers at both ends and is capped at min and max values of -10 and 0
- The registration date is capped at a min value of -27 days



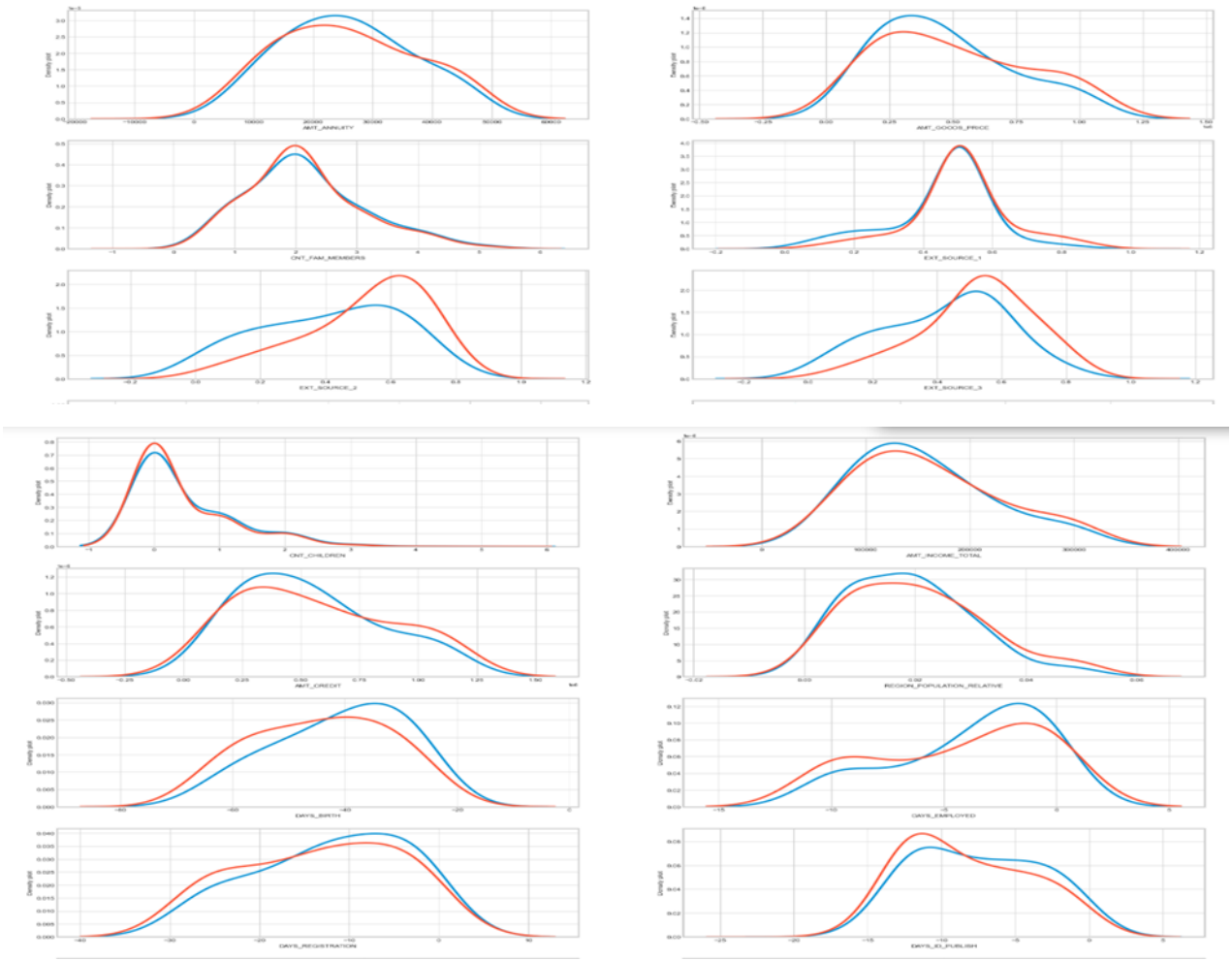
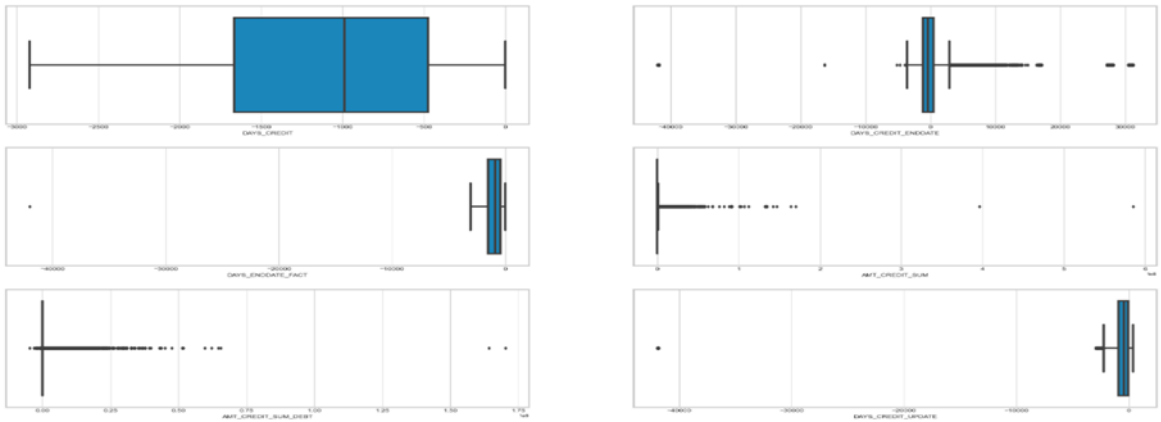


Figure 6.4: Application train dataset features after outlier treatment

Bureau dataset: Out of 11 features, 6 features are numerical. The numerical variables are analysed using univariate analysis and distribution of the individual features:



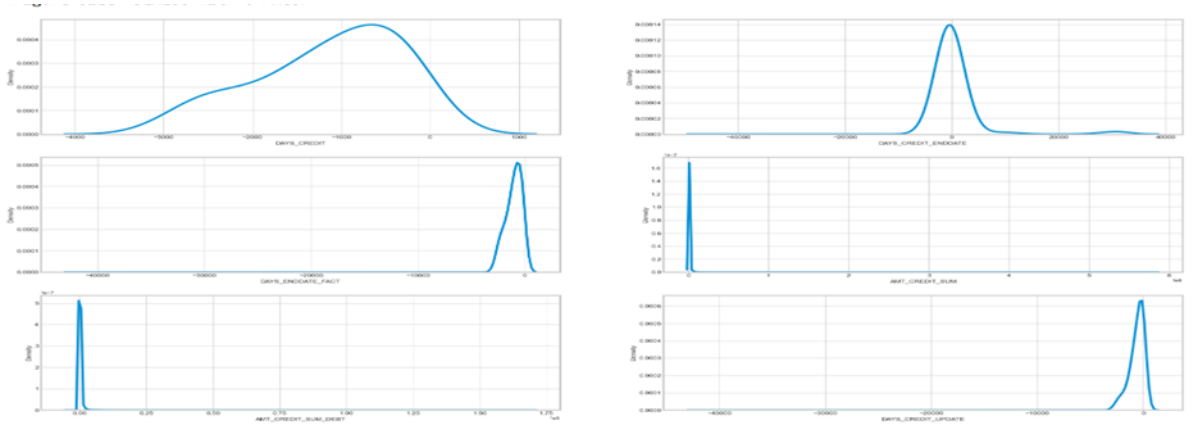


Figure 7.5: Bureau dataset features before outlier treatment

- The credit end-date feature is capped between -2300 and 1600 after distribution analysis and outlier treatment
- The credit amount and credit amount debt features are capped at half million and 50k accordingly
- The credit update feature is capped at a minimum cap of -1500

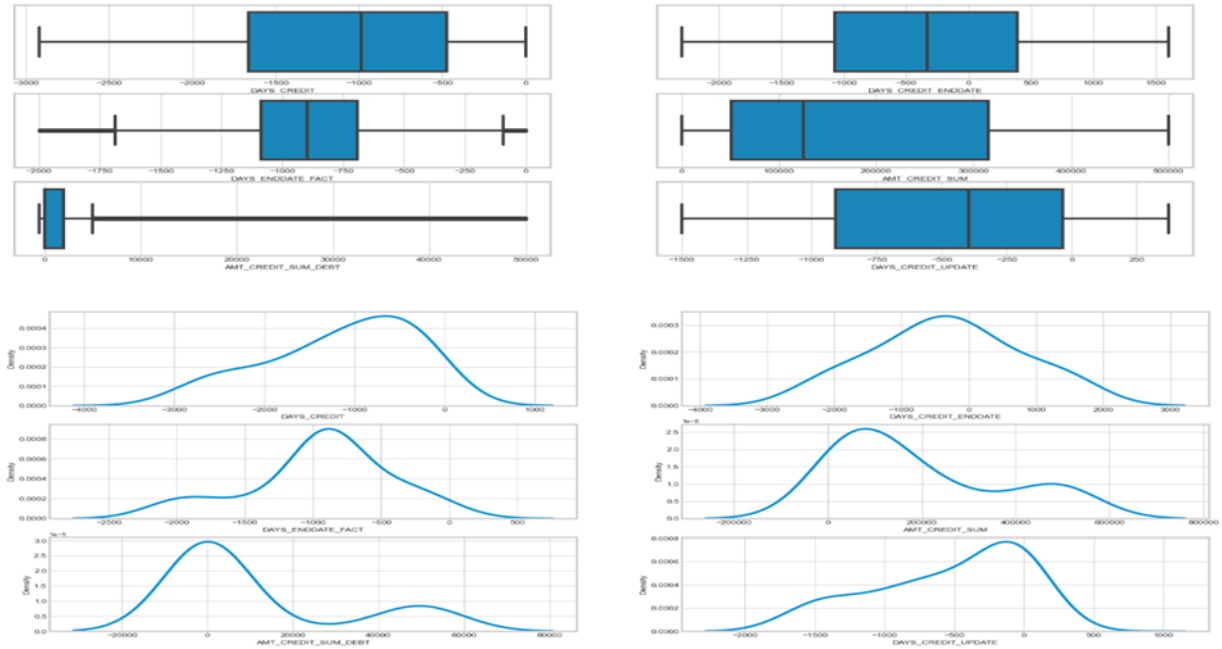


Figure 8.6: Bureau dataset features after outlier treatment

Previous loan application dataset: Out of 37 features, 15 features are numerical. The numerical variables are analysed using univariate analysis and distribution of the individual features:

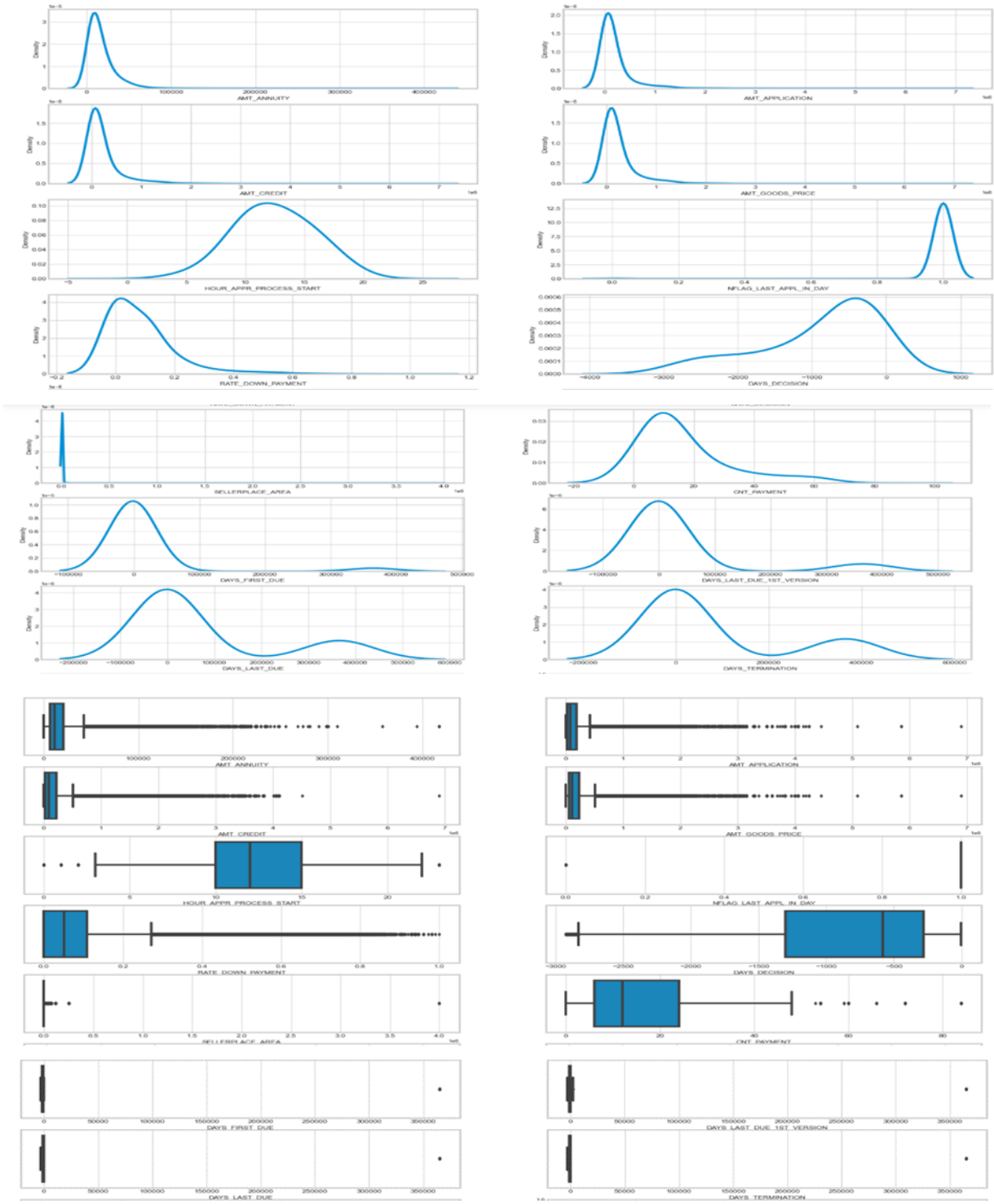


Figure 9.7: Previous loans dataset features before outlier treatment

- The annuity amount is boxed between 5000 and 20K respectively which caps the bottom 5% and upper 8% of values for the feature
- The application amount, goods price and credit amount are capped at the upper end to 0.2m, 0.25m and 0.25m respectively.

- The DAYS related features are capped at upper and lower boundaries in accordance with the approach mentioned at the start of this section. The impact is visible on the graphs below.

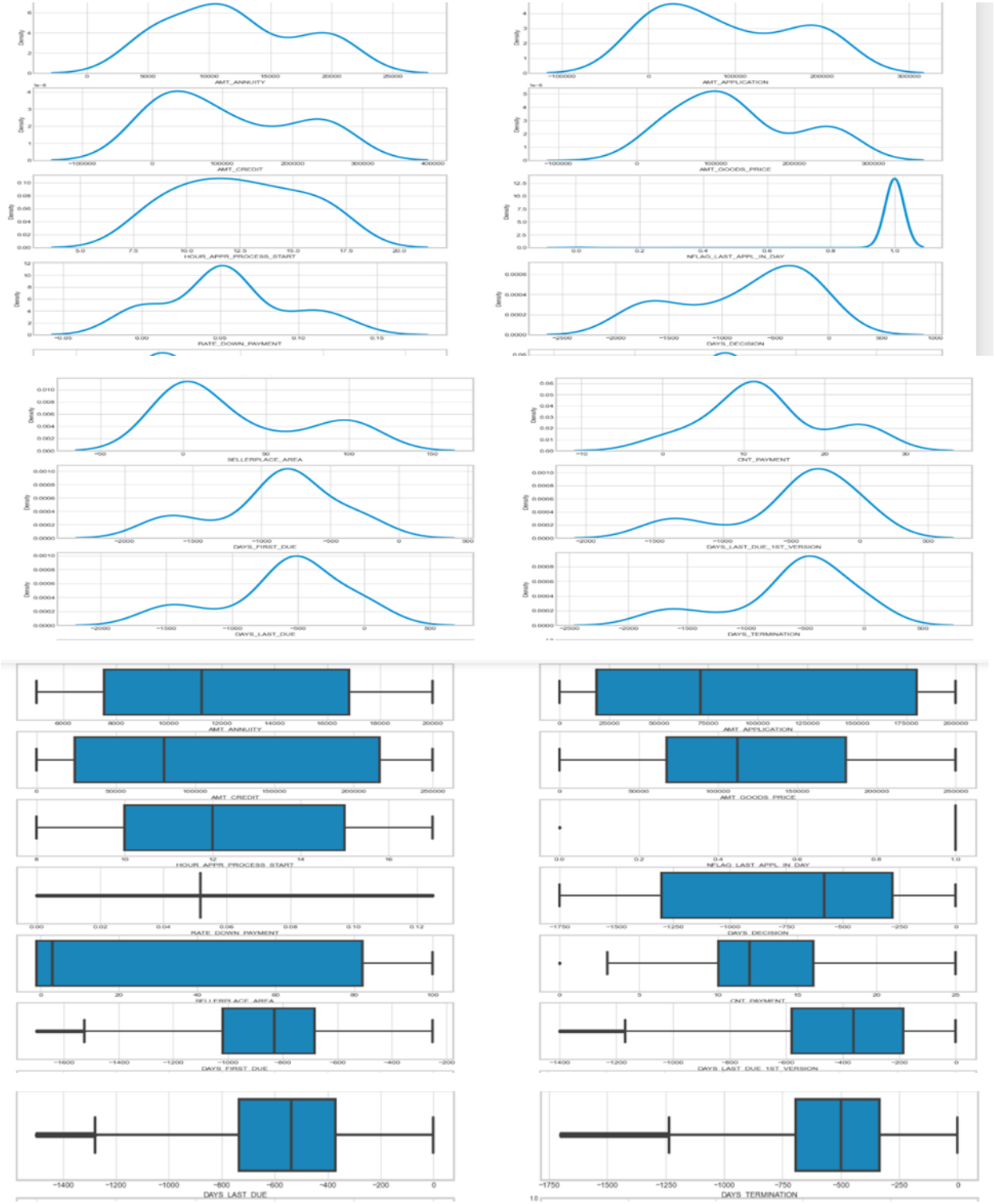
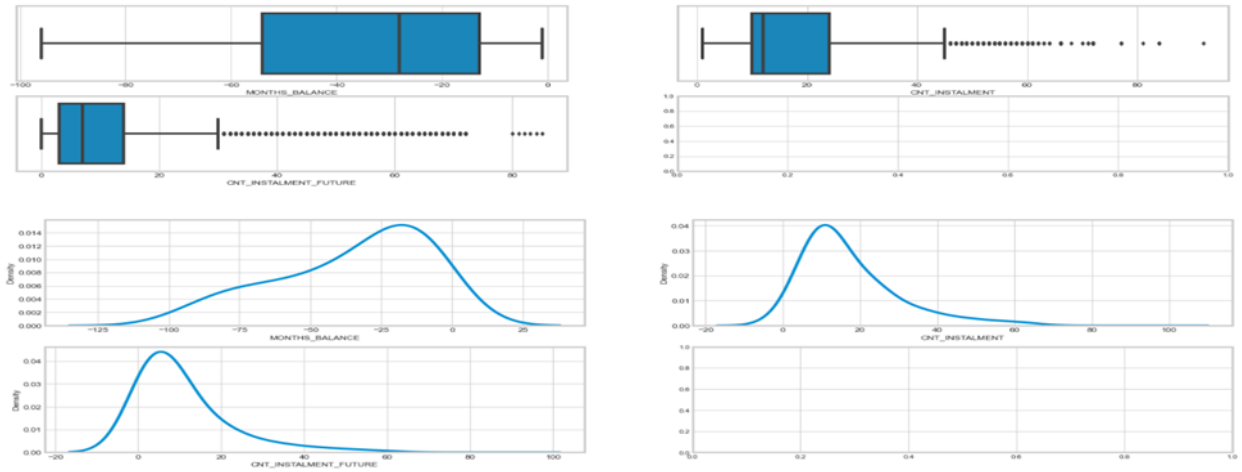


Figure 10.8: Previous loans dataset features after outlier treatment

Point of Sale Dataset: Out of 8 features, three features are numeric and are analysed to assess presence of any outliers and the best way to address the same.

- The instalment count and future instalment count features are capped to an upper boundary of 15 and 25.
- The monthly balance amount is capped to the lower boundary of -70

Before Outlier Treatment



After Outlier Treatment

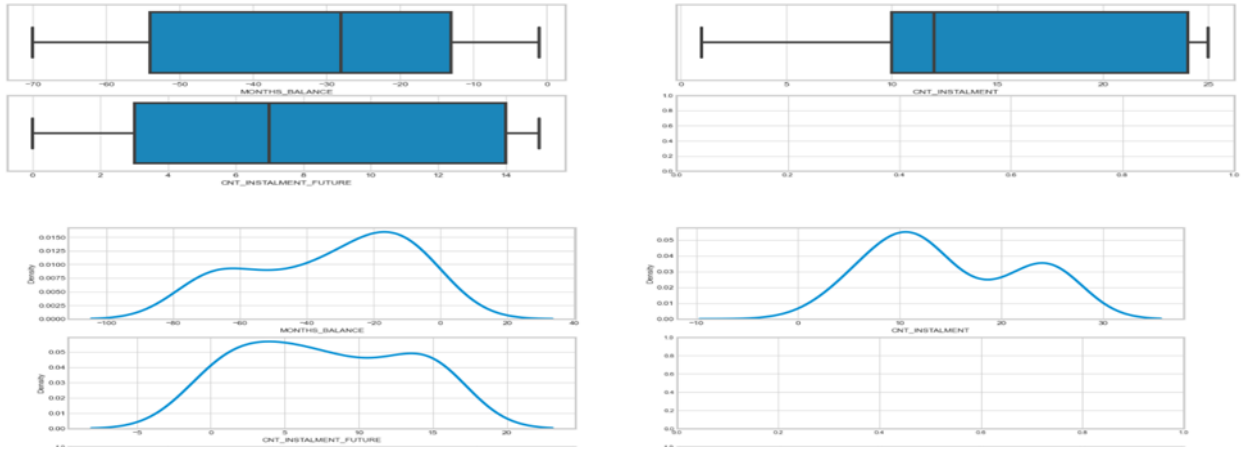
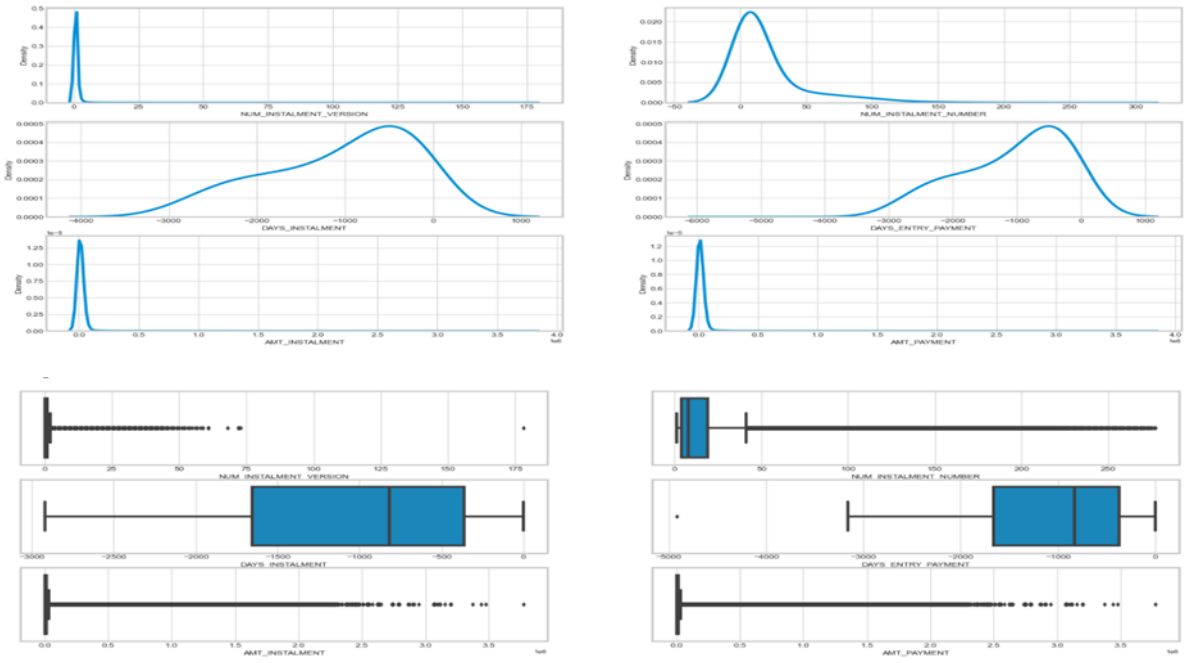


Figure 11.9: Point of sale loans dataset features outlier treatment

Instalment Payment dataset: Out of 8 features, 6 features are numeric and are analysed to assess presence of any outliers and the best way to address the same.

- The payment amount and instalment amount features are boxed in at the upper and lower boundaries in accordance with the outlier treatment approach.
- The payment entry days and instalment days features are capped to a lower value of -2000.
- The instalment number feature is capped to maximum value of 35.

Before Outlier Treatment



After Outlier Treatment

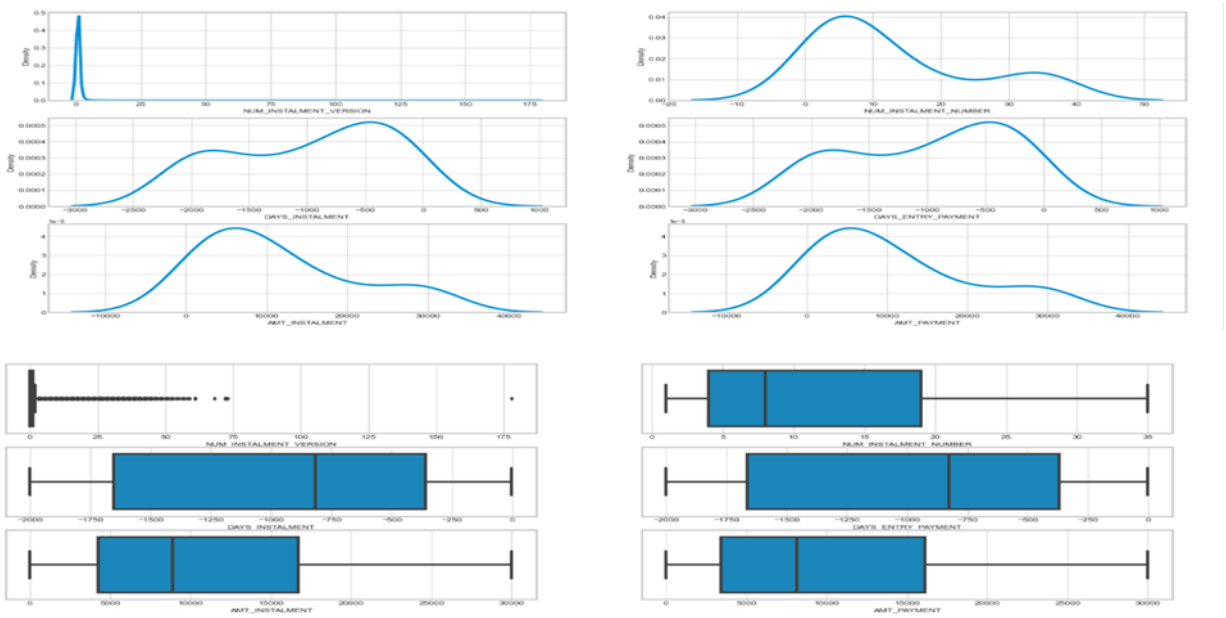


Figure 12.10: Instalment Payments dataset features outlier treatment

4.3.3 Missing Value and Anomaly Treatment

The features under treatment here have lower thresholds of missing values and express a significant corelation with the dependent or target variable of predicting credit default. For numerical columns, the missing value imputation value is calculated post analysis of feature distribution and review of the mean and median value for that feature. For categorical

features, if the missing values are low in frequency and if any other category is the dominant category for that feature, then the missing values are imputed to that category.

For application train,

- The External sources columns missing values are imputed to the median value of the respective feature distributions
- The family count and children count missing values are imputed to a value of zero
- Categorical features with missing values like housetype_mode, Emergencystate_mode, occupation_type and name_suite_type have the missing values imputed to dominant category.
- Categorical feature like name_suite_type have missing values in a relatively higher frequency and are imputed to a separate category "Other" accordingly.
- The DAYS columns missing values are imputed to a value of zero.
- The amount columns missing values are imputed to a value of zero
- The gender column has an anomalous value of 'XNA' which appears in very low frequencies. These values are imputed to the dominant feature of "Female" accordingly

For Bureau dataset,

- The days_credit_update, days_credit_enddate and credit amount feature missing values are imputed to the median value of the respective feature distribution.
- The credit amount debt feature missing values are imputed to zero.

For previous application dataset,

- The annuity amount, credit amount, goods price, rate of down payment, payment count features are imputed to the median value of their respective distributions
- The DAYS related features are also imputed to the median value of their distributions.
- The categorical features of product combination are imputed to the dominant category of 'Cash'
- The Contract type has an anomalous value of 'XNA' with very low frequency and the same is imputed to the dominant category of 'Cash Loans'

For point-of-sale dataset,

- The instalment count and instalment future count features are each imputed to the median value of their respective distributions.

For instalment payment,

- The payment amount, instalment amount and payment entry days feature missing values are imputed to the median value of their respective distributions.

4.4 Exploratory Data Analysis

This section performs a preliminary analysis exploration of the data using graphs and summary statistics. The objective is to identify relationships in the features themselves and also the impact that the features have on the dependent variable. These relationships help identify how the borrowers' quantifiable characteristics influence the creditworthiness of the entity. The relationships and their impact are quantified using visualizations

4.4.1 Data Distribution of dependent variable. Is there a class imbalance?

The dependent feature on the dataset is 'TARGET'. The pie chart illustrates the class distribution for the dependent variable and highlights that only 8% of values show default cases in the total population. This is a strong indicator for class imbalance and may adversely affect model accuracy by causing overfitting. The Home Credit dataset would benefit from data sampling implementation to correct the class imbalance on 'TARGET' variable.

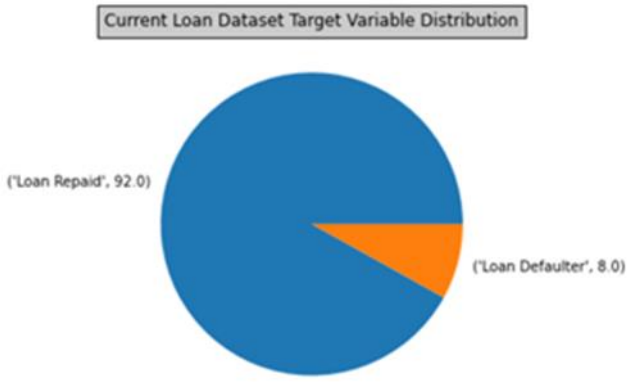


Figure 13.11: Home Credit Dependent/Target Variable distribution

4.4.2 Bivariate Analysis

Bivariate visualization of the dataset features and their relationship / impact on the TARGET variable results in the following observations

- Revolving loans in spite of being only 10% of overall loan contracts have 50% defaults on them

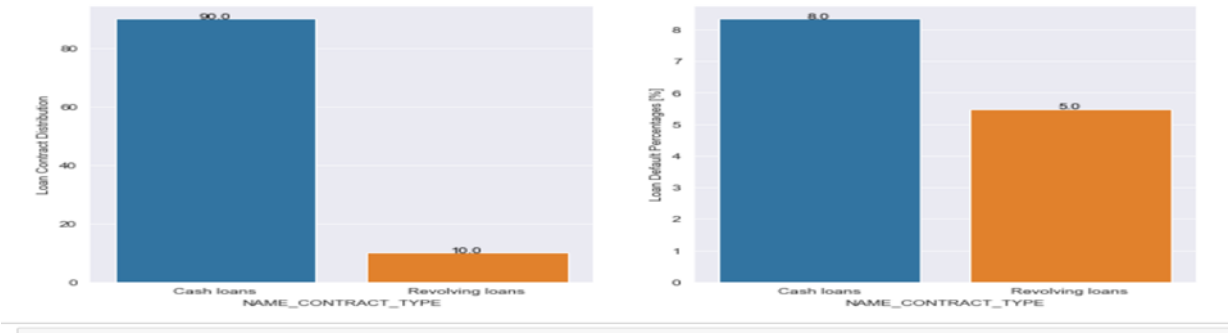


Figure 14.12: Revolving Loans vs Target feature distribution and relationship

- The gender feature-based analysis indicates that while male clients number half to female clients, the number of default cases are higher in that demographic. The number of default cases for males is 10% as compared to females (7%).

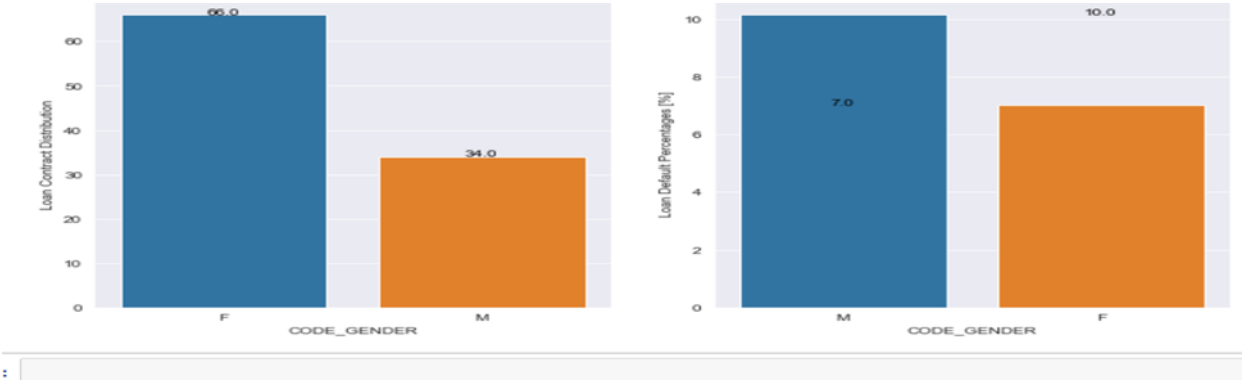


Figure 15.13: Gender vs target variable distribution and relationship

- An analysis of family status of clients reveals that those with status as 'civil marriage' though only 10% of overall population have the highest percentage (10%) of defaults

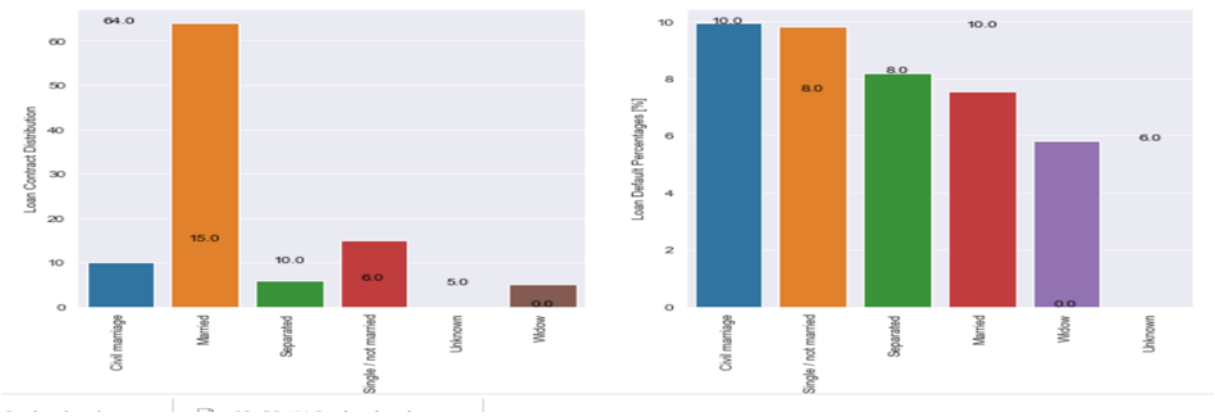


Figure 16.14: Family Status vs Target feature distribution and relationship

- An analysis of family size and children columns reveals that those with large family sizes and children have a 100% loan default history

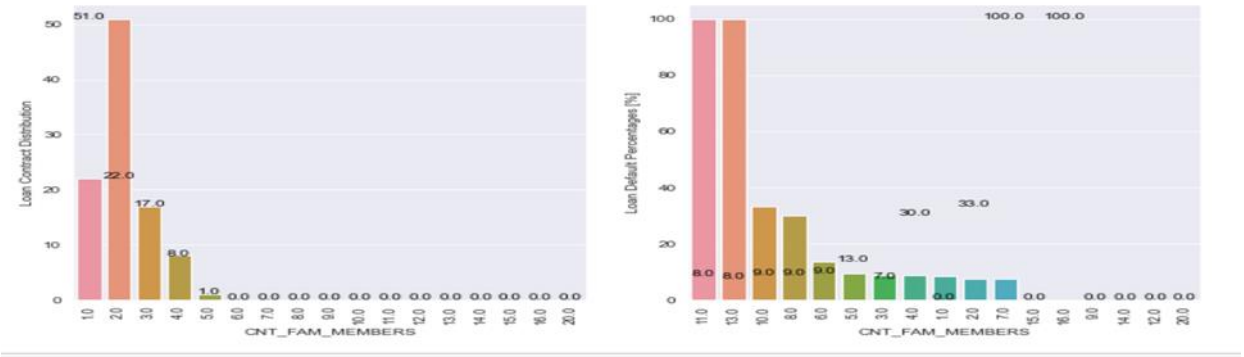


Figure 17.15: Family Count vs Target feature distribution and relationship

- An analysis of the Income Type and its impact on the default probability reveals that applicants on Maternity leave have 40% probability of default followed by Unemployed applicants. This categorical feature is of interest and is subjected to feature engineering to identify more insights.

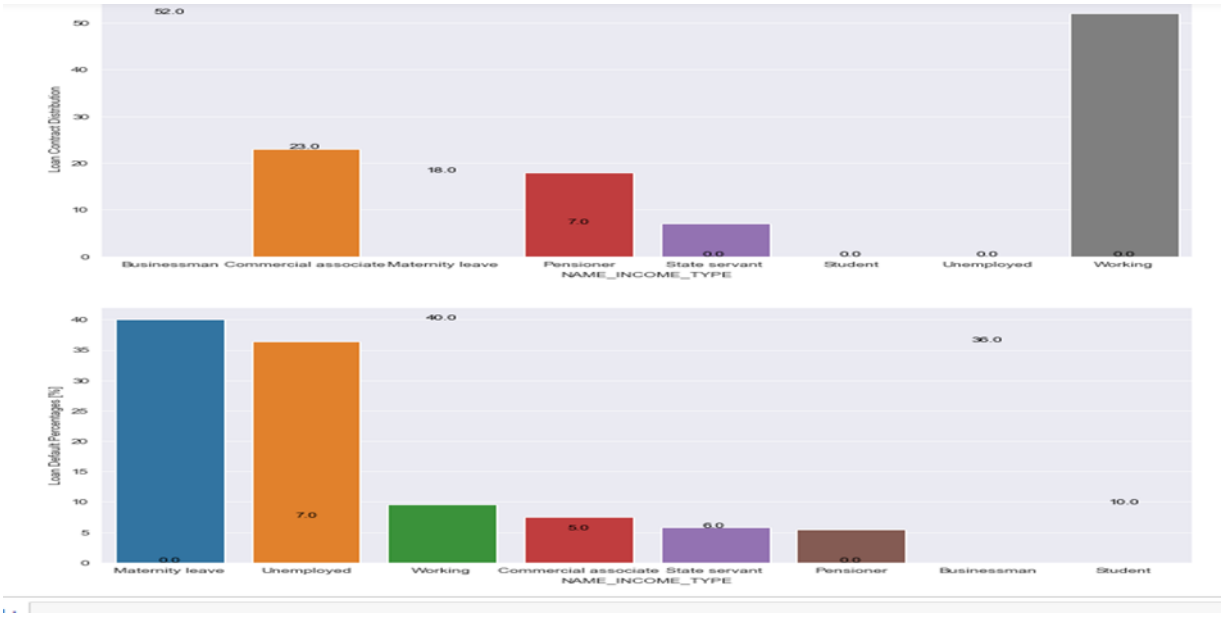


Figure 18.16: Income Type vs Target feature distribution and relationship

- The highest probability of defaults is demonstrated by Low-skill Labourers (17%) followed by the Waiters/barmen and Driver staff, Labourers, Cooking and Security staff. The category applying for loans the most are Labourers.

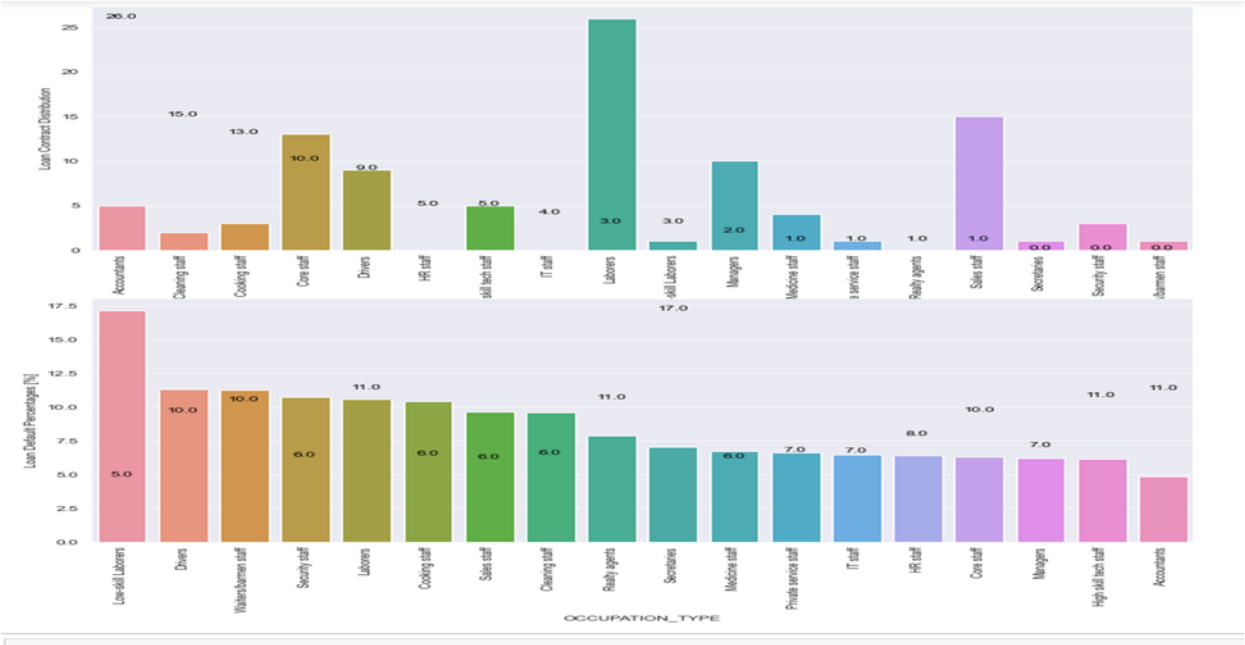


Figure 19.17: Occupation Type vs Target feature distribution and relationship

Transport type3 (16%) category of Organizations tops the list to have highest probability of default followed by Industry 13 at 13.5% and Restaurants at less than 12%.

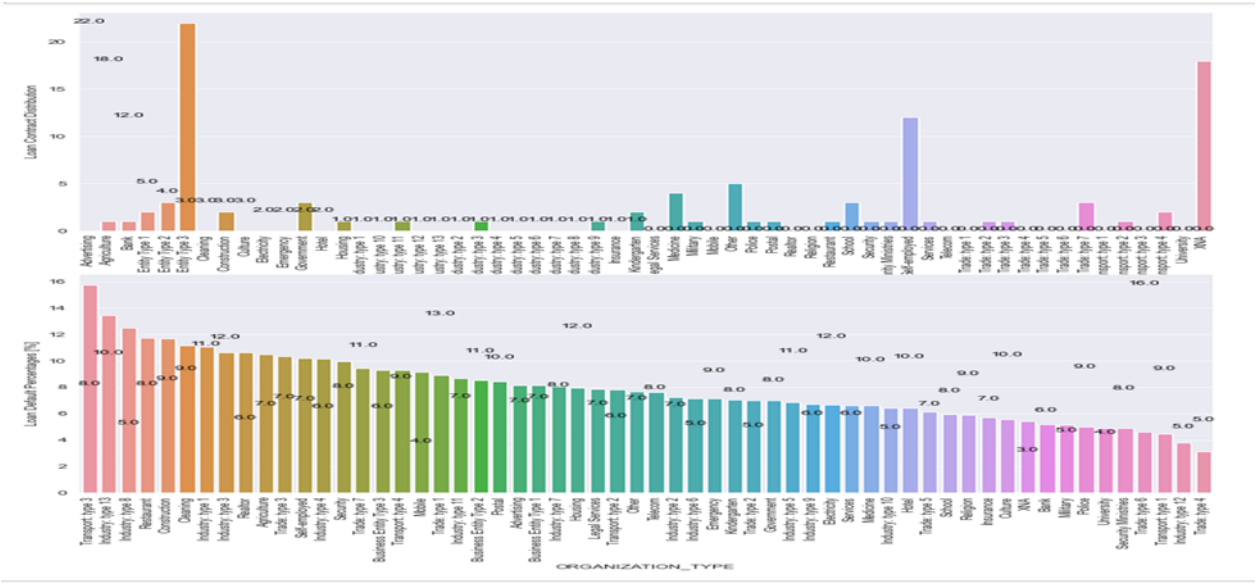


Figure 20.18: Organization Type vs Target feature distribution and relationship

- Applicants holding an academic degree have a 2% probability of default whereas applicants with Lower Secondary category top the list of defaulters with 11% of them having default cases.

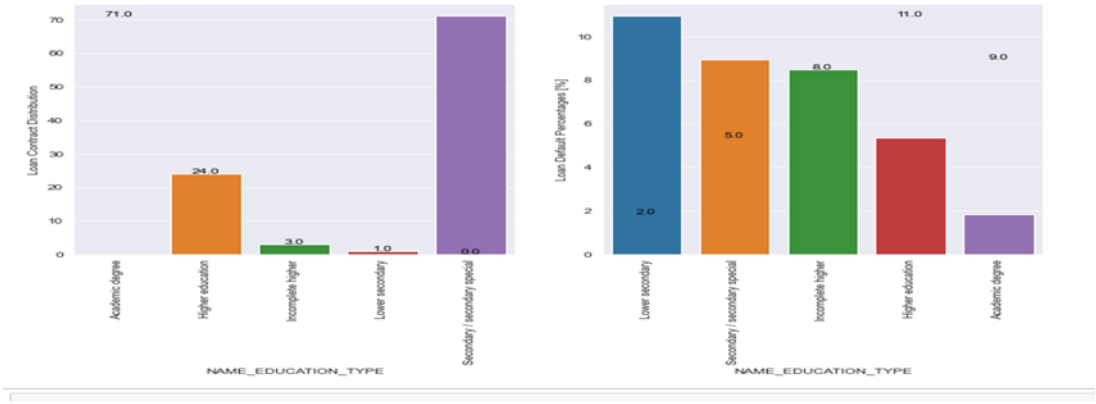


Figure 21.19: Education vs Target feature distribution and relationship

Housing Type of clients indicate that clients living in rented apartments or with parents have over 10% of defaults.

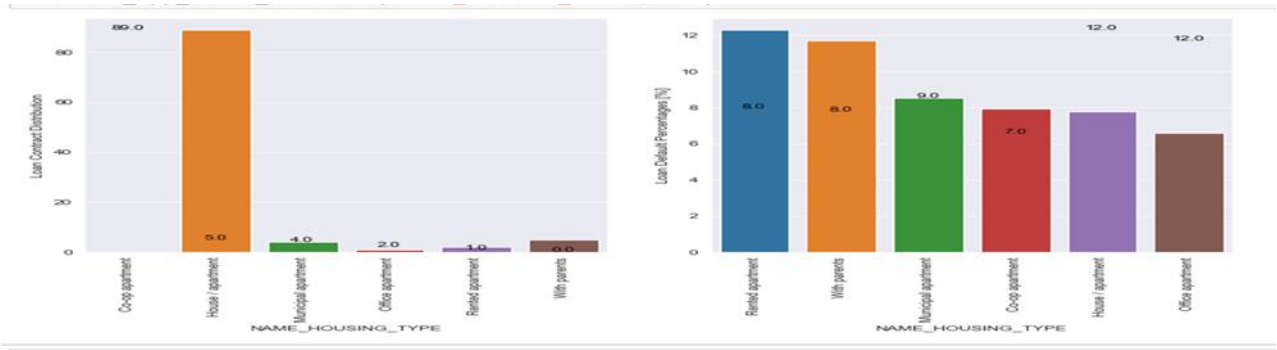
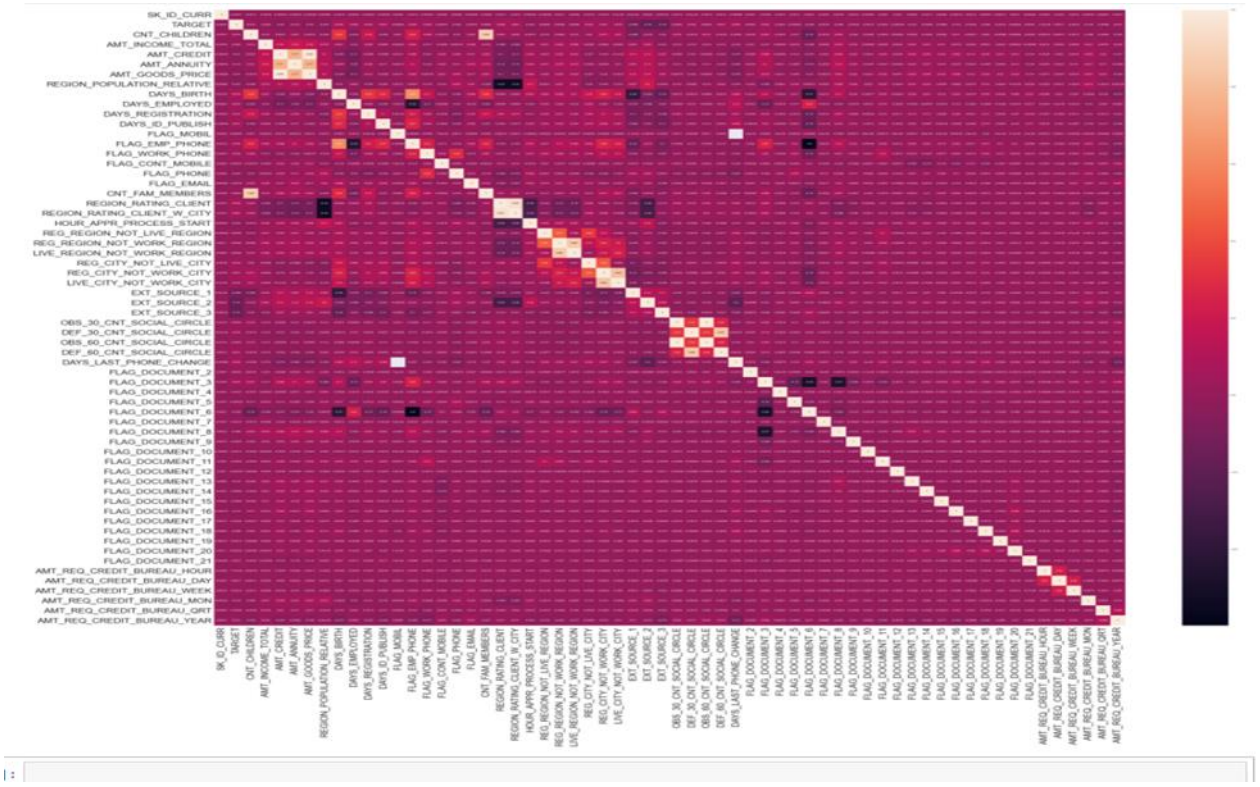


Figure 22.20: Housing Type vs Target feature distribution and relationship

4.4.3 Correlation Analysis

Identifying correlations between the features and the target variable gives insights into data relationships and how features impact the dependent variable. The study calculates Pearson correlation coefficient of every variable to understand possible relationships and relevance of said feature. The three EXT_SOURCE features show strong negative correlation with the dependent variable. A higher EXT_SOURCE value is an indicator of strong credit worthiness of the applicant. The DAYS_BIRTH feature shows positive correlation to one of the external sources feature which is an indication that client age may influence this feature.



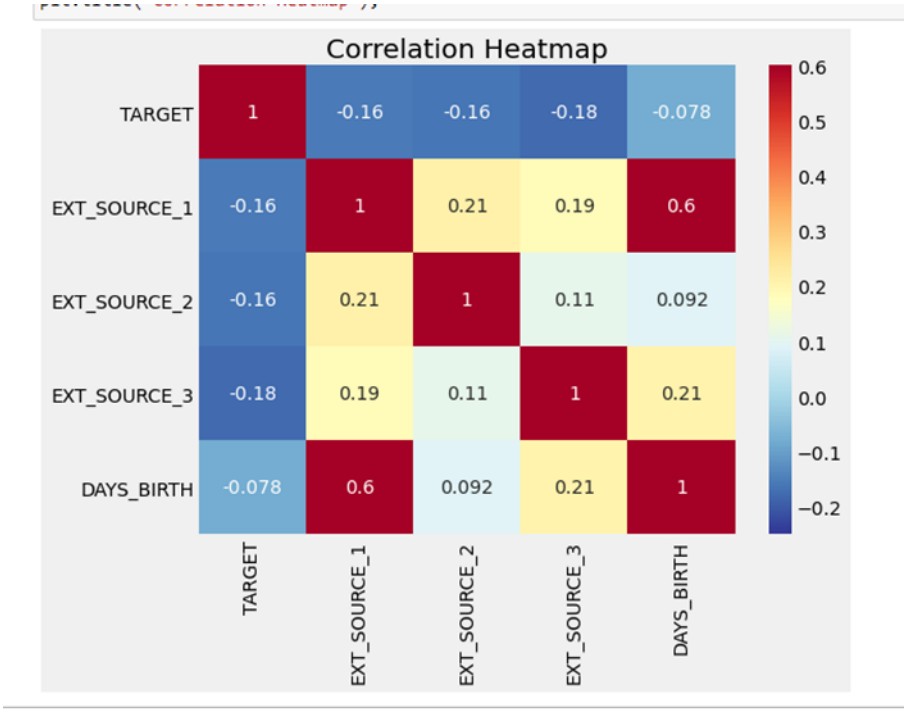


Figure 23.21: Home Credit Application dataset feature corelation

4.5 Data Transformations

4.5.1 Feature Encoding

Since most machine learning models only accept numerical variables, pre-processing the categorical variables becomes a necessary step. These categorical variables are converted to numbers such that the model is able to understand and extract valuable information.

Each of the 7 datasets which are part of the Home Credit case study have already been treated for feature elimination, missing value analysis and redundant feature treatment. There exist multiple categorical features which are of interest to this study in the datasets.

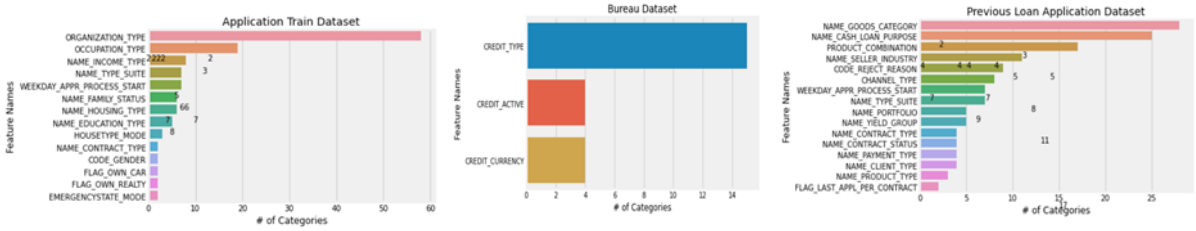


Figure 24.22: Home Credit Categorical Features distributions

This study employs three approaches for feature encoding listed below:

Label Encoding: In Label encoding, each label is converted into an integer value. This study creates a variable that contains the categories representing the categorical feature. If the categorical feature contains a high number of categories, this approach's efficacy suffers and thus, this approach is used only for those features where the number of categories is less than 3.

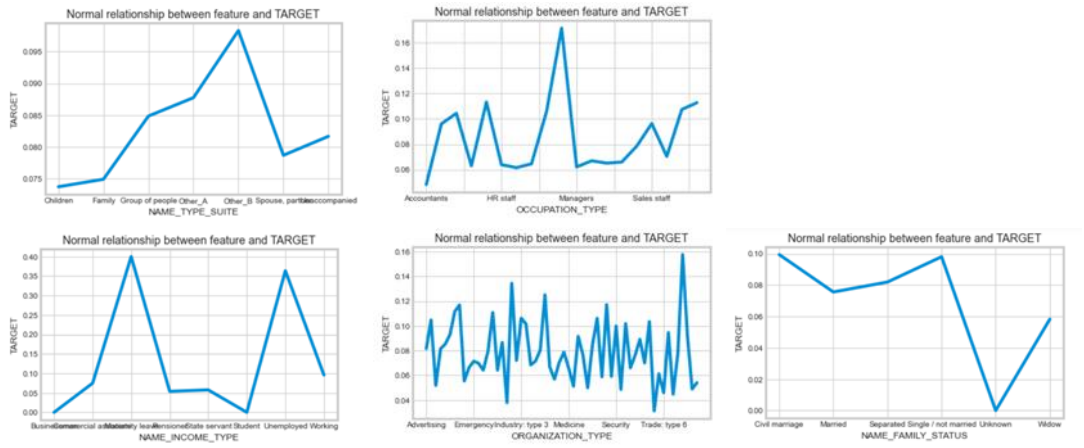
One Hot Encoding: In one hot encoding, for each level of a categorical feature, the study creates a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category. This study uses this form of encoding for the Bureau Balance and Instalment Payments datasets. A limitation of this approach is that when the number of categories for a variable is very high, it results in a large number of dummy features which are sparsely populated and may be corelated to each other. The Home Credit datasets already comprise of over 300 features and this approach of encoding negatively impacts the model performance. Thus, for application train, previous loans and bureau data

Weight of Evidence Encoding: In this approach, categorical variables with large number of categories are transformed to establish a monotonic relationship to the TARGET or dependent variable. Thus, the TARGET variable from application train, is added to the bureau, previous

loans, credit cards and point of sale datasets and weight of evidence transformation is applied to the remaining categorical variables on those datasets.

The table below showcases the relationship between the categorical variables before and after treatment with Weight of Evidence Encoding. As indicated by the charts, there exists a monotonic relationship between the treated categorical features and the TARGET variable.

Feature relationship to Target Before WoE Encoding



Feature relationship to Target After WoE Encoding

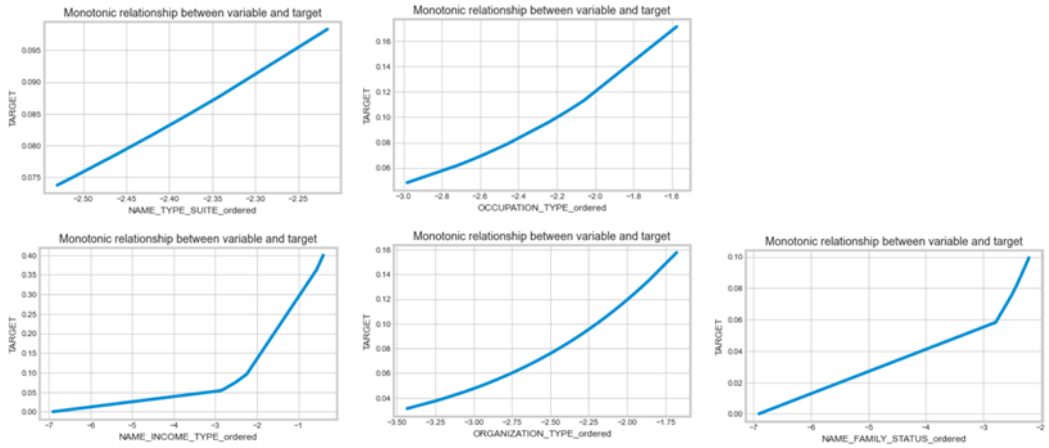


Figure 25.23: Home Credit Categorical Feature vs Target after WoE feature encoding

4.5.2 Feature Engineering

Feature engineering comprises of two parts: the first involves feature construction which is adding new features from the existing data, and the second part is feature selection which involves choosing only the most important features. The Home Credit dataset after data pre-

processing and EDA comprises of 52 features. The feature engineering methods used in this study are:

I. Polynomial Datapoints

This involves generation of new features that are a combination of multiple individual variables. "Interaction items" are features that capture interaction between two or more variables. Another way to generate polynomial features is to generate datapoints that are powers of existing features. Some of the interaction features generated in this study are as below:

- Annuity as a percentage of income: This feature is calculated as a derivative of loan annuity and loan amount.
- Loan Credit Amount as a percentage of Loan Annuity: A interaction between credit vs annuity as a feature.
- Loan Credit Amount as a percentage of Customer Income: A interaction feature between credit amount and income. If the value is very high, it is an anomaly and warrants further investigation to ensure credit safety.
- Loan Credit amount to goods price: Based on EDA graphs, if the credit amount requested is very high as compared to the goods price, the customer is not invested with his/her own capital and the risk of default is high. This polynomial feature would be interesting to analyse for behaviour
- External Sources Mean: A feature calculated as a power of the mean of each of the three external sources features. The external sources features express a high corelation with the target variable and this feature is of peak interest to analyse for its impact on predicting default.
- Employment to Age Ratio: This feature calculates the employment age to age of employee and seeks to determine if this feature impacts the target feature.
- Income to Organization Ratio: Organization Ratio has a number of categories that influences the risk of credit default. This feature explores the relation of Income to organization to explore if the target variable has a significant corelation with this variable.
- Income vs Number of children: The dataset pre-processing establishes that loan default cases are high in cases where number of children exceed 5 in number, this feature tries to establish income vs children relationship with target variable to see if there is a significant relationship or pattern with this hybrid variable.

Ten new polynomial features were generated and added to the Home Credit dataset bringing the feature count up to 62 features.

II. Domain knowledge features

In this method, aggregate features are calculated from the bureau, previous loan application, point of sale & credit card datasets as well as the instalment payment datasets and those aggregate features are merged with the application dataset to determine if those domain aggregate features significantly impact the target features. Towards that objective, the following approach is implemented:

Bureau Dataset:

The Credit days, bureau balance and the credit amount features when aggregated by the current loan application customer presents domain aggregate features of by statistical mean, min, max amounts. These features when merged with application dataset provides insights into how these features impact credit default for the customer.

Previous Application Dataset

- Polynomial feature of credit amount as a percentage of loan application amount is
- derived for previous loan applications
- The annuity amount, application amount, credit amount good price amount are aggregated by the current loan application customer and provide domain aggregate features of mean amounts. These features are merged with application dataset.

Point of Sale Dataset

The month balance amount for point of sales data is aggregated for the customer and this aggregated feature is added to the dataset

Instalment Payment Dataset: Instalment amounts, payment amounts are aggregated for the customer and added to the main dataset.

Credit Card Dataset: The credit card balance and payment amount data is aggregated for the customer and those features are added to the dataset.

Using aggregated mean, min and max values of features from previous application history, bureau history, instalment history adds domain value features to the dataset. The main dataset features total to a value of 168 features after generation of domain based aggregate features.

4.5.3 Feature Scaling

Feature scaling involves scaling the range of data for features in a dataset so that the feature values come down to a specific range. This technique is used in many credit scoring datasets and helps improve results of predictive models and achieve faster convergence. There are two popular methods involved in scaling viz: standardization & normalization. Standardization is to transform the data with mean zero and variance of 1 whereas, Normalization is about bounding the values between two numbers such as [-1,1] or [0,1]. This research applies min-max scalar on all datasets before applying the machine learning algorithms.

4.5.4 Class Balancing

This section analyses the various types of data sampling techniques available and which are applied on the Home Credit Scoring datasets and compares the performance of each sampling techniques to identify the best one for credit scoring datasets. The objective is also to articulate the model performance differences between application and no usage of class imbalance techniques. Another aspect is to identify which is the best performing class imbalance technique for credit scoring datasets. This would address the Research Question 1 of this study.

Data Balancing:

The application dataset dependent variable class distribution shows a large imbalance ratio of 92:8% for the 'TARGET' variable. Using the dataset as is, introduces bias on the model's performance when applied on unseen datasets. To remediate the same, multiple data balancing techniques are applied to the x_train and y_train datasets. This study applies 5 types of sampling techniques and analyse the performance of 13 models against each technique to identify which is the best sampling technique for credit scoring datasets and which model performs the best. Under sampling refers to a group of techniques designed to balance the class distribution for a classification dataset that has a skewed class distribution. Under sampling techniques remove examples from the training dataset that belong to the majority class in order to better balance the class distribution, such as reducing the skew from a 1:100 to a 1:10, 1:2, or even a 1:1 class distribution. This study uses two under sampling techniques: which only removes noisy and ambiguous points along the class boundary of the dataset.

- The Tomek under-sampling method identifies Tomek links which are pairs of instances of opposite classes who are their own nearest neighbours. In other words, they are pairs of opposing instances that have the smallest Euclidean distance to one another.

- The Edited Nearest Neighbours, or ENN technique, involves using *k=3* nearest neighbours to locate those examples in a dataset that are misclassified and that are then removed before a k=1 classification rule is applied.

Both the under-sampling techniques only remove noisy and ambiguous points along the class boundary of the dataset. As such, the resulting transformed dataset is not equally balanced. Table 4.1 showcases the reduction in majority class as implemented by the under-sampling techniques.

Oversampling imbalance algorithms oversample the minority class. The study uses the following two oversampling techniques:

- SMOTE: Synthetic Minority Oversampling Technique works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line thus creating new samples that are plausible
- ADASYN: Adaptive Synthetic Sampling is a modification to SMOTE and involves generating synthetic samples inversely proportional to the density of the examples in the minority class thereby generating more synthetic examples in regions of the feature space where the density of minority examples is low, and fewer or none where the density is high.

This study also uses a hybrid sampling technique which uses a combination of two sampling techniques to create a balanced dataset for modelling.

- The SMOTE+Tomek is implemented using the imblearn.combine library for this study. It reduces the majority class samples using Tomek to remove the noisy data and then applies SMOTE to equalize the class distribution. Table 4.1 showcases the uniform class distribution values post application of SMOTE+Tomek

Table 9.1: Data Balancing sample size and distribution summary

Sampling Category	Sampling method	Training sample size	Distribution of labels
Under-Sampling	Tomek Links	Before: (307511,74) After: (296127,74)	1 24804 0 271323
	Edited Nearest Neighbour (ENN)	Before: (307511,74) After: (245802,74)	1 24804 0 220998
Over-Sampling	SMOTE	Before: (307511,74) After: (564382,74)	1 282416 0 282416
	ADASYN	Before: (307511,74) After: (556895,74)	1 274479 0 282516
Hybrid-Sampling	SMOTE+TomekLink	Before: (307511,74) After: (562710,74)	1 281355 0 281355

4.5.4 Feature Selection

The training dataset was balanced using the SMOTE+Tomek sampling technique and comprised of 168 features across 0.56 million credit applications. The Home Credit datasets already contain a large number of features and feature engineering exercise increased the number of features available to a value of 168 features.

Datasets with high number of features increase training time for algorithms and may result in overfitting for modelling algorithms.

The feature selection methodology employed by this study is in two tranches:

In tranche I, feature importance from LightGBM classifier is employed to reduce the features to a manageable number. The feature selection cut-off is set to be 10% of the value of the feature with highest feature importance number. This exercise reduced the features from 168 to 74 features.

In tranche II, this dataset with 74 features was subjected to three feature selection techniques to identify the top 30 features from those methods. The feature selection methods used are:

- Step Forward Feature Selection: A feature selection wrapper technique that uses a sequence of steps to allow features to enter or leave the regression model one-at-a-time. Often this procedure converges to a subset of features
- Step Backward Feature Selection: A feature selection wrapper technique that starts with the entire set of features and works backward from there, removing features to find the optimal subset of a predefined size.
- Embedded Recursive Feature Elimination: A embedded feature selection technique that uses Random Forest model to identify sensitivity analysis on features and return the optimal subset of 30 features.

Three feature datasets are created based on the feature selection methods.

Furthermore, the common features between two of the above said three feature selection techniques are identified and three more feature-combination datasets are created.

Thus, six feature datasets are created as part of this study which are articulated in table below. During model evaluation, the study analyses which feature dataset results in the highest performance.

Table 10.2: Significant variables from 6 Feature selection methods

	Feature Selection Method	Significant Features	# of Features
Wrapper based Feature Selection Method	Step Forward Selection (RFRegressor)	'AMT_INCOME_TOTAL', 'AMT_ANNUITY', AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'CNT_FAM_MEMBERS', 'HOUR_APPR_PROCESS_START', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'NAME_TYPE_SUITE_ordered', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_DAYS_CREDIT_ENDDATE_MAX', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MEAN', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DBD_MAX', 'INS_DBD_SUM', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'	30
	Step Backward Selection(RandomForestClassifier)	['SK_ID_CURR', 'AMT_CREDIT', 'AMT_ANNUITY', 'DAYS_BIRTH', 'DAYS_ID_PUBLISH', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_3', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'NEW_CREDIT_TO_INCOME_RATIO', 'NEW_SOURCES_PROD', 'NEW_SCORES_STD', 'NAME_TYPE_SUITE_ordered', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'WEEKDAY_APPR_PROCESS_START_ordered', 'ORGANIZATION_TYPE_ordered', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_DAYS_CREDIT_UPDATE_MAX', 'PREV_APP_CREDIT_PERC_MAX', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_DAYS_DECISION_MAX', 'PREV_CNT_PAYMENT_MEAN', 'POS_MONTHS_BALANCE_MAX', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DPD_MAX', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'	30
Embedded Feature Selection Method	RFE Embedded RF Regressor	Unnamed: 0', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_ID_PUBLISH','CNT_FAM_MEMBERS', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'NEW_ANNUITY_TO_INCOME_RATIO', 'NEW_CREDIT_TO_ANNUITY_RATIO', 'NEW_CREDIT_TO_GOODS_RATIO', 'NEW_INC_BY_ORG', 'NEW_SOURCES_PROD', 'NEW_EXT_SOURCES_MEAN', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'OCCUPATION_TYPE_ordered', 'ORGANIZATION_TYPE_ordered', 'BURO_DAYS_CREDIT_ENDDATE_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MAX', PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_HOUR_APPR_PROCESS_START_MAX', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_RATE_DOWN_PAYMENT_MAX', 'PREV_CODE_REJECT_REASON_ordered_MEAN', 'POS_MONTHS_BALANCE_MAX', 'INS_DPD_MEAN', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_SUM'	30
Combination of Wrapper and Embedded Feature Selection Approach. (Identify features common to two approaches)	Features common to step_forward and RFE	['NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'INS_AMT_PAYMENT_SUM', 'INS_AMT_PAYMENT_MIN', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'TARGET']	14
	Features common to step_backward and RFE	['NAME_EDUCATION_TYPE_ordered', 'ORGANIZATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'DAYS_BIRTH', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_APP_CREDIT_PERC_MAX', 'POS_MONTHS_BALANCE_MAX', 'INS_AMT_PAYMENT_SUM', 'OBS_30_CNT_SOCIAL_CIRCLE', 'NEW_SOURCES_PROD', 'DAYS_ID_PUBLISH', 'TARGET']	14
	Features common to step_forward and step backward	['NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'INS_AMT_PAYMENT_MEAN', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_AMT_CREDIT_SUM_MEAN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'AMT_ANNUITY', 'NAME_TYPE_SUITE_ordered', 'INS_DAYS_ENTRY_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'OBS_30_CNT_SOCIAL_CIRCLE', 'TARGET']	16

4.6 Machine Learning Model Implementation

4.6.1. Evaluate data sampling techniques

One of the objectives of this study is to evaluate the impact of class imbalance techniques on credit scoring datasets.

The approach is to use two oversampling techniques (SMOTE and ADASYN), two under-sampling techniques (Tomek Links and ENN) and one hybrid sampling technique (SMOTE+Tomek) on the Home Credit dataset to create sampled datasets. The original class imbalance dataset is also used

Each of the sampled dataset is subjected to 13 classifiers and the performance of the models is compared with the evaluation criteria of Accuracy, AUC, Precision, Recall, F1 score and Kappa to identify which of the sampling techniques gave best results.

The original class imbalance dataset is also subjected to above models to ascertain the performance gap between sampled and class imbalanced Home Credit datasets.

The approach is to generates the following six datasets

Dataset1: The Home Credit imbalanced dataset which does not have any sampling technique applied to it.

Dataset2: SMOTE Oversampling technique applied dataset. The majority and minority samples are equally balanced.

Dataset3: ADASYN technique applied dataset. The majority and minority are not strictly balanced as ADASYN probabilities account of outlying minority datapoints and take those into consideration.

Dataset4: Tomek Links under sampling technique identified datapoints which are noisy or ambiguous and removed them from the majority class. The majority and minority classes are not balanced. Given the small number of majority datapoints removed, it can be ascertained that the Home Credit dataset does not have ambiguity or noise datapoints.

Dataset5: ENN under sampling technique. Again, a small number of majority class datapoints are reduced. This may negatively impact the efficacy of the models.

Dataset6: SMOTE+Tomek hybrid sampling technique. The majority and minority datapoints are equal and slightly lower than those of SMOTE technique which indicates that the under-sampling part of the hybrid model reduced those points.

The SMOTE+Tomek sampling technique is the best performing sampling technique on the Home Credit Dataset.

This study uses the SMOTE+Tomek data balanced dataset for feature selection and model training.

4.6.2. Evaluation and selection of best feature extraction technique

The study used three feature selection techniques to generate three feature datasets each comprising of 30 features. The study generated three more combination or hybrid feature datasets by selecting features that are common to two of the above feature selection techniques. Each of the datasets were tested using 13 base classifiers to ensure the performance does not include any classifier bias. The performance of the classifiers was evaluated on the basis of Accuracy, AUC, Precision, Recall, Kappa and F1 scores respectively.

The below table showcases the performance of the feature selection approaches.

Table 11.3: Feature selection dataset performance summary

Feature DataSet	Model	Accuracy	Test Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Dataset 1 Step Forward Selection (RFRegressor) Features : 30 'AMT_INCOME_TOTAL', 'AMT_ANNUITY', AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'CNT_FAM_MEMBERS', 'HOUR_APPR_PROCESS_START', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'NEW_ING_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'NAME_TYPE_SUITE_ordered', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_DAYS_CREDIT_ENDDATE_MAX', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MEAN', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DBD_MAX', 'INS_DBD_SUM', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'	Extra Trees Classifier	0.9715	0.9692	0.9918	0.9508	0.9816	0.9659	0.933	0.9334
	Light Gradient Boosting Machine	0.9549	0.9549	0.9781	0.9113	0.9984	0.9529	0.9098	0.9133
	Random Forest Classifier	0.9537	0.9541	0.9808	0.9099	0.9972	0.9515	0.9073	0.9108
	Gradient Boosting Classifier	0.9485	0.9354	0.9732	0.902	0.9945	0.946	0.8971	0.901
	Ada Boost Classifier	0.9205	0.9148	0.9636	0.8873	0.9504	0.9178	0.841	0.8429
	Decision Tree Classifier	0.9035	0.8924	0.9035	0.9123	0.8965	0.9043	0.8069	0.8071
	K Neighbors Classifier	0.7845	0.7724	0.8802	0.9408	0.7168	0.8136	0.569	0.599
	Quadratic Discriminant Analysis	0.7353	0.7257	0.8054	0.7706	0.7198	0.7443	0.4706	0.4718
	Ridge Classifier	0.6966	0.6825	0	0.7025	0.6944	0.6984	0.3932	0.3933
	Linear Discriminant Analysis	0.6966	0.6811	0.7608	0.7024	0.6943	0.6984	0.3932	0.3932
	Naive Bayes	0.6261	0.6147	0.6675	0.6802	0.6138	0.6453	0.2522	0.2537
	Logistic Regression	0.5919	0.5814	0.623	0.6273	0.5859	0.6058	0.1838	0.1843
	SVM - Linear Kernel	0.5201	0.5147	0	0.4608	0.6003	0.3879	0.0402	0.0703
Dataset 2 RFE Embedded RF Regressor Features 30 'Unnamed: 0', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_ID_PUBLISH', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'NEW_ANNUITY_TO_INCOME_RATIO', 'NEW_CREDIT_TO_ANNUITY_RATIO', 'NEW_CREDIT_TO_GOODS_RATIO', 'NEW_ING_BY_ORG', 'NEW_SOURCES_PROD', 'NEW_EXT_SOURCES_MEAN', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'OCCUPATION_TYPE_ordered', 'ORGANIZATION_TYPE_ordered', 'BURO_DAYS_CREDIT_ENDDATE_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MAX', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_HOUR_APPR_PROCESS_START_MAX', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_RATE_DOWN_PAYMENT_MAX', 'PREV_CODE_REJECT_REASON_ordered_MEAN', 'POS_MONTHS_BALANCE_MAX', 'INS_DPD_MEAN', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_SUM'	Extra Trees Classifier	0.9607	0.964	0.9895	0.9487	0.9721	0.9602	0.9214	0.9217
	Light Gradient Boosting Machine	0.9543	0.9542	0.9782	0.9122	0.9961	0.9523	0.9086	0.9119
	Random Forest Classifier	0.951	0.952	0.9804	0.912	0.9892	0.949	0.902	0.9048
	Gradient Boosting Classifier	0.9455	0.9351	0.9741	0.9046	0.9853	0.9432	0.891	0.894
	Ada Boost Classifier	0.9143	0.9047	0.9647	0.9018	0.925	0.9133	0.8287	0.829
	Decision Tree Classifier	0.8946	0.8821	0.8946	0.9027	0.8883	0.8954	0.7892	0.7893
	Quadratic Discriminant Analysis	0.718	0.7042	0.8033	0.8289	0.6785	0.7462	0.436	0.4472
	K Neighbors Classifier	0.7088	0.6815	0.7785	0.8161	0.6719	0.737	0.4175	0.4275
	Linear Discriminant Analysis	0.6978	0.6715	0.764	0.7031	0.6957	0.6994	0.3955	0.3955
	Ridge Classifier	0.6977	0.6845	0	0.7032	0.6956	0.6994	0.3955	0.3955
	Naive Bayes	0.6138	0.6025	0.6507	0.7709	0.5866	0.6662	0.2276	0.2398
	Logistic Regression	0.5811	0.5715	0.6029	0.633	0.5734	0.6017	0.1621	0.163
	SVM - Linear Kernel	0.5161	0.5047	0	0.5974	0.5393	0.484	0.0322	0.0533
Dataset 3 Step Backward Selection(RandomForestClassifier) Features 30 ['SK_ID_CURR', 'AMT_CREDIT', 'AMT_ANNUITY', 'DAYS_BIRTH', 'DAYS_ID_PUBLISH', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_3', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'NEW_CREDIT_TO_INCOME_RATIO', 'NEW_SOURCES_PROD', 'NEW_SCORES_STD', 'NAME_TYPE_SUITE_ordered', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'WEEKDAY_APPR_PROCESS_START_ordered', 'ORGANIZATION_TYPE_ordered', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_DAYS_CREDIT_UPDATE_MAX', 'PREV_APP_CREDIT_PERC_MAX', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_DAYS_DECISION_MAX', 'PREV_CNT_PAYMENT_MEAN', 'POS_MONTHS_BALANCE_MAX', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DPD_MAX', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'	Extra Trees Classifier	0.9542	0.935	0.9658	0.9357	0.9864	0.9458	0.9157	0.9158
	Light Gradient Boosting Machine	0.9425	0.9358	0.9542	0.9122	0.9825	0.9425	0.8925	0.9101
	Random Forest Classifier	0.9418	0.9257	0.9425	0.912	0.9754	0.9411	0.8258	0.9089
	Gradient Boosting Classifier	0.9425	0.9215	0.9257	0.9046	0.9158	0.9054	0.8058	0.9075
	Ada Boost Classifier	0.9189	0.9024	0.9217	0.9018	0.8922	0.9025	0.7927	0.9062
	Decision Tree Classifier	0.9052	0.8958	0.8925	0.9027	0.8285	0.8925	0.7157	0.9042
	Quadratic Discriminant Analysis	0.9046	0.8815	0.8588	0.8289	0.6978	0.7425	0.4384	0.8925
	K Neighbors Classifier	0.9014	0.7548	0.7928	0.8161	0.6915	0.7235	0.4057	0.6481
	Linear Discriminant Analysis	0.8914	0.6927	0.7758	0.7031	0.6914	0.6944	0.3854	0.3785
	Ridge Classifier	0.7542	0.6814	0.5488	0.7032	0.6825	0.6815	0.3822	0.3748
	Naive Bayes	0.6054	0.5525	0.5422	0.7709	0.5759	0.6652	0.2125	0.2398
	Logistic Regression	0.5928	0.5324	0.5244	0.633	0.5587	0.6158	0.1553	0.1548
	SVM - Linear Kernel	0.5051	0.4817	0.4587	0.5974	0.5258	0.4758	0.0345	0.0581

Dataset4 Features common to step_forward and RFE Features : ['NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'INS_AMT_PAYMENT_SUM', 'INS_AMT_PAYMENT_MIN', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'TARGET']	Light Gradient Boosting Machine	0.9537	0.9539	0.9762	0.9092	0.998	0.9515	0.9074	0.911
	Extra Trees Classifier	0.9529	0.9483	0.9828	0.9183	0.9866	0.9512	0.9059	0.908
	Random Forest Classifier	0.95	0.9182	0.9764	0.9057	0.9939	0.9477	0.9001	0.9037
	Gradient Boosting Classifier	0.9416	0.9222	0.9717	0.8911	0.9912	0.9385	0.8832	0.8878
	Ada Boost Classifier	0.9081	0.9012	0.957	0.8632	0.9484	0.9038	0.8162	0.8195
	Decision Tree Classifier	0.9029	0.8952	0.9029	0.9095	0.8977	0.9035	0.8058	0.8059
	Quadratic Discriminant Analysis	0.7054	0.7036	0.7753	0.7707	0.6817	0.7235	0.4108	0.4144
	Ridge Classifier	0.6851	0.6751	0	0.6868	0.6844	0.6856	0.3702	0.3702
	Linear Discriminant Analysis	0.6851	0.6651	0.7483	0.6868	0.6845	0.6857	0.3702	0.3702
	K Neighbors Classifier	0.669	0.6597	0.7281	0.7033	0.6582	0.68	0.3381	0.3389
	Naive Bayes	0.5919	0.5824	0.6198	0.7562	0.5692	0.6495	0.1838	0.1946
	Logistic Regression	0.5536	0.5429	0.5663	0.6678	0.5436	0.5994	0.1072	0.1101
	SVM - Linear Kernel	0.5186	0.5084	0	0.4786	0.5467	0.4409	0.0373	0.0507
Dataset5 Features common to step_backward and RFE Features : ['NAME_EDUCATION_TYPE_ordered', 'ORGANIZATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'DAYS_BIRTH', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_APP_CREDIT_PERC_MAX', 'POS_MONTHS_BALANCE_MAX', 'INS_AMT_PAYMENT_SUM', 'OBS_30_CNT_SOCIAL_CIRCLE', 'NEW_SOURCES_PROD', 'DAYS_ID_PUBLISH',	Light Gradient Boosting Machine	0.9516	0.9514	0.9753	0.9056	0.9973	0.9492	0.9031	0.907
	Extra Trees Classifier	0.9514	0.9527	0.9823	0.9166	0.9851	0.9496	0.9027	0.9049
	Random Forest Classifier	0.9488	0.9496	0.9755	0.9031	0.994	0.9464	0.8977	0.9015
	Gradient Boosting Classifier	0.9386	0.9235	0.9706	0.8848	0.9914	0.9351	0.8771	0.8822
	Ada Boost Classifier	0.9038	0.8945	0.9565	0.8661	0.9367	0.9	0.8075	0.8099
	Decision Tree Classifier	0.9019	0.8914	0.9019	0.9078	0.8971	0.9024	0.8037	0.8038
	Quadratic Discriminant Analysis	0.7103	0.7051	0.7764	0.7778	0.6853	0.7286	0.4206	0.4245
	Ridge Classifier	0.6768	0.6628	0	0.7375	0.6576	0.6953	0.3536	0.3562
	Linear Discriminant Analysis	0.6768	0.6684	0.7328	0.7376	0.6576	0.6953	0.3536	0.3562
	Naive Bayes	0.5679	0.5541	0.6013	0.8129	0.5455	0.6529	0.1357	0.1557
	K Neighbors Classifier	0.5555	0.5497	0.5782	0.6074	0.5503	0.5774	0.111	0.1116
	Logistic Regression	0.4999	0.4824	0.5424	0.0001	0.1714	0.0002	-0.0003	-0.009
	SVM - Linear Kernel	0.4956	0.4863	0	0.505	0.362	0.3456	-0.0088	-0.0119
Dataset6 Features common to step_forward and step backward Features : ['NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'INS_AMT_PAYMENT_MEAN', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_AMT_CREDIT_SUM_MEAN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'AMT_ANNUITY', 'NAME_TYPE_SUITE_ordered', 'INS_DAYS_ENTRY_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM',	Extra Trees Classifier	0.955	0.8685	0.9807	0.9136	0.9961	0.9531	0.91	0.9131
	Light Gradient Boosting Machine	0.9547	0.8815	0.9689	0.9098	0.9995	0.9526	0.9094	0.9131
	Random Forest Classifier	0.9536	0.8624	0.9687	0.9092	0.9977	0.9514	0.9071	0.9107
	Gradient Boosting Classifier	0.9498	0.9357	0.9649	0.901	0.9985	0.9472	0.8996	0.904
	Ada Boost Classifier	0.9241	0.9147	0.9552	0.8719	0.9736	0.9199	0.8482	0.8529
	Decision Tree Classifier	0.9066	0.8924	0.9066	0.9143	0.9004	0.9073	0.8132	0.8133
	Quadratic Discriminant Analysis	0.6679	0.6517	0.7246	0.7552	0.643	0.6946	0.3359	0.3411
	K Neighbors Classifier	0.622	0.6188	0.6654	0.6815	0.609	0.6432	0.2439	0.2457
	Ridge Classifier	0.5914	0.5751	0	0.629	0.585	0.6062	0.1829	0.1834
	Linear Discriminant Analysis	0.5914	0.5814	0.6261	0.629	0.5851	0.6062	0.1829	0.1834
	Naive Bayes	0.5904	0.5824	0.6294	0.713	0.5726	0.6351	0.1808	0.1865
	Logistic Regression	0.5449	0.5341	0.5658	0.4463	0.556	0.4951	0.0899	0.0917
	SVM - Linear Kernel	0.5081	0.4915	0	0.4775	0.5102	0.4223	0.0162	0.0184

The following observations are drawn based on the above table:

- The combination or hybrid feature selection datasets show comparable results to the feature selection techniques. The feature intersection approach between two feature selection methods reduces the feature count to 14-15 features but does not significantly impact the performance of the models.
- The best feature selection dataset is the one derived from the step forward feature selection wrapper method.

Going forward, the study uses the step forward feature selection dataset and the SMOTE+Tomek data balancing technique for model evaluation and optimization.

4.6.3. Classifier Training, Tuning, and Evaluation setup

The study having established the training and test datasets after evaluating the best data balancing and feature selection techniques, now compare the performance of base classifiers and ensembles. This study compares the performance of 13 classifiers using the Accuracy, precision, recall, Kappa, AUC and MCC metrics as the evaluation criteria. The model performance summary is depicted in table below:

Table 12.4: Base classifier model performance summary

Model	Accuracy	Test Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Sampling Treatment	Feature DataSet
Extra Trees Classifier	0.9665	0.9552	0.9918	0.9508	0.9816	0.9659	0.933	0.9334	SMOTE+Tomek majority:281355, minority:281355 Train after : (562710, 30)	Dataset 1 Step Forward Selection (RFRegressor) Features : 30 'AMT_INCOME_TOTAL', 'AMT_ANNUITY', AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'CNT_FAM_MEMBERS', 'HOUR_APPR_PROCESS_START', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'NAME_TYPE_SUITE_ordered', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_DAYS_CREDIT_ENDDATE_MAX', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MEAN', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DBD_MAX', 'INS_DBD_SUM', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'
Light Gradient Boosting Machine	0.9549	0.9549	0.9781	0.9113	0.9984	0.9529	0.9098	0.9133		
Random Forest Classifier	0.9537	0.9541	0.9808	0.9099	0.9972	0.9515	0.9073	0.9108		
Gradient Boosting Classifier	0.9485	0.9354	0.9732	0.902	0.9945	0.946	0.8971	0.901		
Ada Boost Classifier	0.9205	0.9148	0.9636	0.8873	0.9504	0.9178	0.841	0.8429		
Decision Tree Classifier	0.9035	0.8924	0.9035	0.9123	0.8965	0.9043	0.8069	0.8071		
K Neighbors Classifier	0.7845	0.7724	0.8802	0.9408	0.7168	0.8136	0.569	0.599		
Quadratic Discriminant Analysis	0.7353	0.7257	0.8054	0.7706	0.7198	0.7443	0.4706	0.4718		
Ridge Classifier	0.6966	0.6825	0	0.7025	0.6944	0.6984	0.3932	0.3933		
Linear Discriminant Analysis	0.6966	0.6811	0.7608	0.7024	0.6943	0.6984	0.3932	0.3932		
Naive Bayes	0.6261	0.6147	0.6675	0.6802	0.6138	0.6453	0.2522	0.2537		
Logistic Regression	0.5919	0.5814	0.623	0.6273	0.5859	0.6058	0.1838	0.1843		
SVM - Linear Kernel	0.5201	0.5147	0	0.4608	0.6003	0.3879	0.0402	0.0703		

The study having established the training and test datasets after evaluating the best data balancing and feature selection techniques compared the performance of 13 base classifiers and concluded that the best performing classifiers are:

- Light Gradient Boosting Machine
- Extra Trees Classifier
- Random Forest Classifier

These three classifiers are subjected to hyper-parameter optimization and for building ensemble models.

Optimization of Hyperparameters: This method involves tuning the parameters for the model to identify the ones that result in optimum performance and lowest error rate. For Hyperparameter tuning, pycaret.classification provides a method called tune_model. With tune_model, the study provides a custom set of hyper-parameters that needs to be tuned. The hyperparameters for each model however vary and thus a manual set of hyperparameters needs to be pre-determined for each model and passed onto the tune_model method. Each model is retrained with the optimized parameters to assess improvement in performance.

This step is repeated in a iterative manner dependent on the different ranges and sets of parameters used in tuning, till the final results are acceptable

Model evaluation:

Model Evaluation is the process through the quality of a system's predictions are quantified. To do this, the study measures the newly trained model performance on a new and independent dataset. This model compares labelled data with its own predictions.

Although training a model is an important step, generalizing the model to unseen data is an equally vital aspect to consider in any machine learning pipeline. While train accuracy helps to identify model over-fitting, the test metric identifies under-fitting probabilities and model performance in predicting the new data. The various models developed in this research are assessed by these measures to determine the best approach and solution to the problem of default prediction. The training accuracy, test accuracy, precision, recall, and f1 score are computed with 10-fold cross-validation using the cross_val_score from sklearn.model_selection. The Roc and Auc coefficient values are derived from sklearn.metrics using roc_curve and AUC functions and the kappa measure using cohen_kappa_score.

Light Gradient Boosting Machine

The LightGBM classifier was trained with the step forward feature selection dataset and tested against the test data. Following that, the model was further tuned to optimize for better AUC score.

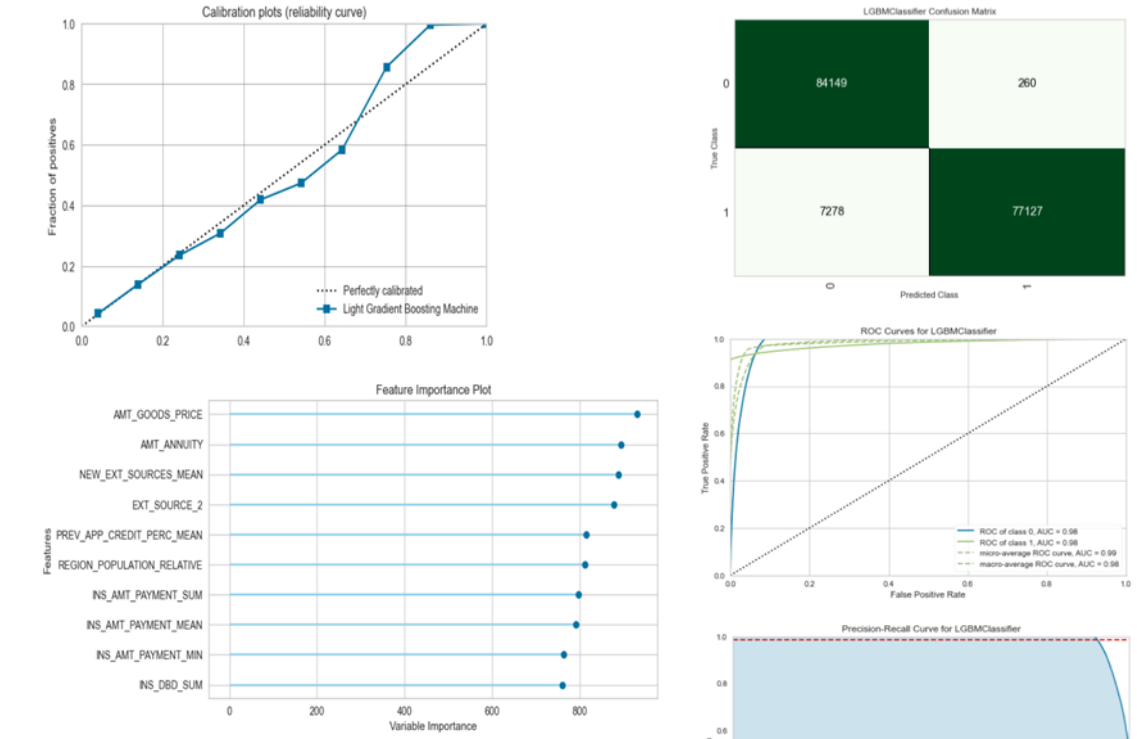
Default parameters:

```
boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
importance_type='split', learning_rate=0.1, max_depth=-1,
min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0
n_estimators=100, n_jobs=-1, num_leaves=31, objective=None,
random_state=123, reg_alpha=0.0, reg_lambda=0.0, silent='warn',
subsample=1.0, subsample_for_bin=200000, subsample_freq=0)
```

Tuned Parameters:

```
bagging_fraction=0.6, bagging_freq=2, feature_fraction=0.4,
boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
importance_type='split', learning_rate=0.1, max_depth=-1,
min_child_samples=41, min_child_weight=0.001, min_split_gain=0.9,
n_estimators=260, n_jobs=-1, num_leaves=70, objective=None,
random_state=123, reg_alpha=2, reg_lambda=3, silent='warn',
subsample=1.0, subsample_for_bin=200000, subsample_freq=0
```

Post tuning, the feature & bagging fraction parameters were modified along with the # estimators and leaves for the model. These changes resulted in a performance improvement for the model. The best performance score is Accuracy=0.9555 AUC = 0.9783 and Recall = 0.9138.



Parameters	
boosting_type	gbdt
class_weight	None
colsample_bytree	1.0
importance_type	split
learning_rate	0.1
max_depth	-1
min_child_samples	41
min_child_weight	0.001
min_split_gain	0.9
n_estimators	260
n_jobs	-1
num_leaves	70
objective	None
random_state	123
reg_alpha	2
reg_lambda	3
silent	warn
subsample	1.0
subsample_for_bin	200000
subsample_freq	0
bagging_fraction	0.6
bagging_freq	2
feature_fraction	0.4

Figure 26.24: LightGBM Model Performance Metrics

a) Extra Trees Classifier:

The Extra Trees classifier was trained with the step forward feature selection dataset and tested against the test data. Following that, the model was further tuned to optimize for better AUC score.

Default Parameters

```
bootstrap=False, ccp_alpha=0.0, class_weight=None,max_depth=None,
criterion='gini',max_features='auto',max_leaf_nodes=None,
```

```
max_samples=None,min_impurity_decrease=0.0,min_impurity_split=None
min_samples_leaf=1,min_samples_split=2,min_weight_fraction_leaf=0.
0, n_estimators=100, n_jobs=-1,oob_score=False, random_state=123,
verbose=0, warm_start=False)

Tuned Parameters
bootstrap=False, ccp_alpha=0.0, class_weight='balanced_subsample',
criterion='gini', max_depth=6, max_features=1.0, max_leaf_nodes=No
ne, max_samples=None, min_impurity_decrease=0, min_impurity_split=
None, min_samples_leaf=4, min_samples_split=7, min_weight_fraction
_leaf=0.0,n_estimators=200, n_jobs=-1, oob_score=False, random_sta
te=123, verbose=0, warm_start=False
```

Post tuning, the class weight, sample leaf size and split parameters were modified. The model did not show any performance improvement with the tuned parameters. The best performance score was 0.9665 whereas the accuracy post tuning parameters drops to 0.7913. Going forward, only the default parameters are chosen for this model.

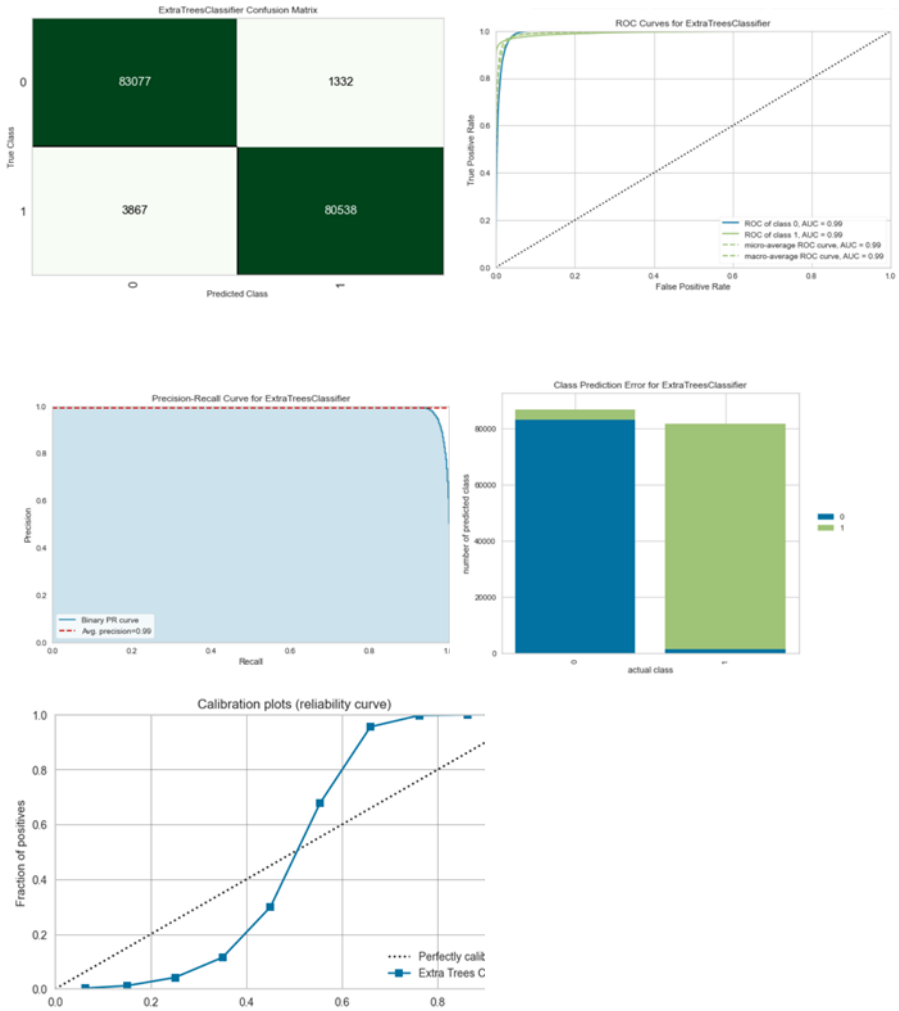


Figure 27.25: Extra Trees Classifier Model Performance Metrics

b) Random Forest Classifier:

The Random Forest classifier was trained with the step forward feature selection dataset and tested against the test data. Following that, the model was further tuned to optimize for better AUC score.

Default parameters

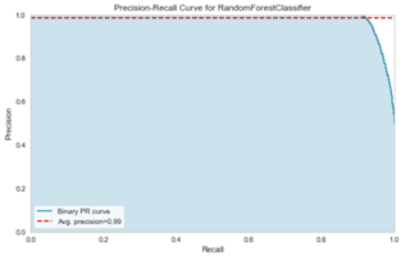
```
bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini'
, max_depth=None, max_features='auto', max_leaf_nodes=None, max_sa
mples=None, min_impurity_decrease=0.0, min_impurity_split=None, mi
n_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.
0, n_estimators=100, n_jobs=-1, oob_score=False, random_state=123,
verbose=0, warm_start=False
```

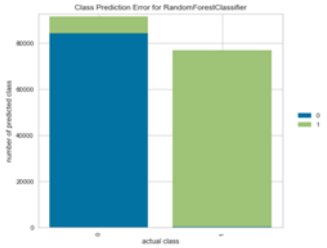
Tuned parameters

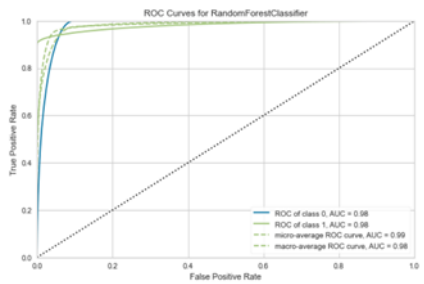
```
bootstrap=False, ccp_alpha=0.0,class_weight='balanced_subsample',
criterion='gini',max_depth=6, max_features='log2', max_leaf_nodes=
None,max_samples=None, min_impurity_decrease=0.001,min_impurity_sp
lit=None, min_samples_leaf=6,min_samples_split=9, min_weight_fract
ion_leaf=0.0,n_estimators=190, n_jobs=-1, oob_score=False, random_
state=123, verbose=0, warm_start=False
```

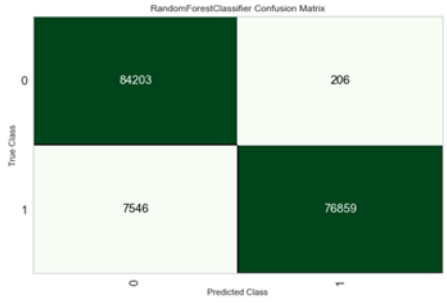
Post tuning, the class weight, sample leaf size and split parameters were modified. The model did not show any performance improvement with the tuned parameters. The best performance score was 0.9665 whereas the accuracy post tuning parameters drops to 0.8413. Going forward, only the default parameters are chosen for this model.

Parameters	
bootstrap	True
ccp_alpha	0.0
class_weight	None
criterion	gini
max_depth	None
max_features	auto
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0.0
n_estimators	100
n_jobs	-1
oob_score	False
random_state	123
verbose	0
warm_start	False









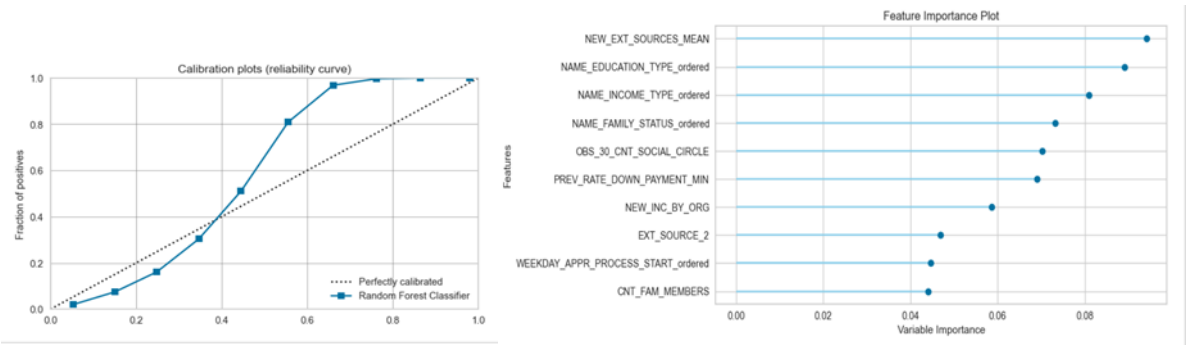


Figure 28.26: Random Forest Classifier Model Performance Metrics

4.6.3. Ensemble model and comparison to best performing classifier.

The top performing base classifiers on Home Credit dataset are 1) Light Gradient Boosting Machine (2) Extra Trees Classifier and (3) Random Forest Classifier.

The literature review of this study emphasized the focus on ensemble creation in credit scoring studies. Using the top three base classifiers, this study creates heterogenous ensembles and compare the performance of the ensembles to the base classifiers and compare the performances of the ensembles.

Bagging Ensemble:

Bagging or Bootstrap aggregating is an ensemble algorithm that reduces variance and avoids overfitting. The three base classifiers are submitted to the bagging classifier to build a heterogenous ensemble tested across 10 folds. The default parameters and performance results from bagging ensemble are given below:

Ensemble Parameters:

```
BaggingClassifier      base_estimator=ExtraTreesClassifier
      (bootstrap=False,ccp_alpha=0.0, class_weight=None,criterion='gini'
      ,max_depth=None, max_features='auto',max_leaf_nodes=None,max_sampl
      es=None,min_impurity_decrease=0.0,min_impurity_split=None,min_samp
      les_leaf1,min_samples_split=2,min_weight_fraction_leaf=0.0,n_estim
      ators=100,n_jobs=-1,oob_score=False,random_state=123,verbose=0,war
      m_start=False),bootstrap=True, bootstrap_features=False, max_featu
      res=1.0,max_samples=1.0, n_estimators=10, n_jobs=None,oob_score=Fa
      lse, random_state=123, verbose=0,warm_start=False),
BaggingClassifier   (base_estimator=LGBMClassifier
```

```
      (bagging_fraction=0.6,bagging_freq=2, boosting_type='gbdt',class_w
      eight=None,colsample_bytree=1.0,feature_fraction=0.4,importance_ty
      pe='split',learning_rate=0.1, max_depth=-1,min_child_samples=41,mi
      n_child_weight=0.001,min_split_gain=0.9,n_estimators=260, n_jobs=-
      1,num_leaves=70, objective=None,random_state=123, reg_alpha=2,reg_
      lambda=3, silent='warn',subsample=1.0,subsample_for_bin=200000,sub
      sample_freq=0),bootstrap=True, bootstrap_features=False, max_featu
      res=1.0,max_samples=1.0, n_estimators=10, n_jobs=None,oob_score=Fa
      lse, random_state=123, verbose=0,warm_start=False),
BaggingClassifier(base_estimator=RandomForestClassifier
      (bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gi
      ni',max_depth=None, max_features='auto', max_leaf_nodes=None, m
      ax_samples=None, min_impurity_decrease=0.0, min_impurity_split=N
      one, min_samples_leaf=1, min_samples_split=2, min_weight_fracti
      on_leaf=0.0, n_estimators=100, n_jobs=-1, oob_score=False, ran
      dom_state=123, verbose=0, warm_start=False),bootstrap=True, boot
      strap_features=False, max_features=1.0,max_samples=1.0, n_estimato
      rs=10, n_jobs=None,oob_score=False, random_state=123, verbose=0,wa
      rm_start=False)]
```

Table 13.5: Bagging Ensemble Performance Summary

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9523	0.9806	0.9072	0.9971	0.9500	0.9046	0.9083
1	0.9547	0.9818	0.9123	0.9968	0.9527	0.9093	0.9126
2	0.9518	0.9810	0.9070	0.9963	0.9496	0.9036	0.9073
3	0.9520	0.9807	0.9066	0.9972	0.9498	0.9041	0.9078
4	0.9536	0.9808	0.9096	0.9974	0.9515	0.9073	0.9108
5	0.9519	0.9801	0.9070	0.9965	0.9496	0.9038	0.9075
6	0.9520	0.9806	0.9069	0.9967	0.9497	0.9039	0.9076
7	0.9523	0.9809	0.9072	0.9971	0.9500	0.9045	0.9082
8	0.9519	0.9809	0.9068	0.9968	0.9497	0.9039	0.9076
9	0.9524	0.9805	0.9074	0.9972	0.9502	0.9048	0.9085
Mean	0.9525	0.9808	0.9078	0.9969	0.9503	0.9050	0.9086
SD	0.0009	0.0004	0.0017	0.0003	0.0010	0.0018	0.0016

Blending Ensemble

Blending or Voting Classifier is an ensemble algorithm that combines the above mentioned three base classifiers and generates average predicted probabilities to predict the risk of credit default. The three base classifiers are submitted to the blending classifier to build a heterogenous ensemble tested across 10 folds. The default parameters and performance results from bagging ensemble are given below:

Ensemble Parameters:

```
VotingClassifier(estimators=[('et',
```

```
                                ExtraTreesClassifier(bootstrap=False,
ccp_alpha=0.0,class_weight=None,criterion='gini',max_depth=None,
max_features='auto',max_leaf_nodes=None,max_samples=None,min_impurity_d
ecrease=0.0,min_impurity_split=None,min_samples_leaf=1,min_samples_spli
t=2,min_weight_fraction_leaf=0.0,n_estimators=100, n_jobs=-1,oob_score=
Fa...max_depth=None,max_features='auto',max_leaf_nodes=None,max_samples
=None,min_impurity_decrease=0.0,min_impurity_split=None,min_samples_lea
f=1,min_samples_split=2,min_weight_fraction_leaf=0.0,n_estimators=100,
n_jobs=-1, oob_score=False,random_state=123,verbose=0,warm_start=False)
)], flatten_transform=True, n_jobs=-1, verbose=False, voting='soft', we
ights=None)
```

Table 14.6: Blending Ensemble Performance Summary

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9587	0.9890	0.9189	0.9983	0.9570	0.9174	0.9203
1	0.9598	0.9899	0.9216	0.9978	0.9582	0.9196	0.9223
2	0.9573	0.9894	0.9167	0.9977	0.9555	0.9146	0.9176
3	0.9572	0.9895	0.9166	0.9976	0.9554	0.9144	0.9174
4	0.9588	0.9892	0.9195	0.9980	0.9571	0.9176	0.9204
5	0.9576	0.9893	0.9175	0.9974	0.9558	0.9151	0.9180
6	0.9576	0.9893	0.9175	0.9975	0.9558	0.9152	0.9182
7	0.9581	0.9894	0.9183	0.9977	0.9564	0.9162	0.9191
8	0.9583	0.9891	0.9185	0.9978	0.9565	0.9165	0.9194
9	0.9580	0.9886	0.9175	0.9984	0.9563	0.9161	0.9191
Mean	0.9581	0.9893	0.9183	0.9978	0.9564	0.9163	0.9192
SD	0.0008	0.0003	0.0014	0.0003	0.0008	0.0015	0.0014

Stacking Ensemble:

Stacking is an ensemble algorithm that uses a meta model and generates the final credit default prediction using the prediction of the above mentioned three base classifiers. The meta model for the Stacking ensemble is set as the LightGBM model. The three base classifiers are submitted to the blending classifier to build a heterogenous ensemble tested across 10 folds. The default parameters and performance results from bagging ensemble are given below:

Ensemble Parameters:

```
StackingClassifier(cv=StratifiedKFold(n_splits=10,
        random_state=RandomState(MT19937) at 0x27776DCBD40, shuffle=Fal
se), estimators=[('et', ExtraTreesClassifier(bootstrap=False,
ccp_alpha=0.0,class_weight=None,criterion='gini',max_depth=None,max_fea
tures='auto',max_leaf_nodes=None,max_samples=None,min_impurity_decrease
=0.0,min_impurity_split=None,min_sample... random_state=123, verbose=0,
 warm_start=False))], final_estimator=LogisticRegression(C=1.0, class_w
eight=None,dual=False,fit_intercept=True,intercept_scaling=1,l1_ratio=N
one,max_iter=1000,multi_class='auto',n_jobs=None, penalty='l2',random_s
```

```
tate=123,solver='lbfgs',tol=0.0001, verbose=0,warm_start=False),
n_jobs=-1, passthrough=True, stack_method='auto', verbose=0)
```

Table 15.7: Stacking Ensemble Performance Summary

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	
0	0.9486	0.9825	0.9125	0.9915	0.9531	0.9148	0.9205	
1	0.9476	0.9815	0.9135	0.9925	0.9564	0.9175	0.9215	
2	0.9472	0.9865	0.9145	0.9935	0.9597	0.9128	9189	
3	0.9458	0.9845	0.9175	0.9976	0.9586	0.9136	0.9148	
4	0.9478	0.9848	0.9185	0.9968	0.9542	0.9148	0.9125	
5	0.9456	0.9857	0.9168	0.9943	0.9548	0.9146	0.9145	
6	0.9465	0.9836	0.9135	0.9968	0.9535	0.9143	0.9157	
7	0.9482	0.9875	0.9146	0.9943	0.9568	0.9184	0.9156	
8	0.9468	0.9814	0.9137	0.9937	0.9518	0.9168	0.9158	
9	0.9457	0.9835	0.9159	0.9918	0.9528	0.9128	0.9168	
Mean	0.9478	0.9854	0.9155	0.9924	0.9518	0.9134	0.9124	
SD	0.0008	0.0003	0.0014	0.0003	0.0008	0.0015	0.0014	

Overall, the generation of the ensemble models is time & computationally intensive. Preliminary observations showcase the performance of the ensembles to be at par with that of the base classifier learners.

4.7 Model Interpretability

This section details the methods and steps followed to generate feature explanations for the best performant model on the Home Credit dataset. The study employs LIME to generate local explanations of the features and Tree SHAP to generate both global and local explanations using the nuanced services and visualizations provided by the above mentioned two libraries.

4.7.1 LIME Explanation Generation

LIME generates an explanation for a prediction from the components of an interpretable model which resembles the black-box model at the vicinity of the point of interest and which is trained over a new data representation to ensure interpretability. To ensure that the explanation is interpretable, LIME distinguishes an *interpretable representation* from the original feature space that the model uses. The interpretable representation has to be understandable to humans, so its dimension is not necessarily the same as the dimension of the original feature space.
The approach taken to generate LIME Explanations are:

I. The featured and data sampled dataset and test datasets are prepared and the best performing model which is LightGBM is selected. The predicted probability of default is generated using the LightGBM's predict probability function. Lime libraries are imported.

II. Using the model and the dataset, a LIMETabularExplainer instance is generated which can further be used to generate explanations for instances. The setup uses the following parameters:

    - Training data = x_train (the feature rich and sampled dataset)
    - feature_names = the list of features generated by the top feature selection algorithm for this study
    - class names = 'TARGET' (the dependent variable for this study)
    - mode = 'classification' (given the problem statement of predicting default is a binary classification problem)
    - verbose = 'true' (This prints linear mode's local prediction values)

III. The instances of interest are determined from the dataset for generation of local interpretations. The test dataset presents unforeseen data for the model and two datapoints from the test dataset are used in this study.

- Default datapoint. This is a datapoint from the test dataset where the actual loan application was a default case and the LightGBM model also predicted the label correctly as a default case.
- Non-Default datapoint. A loan application from the test dataset where the model as well as the actual application determined this application to be a non-default case.

LIME explains the various features of interest and their influence on the prediction for both the above datasets

IV. The study uses the following visualization features from LIME to interpret results from the features that influenced prediction by the model in the default and non-default datapoints.

show_in_notebook: This explanation plot visualization comprises of three parts. The most important features influencing the prediction are detailed in the middle section. Given the Home Credit dataset prediction case is one of binary classification, the feature influence is depicted in two colours. The default class attributes are highlighted in orange whereas the non-default attributes are highlighted in blue. The prediction probabilities are highlighted on the left section of the report.

Pyplot Feature report: This visualization technique shows the feature importances and their influence on the datapoint. Attributes influencing towards default are highlighted in green whereas attributes influencing towards non-default are highlighted in red.

4.7.2 SHAP Explanation Generation

The SHAP explainer calculates the variable importance values using the Shapley values from game theory. This study uses the Tree SHAP to generate global and local explanations in the following manner.

I. Shap libraries are imported, the train and test datasets and the best performing model which is LightGBM is used to generate a explainer instance using shap.TreeExplainer()
II. The shap_values and expected values are calculated using the explainer on the test dataset. These values can be used to visualize feature importance and influence of attributes on the dependent variable at a global interpretation level.
III. Local interpretability is generated by using a subset of datapoints from the test dataset that includes both default case as well as non-default case and using the visualization technique of force_plots and dependence plots, the influence of the features is explained using the Tree SHAP library.

4.8 Resources

4.8.1 Hardware Resources

In this study, all the implementations are performed on a personal computer with the following configurations:

Processor: Intel® Core™ i7 8th gen 8550U Processor

Memory: 8 GB LPDDR3 2133MHz

GPU: Integrated Intel UHD Graphics 620

Operating System. Windows 10 64-bit

4.8.2 Software Resources

The model for classifying home credit default is implemented using the Python framework. Several open source libraries such as pandas, numpy, matplotlib, seaborn,pylab and scipy are leveraged to implement the data-preprocessing steps and also for visualization aid during EDA. Data balancing evaluation leverages imblearn whereas feature extraction and scaling were implemented using Sklearn. The predictive modelling steps were implemented using a

combination of Sklearn and pycaret. Model explainability leverages the LIME and Tree SHAP libraries.

The important Python packages referred to along with versions are listed in table 22.

Table 16.8: Python packages

Library name	Version
imblearn	0.7.0
lime	0.2.0.1
matplotlib	3.1.3
Pycaret	0.17.3
NumPy	1.18.1
pandas	1.0.1
python	3.7.6
scikit-learn	0.23.2
scipy	1.4.1
seaborn	0.11.0
shap	0.36.0
xgboost	1.1.1

4.9 Summary

The design and execution of the study can be broadly divided into three steps. The first stage involves cleaning anomalies in the data, transformed by outlier fix, feature encoded, detailed analysis through EDA graphs and plots. The dataset is then transformed with the implementation of various categories of class sampling methods to smooth out the unbalanced target class, along with the application of feature extraction methods for choosing only the significant features to make them suitable for modelling.

In the next step, thirteen base classifier models are built on the data balanced and feature selected dataset and the model performance is evaluated and compared to identify the top three performing base classifiers. These top three models are subjected to hyperparameter optimization. Three heterogenous ensembles are built using the top three base classifiers. The performance of the ensembles and the top performing classifier are presented for comparison.

The generation of explanations is another key focus of this study. Tree SHAP and LIME are used to generate global as well as local interpretations and to identify feature importance and the impact of features on model's predictability.

CHAPTER 5: RESULTS AND DISCUSSIONS

5.1 Introduction

In this section, the results from the approach /implementations presented in the study are discussed. Firstly, the chapter reviews the results achieved from various data balancing techniques and compares the results to identify the pros and cons of each technique.

Secondly, a comparison of results of the three feature selection approaches used is presented and the results are discussed. Having identified the best data sampled and feature engineered dataset, this chapter then compares the results of thirteen base classifiers and identifies the top three performing classifiers. The optimization results on those three classifiers are discussed and inferences drawn on which approach is the best. Given the trend of building ensembles, this study builds three heterogenous ensembles using the top three classifiers and compares the results of the ensembles to the top performing classifier. The feature importance and impact on the predictive correctness of the model is explained using LIME and SHAP.

5.2 Feature Selection technique evaluation and results

The study created six feature selected datasets using three feature selection methods (two wrapper and one embedded), and a combination of features that intersect the features between two feature selection methods each.

The conclusions drawn from the feature selection datasets are as follows:

- All three-feature selection method-based datasets highlight the features generated through feature engineering prominently as significant features. This reflects positively on the manual feature engineering approach of creating polynomial features and domain based aggregated features
- Step forward selection wrapper-based feature selection approach stands out as the best performing feature dataset for Home Credit. It comprises of 30 features.
- The combination datasets that comprise of intersection features from two feature selection techniques comprise of 14 to 16 features. While this reduces the number of features from those of individual feature selection techniques, there is no performance gains observed on the models. Selecting features from multiple feature selection techniques is a good practice but this study results show that it does not always result in performance gains for the models.

Table 17.1: Feature Selection Technique Performance Summary

Feature DataSet	Model	Accuracy	Test Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Dataset 1 Step Forward Selection (RFRegressor) Features : 30 'AMT_INCOME_TOTAL', 'AMT_ANNUITY', AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'CNT_FAM_MEMBERS', 'HOUR_APPR_PROCESS_START', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'NAME_TYPE_SUITE_*ordered*', 'NAME_INCOME_TYPE_*ordered*', 'NAME_EDUCATION_TYPE_*ordered*', 'NAME_FAMILY_STATUS_*ordered*', 'WEEKDAY_APPR_PROCESS_START_*ordered*', 'BURO_DAYS_CREDIT_ENDDATE_MAX', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MEAN', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DBD_MAX', 'INS_DBD_SUM', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'	Extra Trees Classifier	0.9665	0.9552	0.9918	0.9508	0.9816	0.9659	0.933	0.9334
	Light Gradient Boosting Machine	0.9549	0.9549	0.9781	0.9113	0.9984	0.9529	0.9098	0.9133
	Random Forest Classifier	0.9537	0.9541	0.9808	0.9099	0.9972	0.9515	0.9073	0.9108
	Gradient Boosting Classifier	0.9485	0.9354	0.9732	0.902	0.9945	0.946	0.8971	0.901
	Ada Boost Classifier	0.9205	0.9148	0.9636	0.8873	0.9504	0.9178	0.841	0.8429
	Decision Tree Classifier	0.9035	0.8924	0.9035	0.9123	0.8965	0.9043	0.8069	0.8071
	K Neighbors Classifier	0.7845	0.7724	0.8802	0.9408	0.7168	0.8136	0.569	0.599
	Quadratic Discriminant Analysis	0.7353	0.7257	0.8054	0.7706	0.7198	0.7443	0.4706	0.4718
	Ridge Classifier	0.6966	0.6825	0	0.7025	0.6944	0.6984	0.3932	0.3933
	Linear Discriminant Analysis	0.6966	0.6811	0.7608	0.7024	0.6943	0.6984	0.3932	0.3932
	Naive Bayes	0.6261	0.6147	0.6675	0.6802	0.6138	0.6453	0.2522	0.2537
	Logistic Regression	0.5919	0.5814	0.623	0.6273	0.5859	0.6058	0.1838	0.1843
	SVM - Linear Kernel	0.5201	0.5147	0	0.4608	0.6003	0.3879	0.0402	0.0703
Dataset 2 RFE Embedded RF Regressor Features 30 'Unnamed: 0', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_ID_PUBLISH', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'NEW_ANNUITY_TO_INCOME_RATIO', 'NEW_CREDIT_TO_ANNUITY_RATIO', 'NEW_CREDIT_TO_GOODS_RATIO', 'NEW_INC_BY_ORG', 'NEW_SOURCES_PROD', 'NEW_EXT_SOURCES_MEAN', 'NAME_INCOME_TYPE_*ordered*', 'NAME_EDUCATION_TYPE_*ordered*', 'NAME_FAMILY_STATUS_*ordered*', 'OCCUPATION_TYPE_*ordered*', 'ORGANIZATION_TYPE_*ordered*', 'BURO_DAYS_CREDIT_ENDDATE_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MAX', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_HOUR_APPR_PROCESS_START_MAX', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_RATE_DOWN_PAYMENT_MAX', 'PREV_CODE_REJECT_REASON_*ordered*_MEAN', 'POS_MONTHS_BALANCE_MAX', 'INS_DPD_MEAN', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_SUM'	Extra Trees Classifier	0.9607	0.964	0.9895	0.9487	0.9721	0.9602	0.9214	0.9217
	Light Gradient Boosting Machine	0.9543	0.9542	0.9782	0.9122	0.9961	0.9523	0.9086	0.9119
	Random Forest Classifier	0.951	0.952	0.9804	0.912	0.9892	0.949	0.902	0.9048
	Gradient Boosting Classifier	0.9455	0.9351	0.9741	0.9046	0.9853	0.9432	0.891	0.894
	Ada Boost Classifier	0.9143	0.9047	0.9647	0.9018	0.925	0.9133	0.8287	0.829
	Decision Tree Classifier	0.8946	0.8821	0.8946	0.9027	0.8883	0.8954	0.7892	0.7893
	Quadratic Discriminant Analysis	0.718	0.7042	0.8033	0.8289	0.6785	0.7462	0.436	0.4472
	K Neighbors Classifier	0.7088	0.6815	0.7785	0.8161	0.6719	0.737	0.4175	0.4275
	Linear Discriminant Analysis	0.6978	0.6715	0.764	0.7031	0.6957	0.6994	0.3955	0.3955
	Ridge Classifier	0.6977	0.6845	0	0.7032	0.6956	0.6994	0.3955	0.3955
	Naive Bayes	0.6138	0.6025	0.6507	0.7709	0.5866	0.6662	0.2276	0.2398
	Logistic Regression	0.5811	0.5715	0.6029	0.633	0.5734	0.6017	0.1621	0.163
	SVM - Linear Kernel	0.5161	0.5047	0	0.5974	0.5393	0.484	0.0322	0.0533
Dataset 3 Step Backward Selection(RandomForestClassifier) Features 30 ['SK_ID_CURR', 'AMT_CREDIT', 'AMT_ANNUITY', 'DAYS_BIRTH', 'DAYS_ID_PUBLISH', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_3', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'NEW_CREDIT_TO_INCOME_RATIO', 'NEW_SOURCES_PROD', 'NEW_SCORES_STD', 'NAME_TYPE_SUITE_*ordered*', 'NAME_INCOME_TYPE_*ordered*', 'NAME_EDUCATION_TYPE_*ordered*', 'NAME_FAMILY_STATUS_*ordered*', 'WEEKDAY_APPR_PROCESS_START_*ordered*', 'ORGANIZATION_TYPE_*ordered*', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_DAYS_CREDIT_UPDATE_MAX', 'PREV_APP_CREDIT_PERC_MAX', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_DAYS_DECISION_MAX', 'PREV_CNT_PAYMENT_MEAN', 'POS_MONTHS_BALANCE_MAX', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DPD_MAX', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'	Extra Trees Classifier	0.9542	0.935	0.9658	0.9357	0.9864	0.9458	0.9157	0.9158
	Light Gradient Boosting Machine	0.9425	0.9358	0.9542	0.9122	0.9825	0.9425	0.8925	0.9101
	Random Forest Classifier	0.9418	0.9257	0.9425	0.912	0.9754	0.9411	0.8258	0.9089
	Gradient Boosting Classifier	0.9425	0.9215	0.9257	0.9046	0.9158	0.9054	0.8058	0.9075
	Ada Boost Classifier	0.9189	0.9024	0.9217	0.9018	0.8922	0.9025	0.7927	0.9062
	Decision Tree Classifier	0.9052	0.8958	0.8925	0.9027	0.8285	0.8925	0.7157	0.9042
	Quadratic Discriminant Analysis	0.9046	0.8815	0.8588	0.8289	0.6978	0.7425	0.4384	0.8925
	K Neighbors Classifier	0.9014	0.7548	0.7928	0.8161	0.6915	0.7235	0.4057	0.6481
	Linear Discriminant Analysis	0.8914	0.6927	0.7758	0.7031	0.6914	0.6944	0.3854	0.3785
	Ridge Classifier	0.7542	0.6814	0.5488	0.7032	0.6825	0.6815	0.3822	0.3748
	Naive Bayes	0.6054	0.5525	0.5422	0.7709	0.5759	0.6652	0.2125	0.2398
	Logistic Regression	0.5928	0.5324	0.5244	0.633	0.5587	0.6158	0.1553	0.1548
	SVM - Linear Kernel	0.5051	0.4817	0.4587	0.5974	0.5258	0.4758	0.0345	0.0581
Dataset4 Features common to step_forward and RFE Features : ['NAME_EDUCATION_TYPE_*ordered*', 'NAME_FAMILY_STATUS_*ordered*', 'NAME_INCOME_TYPE_*ordered*', 'CNT_FAM_MEMBERS', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'INS_AMT_PAYMENT_SUM', 'INS_AMT_PAYMENT_MIN', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'TARGET']	Light Gradient Boosting Machine	0.9537	0.9539	0.9762	0.9092	0.998	0.9515	0.9074	0.911
	Extra Trees Classifier	0.9529	0.9483	0.9828	0.9183	0.9866	0.9512	0.9059	0.908
	Random Forest Classifier	0.95	0.9182	0.9764	0.9057	0.9939	0.9477	0.9001	0.9037
	Gradient Boosting Classifier	0.9416	0.9222	0.9717	0.8911	0.9912	0.9385	0.8832	0.8878
	Ada Boost Classifier	0.9081	0.9012	0.957	0.8632	0.9484	0.9038	0.8162	0.8195
	Decision Tree Classifier	0.9029	0.8952	0.9029	0.9095	0.8977	0.9035	0.8058	0.8059
	Quadratic Discriminant Analysis	0.7054	0.7036	0.7753	0.7707	0.6817	0.7235	0.4108	0.4144
	Ridge Classifier	0.6851	0.6751	0	0.6868	0.6844	0.6856	0.3702	0.3702
	Linear Discriminant Analysis	0.6851	0.6651	0.7483	0.6868	0.6845	0.6857	0.3702	0.3702
	K Neighbors Classifier	0.669	0.6597	0.7281	0.7033	0.6582	0.68	0.3381	0.3389
	Naive Bayes	0.5919	0.5824	0.6198	0.7562	0.5692	0.6495	0.1838	0.1946
	Logistic Regression	0.5536	0.5429	0.5663	0.6678	0.5436	0.5994	0.1072	0.1101
	SVM - Linear Kernel	0.5186	0.5084	0	0.4786	0.5467	0.4409	0.0373	0.0507
Dataset5 Features common to step_backward and RFE Features : ['NAME_EDUCATION_TYPE_*ordered*', 'ORGANIZATION_TYPE_*ordered*', 'NAME_FAMILY_STATUS_*ordered*', 'NAME_INCOME_TYPE_*ordered*', 'CNT_FAM_MEMBERS', 'DAYS_BIRTH', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_APP_CREDIT_PERC_MAX', 'POS_MONTHS_BALANCE_MAX', 'INS_AMT_PAYMENT_SUM', 'OBS_30_CNT_SOCIAL_CIRCLE', 'NEW_SOURCES_PROD', 'DAYS_ID_PUBLISH',	Light Gradient Boosting Machine	0.9516	0.9514	0.9753	0.9056	0.9973	0.9492	0.9031	0.907
	Extra Trees Classifier	0.9514	0.9527	0.9823	0.9166	0.9851	0.9496	0.9027	0.9049
	Random Forest Classifier	0.9488	0.9496	0.9755	0.9031	0.994	0.9464	0.8977	0.9015
	Gradient Boosting Classifier	0.9386	0.9235	0.9706	0.8848	0.9914	0.9351	0.8771	0.8822
	Ada Boost Classifier	0.9038	0.8945	0.9565	0.8661	0.9367	0.9	0.8075	0.8099
	Decision Tree Classifier	0.9019	0.8914	0.9019	0.9078	0.8971	0.9024	0.8037	0.8038
	Quadratic Discriminant Analysis	0.7103	0.7051	0.7764	0.7778	0.6853	0.7286	0.4206	0.4245
	Ridge Classifier	0.6768	0.6628	0	0.7375	0.6576	0.6953	0.3536	0.3562
	Linear Discriminant Analysis	0.6768	0.6684	0.7328	0.7376	0.6576	0.6953	0.3536	0.3562
	Naive Bayes	0.5679	0.5541	0.6013	0.8129	0.5455	0.6529	0.1357	0.1557
	K Neighbors Classifier	0.5555	0.5497	0.5782	0.6074	0.5503	0.5774	0.111	0.1116
	Logistic Regression	0.4999	0.4824	0.5424	0.0001	0.1714	0.0002	-0.0003	-0.009
	SVM - Linear Kernel	0.4956	0.4863	0	0.505	0.362	0.3456	-0.0088	-0.0119

Dataset6 Features common to step_forward and step backward Features : ['NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'INS_AMT_PAYMENT_MEAN', 'NAME_INCOME_TYPE_ordered', 'CNT_FAM_MEMBERS', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_AMT_CREDIT_SUM_MEAN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'AMT_ANNUITY', 'NAME_TYPE_SUITE_ordered', 'INS_DAYS_ENTRY_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM',	Extra Trees Classifier	0.955	0.8685	0.9807	0.9136	0.9961	0.9531	0.91	0.9131
	Light Gradient Boosting Machine	0.9547	0.8815	0.9689	0.9098	0.9995	0.9526	0.9094	0.9131
	Random Forest Classifier	0.9536	0.8624	0.9687	0.9092	0.9977	0.9514	0.9071	0.9107
	Gradient Boosting Classifier	0.9498	0.9357	0.9649	0.901	0.9985	0.9472	0.8996	0.904
	Ada Boost Classifier	0.9241	0.9147	0.9552	0.8719	0.9736	0.9199	0.8482	0.8529
	Decision Tree Classifier	0.9066	0.8924	0.9066	0.9143	0.9004	0.9073	0.8132	0.8133
	Quadratic Discriminant Analysis	0.6679	0.6517	0.7246	0.7552	0.643	0.6946	0.3359	0.3411
	K Neighbors Classifier	0.622	0.6188	0.6654	0.6815	0.609	0.6432	0.2439	0.2457
	Ridge Classifier	0.5914	0.5751	0	0.629	0.585	0.6062	0.1829	0.1834
	Linear Discriminant Analysis	0.5914	0.5814	0.6261	0.629	0.5851	0.6062	0.1829	0.1834
	Naive Bayes	0.5904	0.5824	0.6294	0.713	0.5726	0.6351	0.1808	0.1865
	Logistic Regression	0.5449	0.5341	0.5658	0.4463	0.556	0.4951	0.0899	0.0917
	SVM - Linear Kernel	0.5081	0.4915	0	0.4775	0.5102	0.4223	0.0162	0.0184

5.3 Model performance Analysis

The study identified the best sampling technique and feature selection technique on the Home Credit dataset and generated the training and test datasets.

The dataset was subjected to 13 classifiers and the performance of the models was evaluated on the basis of Accuracy, Precision, Recall, AUC, F1 Score and Kappa. The top three classifiers were Light Gradient Boosting Machine, Extra Trees Classifier and Random Forest Classifiers. The performance of all 13 classifiers is listed in the table below against the SMOTE+Tomek balanced and Step forward Selection based feature dataset:

Table 18.2: Model Performance Metrics against Balanced and Feature Selected Dataset

Model	Accuracy	Test Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Sampling Treatment	Feature DataSet
Extra Trees Classifier	0.9665	0.9552	0.9918	0.9508	0.9816	0.9659	0.933	0.9334	SMOTE+Tomek majority:281355, minority:281355 Train after : (562710, 30)	Dataset 1 Step Forward Selection (RFRegressor) Features : 30 'AMT_INCOME_TOTAL', 'AMT_ANNUITY', AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'CNT_FAM_MEMBERS', 'HOUR_APPR_PROCESS_START', 'EXT_SOURCE_2', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'NEW_INC_BY_ORG', 'NEW_EXT_SOURCES_MEAN', 'NAME_TYPE_SUITE_ordered', 'NAME_INCOME_TYPE_ordered', 'NAME_EDUCATION_TYPE_ordered', 'NAME_FAMILY_STATUS_ordered', 'WEEKDAY_APPR_PROCESS_START_ordered', 'BURO_DAYS_CREDIT_ENDDATE_MAX', 'BURO_AMT_CREDIT_SUM_MEAN', 'BURO_AMT_CREDIT_SUM_SUM', 'PREV_APP_CREDIT_PERC_MEAN', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'PREV_RATE_DOWN_PAYMENT_MIN', 'INS_NUM_INSTALMENT_VERSION_NUNIQUE', 'INS_DBD_MAX', 'INS_DBD_SUM', 'INS_AMT_PAYMENT_MIN', 'INS_AMT_PAYMENT_MEAN', 'INS_AMT_PAYMENT_SUM', 'INS_DAYS_ENTRY_PAYMENT_MEAN'
Light Gradient Boosting Machine	0.9549	0.9549	0.9781	0.9113	0.9984	0.9529	0.9098	0.9133		
Random Forest Classifier	0.9537	0.9541	0.9808	0.9099	0.9972	0.9515	0.9073	0.9108		
Gradient Boosting Classifier	0.9485	0.9354	0.9732	0.902	0.9945	0.946	0.8971	0.901		
Ada Boost Classifier	0.9205	0.9148	0.9636	0.8873	0.9504	0.9178	0.841	0.8429		
Decision Tree Classifier	0.9035	0.8924	0.9035	0.9123	0.8965	0.9043	0.8069	0.8071		
K Neighbors Classifier	0.7845	0.7724	0.8802	0.9408	0.7168	0.8136	0.569	0.599		
Quadratic Discriminant Analysis	0.7353	0.7257	0.8054	0.7706	0.7198	0.7443	0.4706	0.4718		
Ridge Classifier	0.6966	0.6825	0	0.7025	0.6944	0.6984	0.3932	0.3933		
Linear Discriminant Analysis	0.6966	0.6811	0.7608	0.7024	0.6943	0.6984	0.3932	0.3932		
Naive Bayes	0.6261	0.6147	0.6675	0.6802	0.6138	0.6453	0.2522	0.2537		
Logistic Regression	0.5919	0.5814	0.623	0.6273	0.5859	0.6058	0.1838	0.1843		
SVM - Linear Kernel	0.5201	0.5147	0	0.4608	0.6003	0.3879	0.0402	0.0703		

Hyper parameter Optimization was performed on the top three classifiers and LightGBM emerged as the best performing classifier post hyper parameter optimization. The conclusion drawn from the analysis and results from section 4.6.3 are:

- Hyper parameter optimization positively influenced the performance of the LightGBM classifier.

- Hyper parameter optimization does not always result in positive gains. The performance of the Extra Trees Classifier as well as the Random Forest Classifier showed poor results as compared to the default parameters.

The study created three heterogenous ensembles using bagging, blending and stacking classifiers and compared the performance of the ensembles to the LightGBM classifier. The ensemble scores were comparable to the top performing base classifiers of LightGBM and Extra Trees Classifier. The study concludes that application of feature engineering, feature selection and sampling techniques on the Home Credit dataset generated strong learners from the LightGBM and Extra Trees base classifiers. Ensembles are relevant and their efficacy stands out when they are run on weak learners which is not the scenario on the Home Credit dataset. In the context of this dissertation, the conclusion drawn is that strong base classifiers when tested on feature engineered, feature selected, data balanced dataset showcase performance metrics that are comparable to ensembles.

5.4 LIME and SHAP explanation interpretation

This section describes the results from the two model interpretability techniques used by this study. The train dataset is balanced using SMOTE+Tomek and the best feature engineering technique applied is step forward feature selection. The best performing classifier identified is LightGBM and has been used to for model interpretability explainers.

5.4.1 LIME explanations

The two datapoints used to generate local interpretations are created from the test dataset as that presents unforeseen data for the model. The two datapoints and the local interpretability results are explained below:

- Default datapoint. This is a datapoint from the test dataset where the actual loan application was a default case and the LightGBM model also predicted the label correctly as a default case.

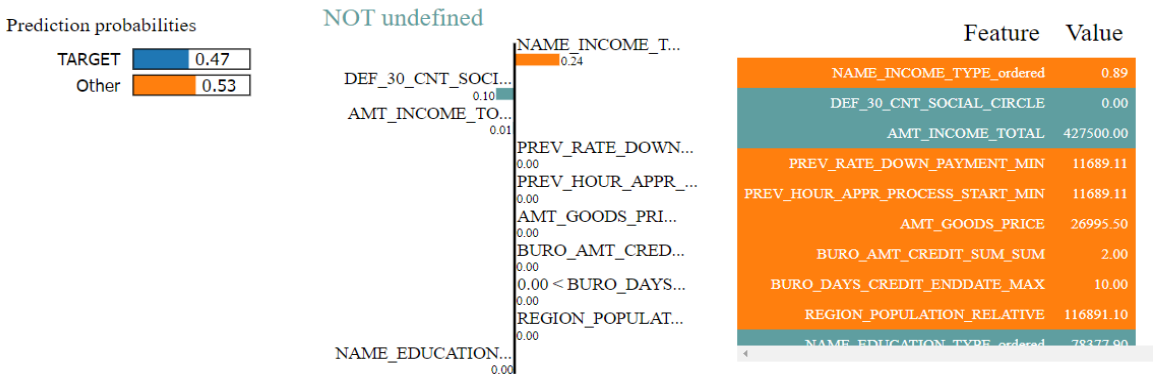


Figure 29.1: Variable importance and prediction probability visualization in LIME

The figures 5.1 & 5.2shows the explanations for the model predictions on a default case that is correctly identified by the model. Figure 5.1 is the 'show_in_notebook' view which should be analysed in three regions. The left side of the plot depicts predicted probabilities which are in close vicinity of each other but trend more towards the 'other' value which is why the model predicted the value correctly as a default case. On the right side, the 'NAME_INCOME_TYPE_ORDERED', 'PREV_RATE_DOWN_PAYMENT_MIN', 'PREV_HOUR_APPR_PROCESS_START_MIN', 'AMT_GOODS_PRICE' features which are marked in orange influence the model to predict this datapoint as a default whereas 'DEF_30_CNT_SOCIAL_CIRCLE', 'AMT_INCOME_TOTAL' have a tendency to influence the model towards non-default case. Feature importance is depicted in the middle portion of the plot.

Figure 5.2 is the pyplot visualization which lists the feature importance that influenced the results. Features in green have a positive corelation with the target whereas features in red have a negative co-relation. The inferences and explanations drawn from the feature importance are given below:

- Name_income_type_ordered feature has a positive co-relation with default cases. Thus, applicants with income type categories which have higher value for this feature are more likely to default.
- Amount_Income feature has a negative co-relation with default cases. Applicants with higher incomes are less likely to default

- DEF_30_CNT_SOCIAL_CIRCLE feature has a strong negative corelation in this case. Applicants with lower negative features are less likely to default.

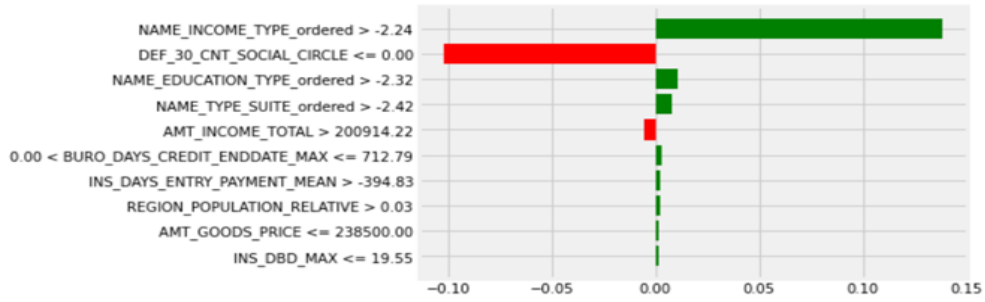


Figure 30.2: Variable importance visualization in LIME

- Non-Default datapoint. A loan application from the test dataset where the model as well as the actual application determined this application to be a non-default case.

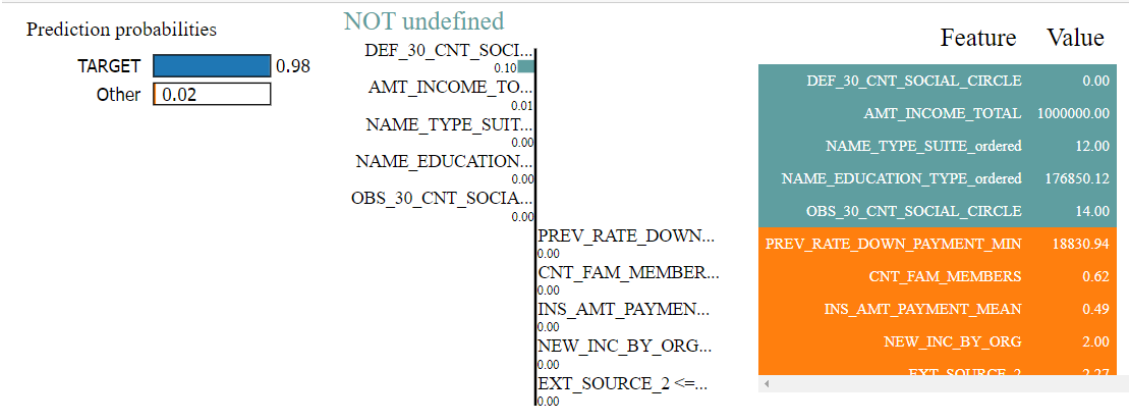


Figure 31.3: Variable importance and prediction probability visualization in LIME

Figure 5.3 and Figure 5.4 show the local explanations, probabilities and feature importance for a datapoint which is non-default case in real and also correctly predicted as non-default by the model. Figure5.3 is the 'show_in_notebook' view of the datapoint. The prediction probabilities depicted on the left side of the visualization show a resounding 0.98% view towards 'target' which led the model to predict this as a non-default case. On the right side of the plot, the features in blue influence the model to predict this as a non-default case whereas the features in orange influence this model towards a default case.

Figure 5.4 is the pyplot that elaborates on the features by importance. The following inferences can be drawn for this datapoint.

- DEF_30_CNT_SOCIAL_CIRCLE has a negative corelation with the prediction to default. Applicants with higher negative values for this feature have a lesser probability of default.

- Amount Income total feature has negative corelation. Thus, applicants with higher income have lesser probability of default.
- Name_Type_Suite feature has a positive corelation with the probability of default. Thus, applicants with higher values of this feature are more likely to default.

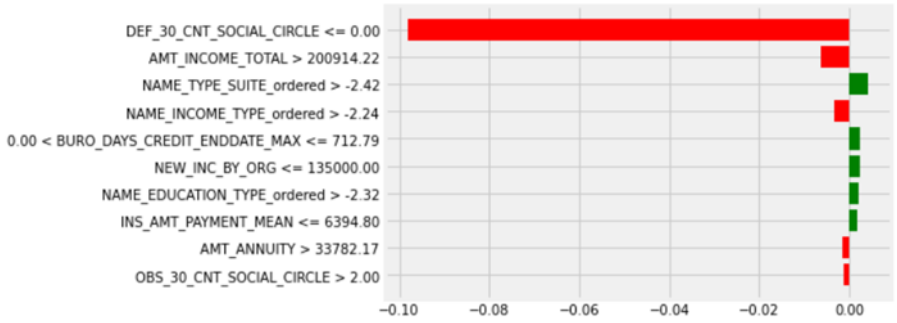


Figure 32.4: Variable importance visualization in LIME

5.4.2 SHAP explanations

Using the shap_values generated on the dataset, the following visualization plots show feature importance and influence of features on the dependent variable.

The variable importance plot in SHAP lists the most significant features in descending order. The features are ranked by averaging their absolute SHAP values. 'New_Ext_Sources_Mean', 'Weekday_apr_process_start_ordered', 'Name_Income_Type_ordered' and 'Name_Family_status_ordered' are four features that are in the top five features ranked by SHAP variable importance plot. Each of those four features are generated by this study using either polynomial feature generation or domain-based feature generation. This concludes that the feature engineering exercise of generating features is a good practice for credit scoring datasets.

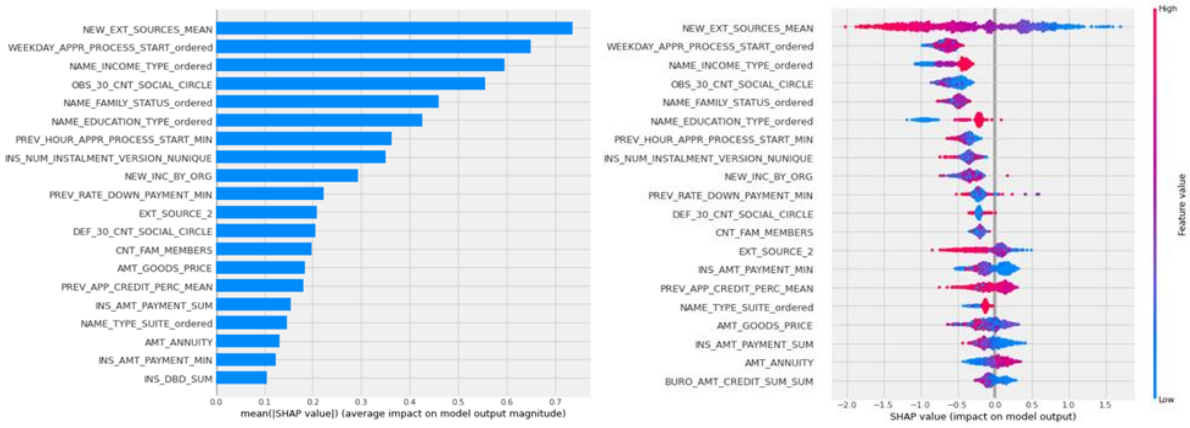


Figure 33.5: SHAP – Variable importance summary plot

The Shap summary value plot using the collective SHAP values for features shows how much each predictor contributes, either positively or negatively, to the target variable. This provides global interpretability to the features by being able to show the positive or negative relationship for each variable with the target. The 'New_Ext_Sources_Mean' feature shows a high and negative corelation with the target variable. The 'Weekday_apr_process_start_ordered', 'Name_Income_Type_ordered' and 'Name_Family_status_ordered' features each have a negative corelation with the target variable for their distributions.

For local interpretability using SHAP, a subset of the test dataset that includes loan application that are default as well as 9 other applications which are marked as non-defaults are chosen. Each observation has its own SHAP values which greatly increases its transparency. The Home Credit business can use these local interpretations to explain the rationale behind approval or rejection for a loan application. The following visualization techniques from SHAP are used to generate the rationale for model's prediction results and conclusions:

- Decision plot: The plot shows how the model arrived at the final prediction and the impact of each feature towards arriving at that decision. Each of the lines depict a data point. The 9 solid blue lines are for 9 datapoints which are non-default cases. The red dashed line is for a default case. For all cases, the decision plot starts at the baseline near 1 at the bottom of the graph and each new feature adds its influence from bottom to top. Following are the observations and inferences from the same.
    - For the non-default cases, the decision plot started at baseline level which is near 0.9 and most of the features pulled the decision towards a non-default case. The new_ext_sources_mean, ext_source_2 and income_by_organization features influenced the model decision towards default but the influence was weak as

compared to the collective influence of other features. The model thus predicted these datapoints to be non-default cases.

- For the default case, the decision plot started at the baseline level which is near default case. Features from the 'prev_rate_down_payment_min' started influencing the model prediction towards non-default with 'weekday_appr_process_start' being the strongest influencer towards non-default. The ext_source2 and income_by_organization features have a weak influence on the model towards default. The 'new_ext_sources_mean' feature has a strong influence towards default and resulted in this datapoint being predicted as a default case

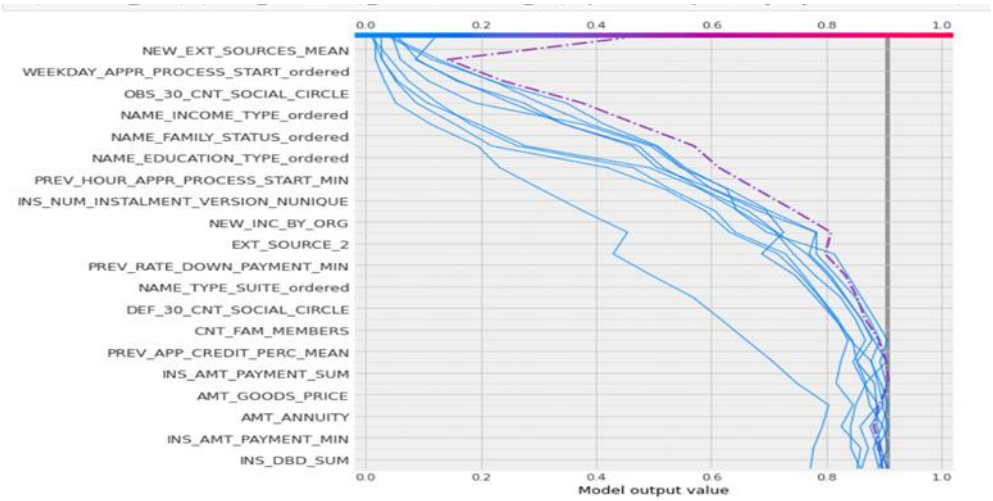


Figure 34.6: SHAP – Decision plot

- Force plot: This plot shows how the probability of prediction is impacted by features. For the default datapoint below, the force plot actual value of -0.48 is much lower than the base value of 2.34 which is an indication that the model predicts the result for said observation as a default value. The weekday_appr_process_start ordered and obs_30_cnt_social_circle features push the prediction to the left side and have a negative impact on the prediction. The 'new_ext_sources_mean' feature value of 0.1618 is lower than the mean value and thus has a positive impact on the prediction value and thus pushes the prediction to the right side

Figure 35.7: SHAP – Force plot

5.5 Answering Research Questions

Question 1: What is the impact of class imbalance on credit scoring datasets. Which sampling technique works best for this domain?

Answer:

The study used two oversampling techniques, two under sampling techniques, one hybrid sampling technique on the Home Credit dataset.

The performance of the sampling techniques is measured by comparing the performance of 13 classifiers. This ensures there is no bias towards one particular classifier. The evaluation criteria for identifying the best model are Accuracy, Precision, AUC, F1 and Kappa.

The below table summarizes the performance of thirteen machine learning algorithms when applied to five balanced datasets as well as the original class imbalanced dataset. The following observations or results are concluded based on the below:

- The dataset with no sampling technique applied shows clear underperformance as compared to other sampled dataset performance. This highlights the importance of class imbalance treatment on credit scoring datasets
- Under sampling techniques like Tomek and ENN show poor performance as compared to oversampling techniques. The under-sampling techniques only removed noisy and ambiguous samples (e.g., Tomek links) and thus for this credit scoring dataset, that sampling technique did not yield major performance improvement.
- Oversampling techniques like SMOTE and ADASYN both show strong performance on the model accuracy and other estimators. SMOTE marginally performs better on this dataset.
- The hybrid sampling technique of SMOTE+Tomek yields the best results out of all sampling techniques used. This hybrid method of under-sampling and over-sampling technique helps overcome the disadvantage of e.g., not taking into account adjacent examples of other classes that lead to over-fitting. SMOTE+Tomek is finalized as the best sampling technique in this research.

The study concludes that the hybrid approach of SMOTE+Tomek is the best performing sampling technique for the Home Credit dataset.

Table 19.3: Data Balancing method performance summary

	Model	Accuracy	Test Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
No Class Imbalance Applied	Logistic Regression	0.9197	0.9183	0.5293	0	0	0	0	0
	Ridge Classifier	0.9197	0.9183	0	0	0	0	0	0
	Naive Bayes	0.9196	0.9182	0.5801	0.0001	0.083	0.0002	0	0.0004
	Ada Boost Classifier	0.9196	0.9022	0.7359	0.0065	0.482	0.0129	0.0107	0.0484
	Gradient Boosting Classifier	0.9196	0.901	0.7383	0.0046	0.48	0.009	0.0075	0.0403
	Light Gradient Boosting Machine	0.9196	0.895	0.737	0.0031	0.478	0.0062	0.0051	0.0332
	Random Forest Classifier	0.9195	0.897	0.7085	0.0083	0.443	0.0163	0.0135	0.0518
	Linear Discriminant Analysis	0.9193	0.9052	0.7278	0.0121	0.426	0.0235	0.0191	0.0606
	Extra Trees Classifier	0.9193	0.9023	0.6997	0.0065	0.38	0.0127	0.0101	0.0406
	Quadratic Discriminant Analysis	0.916	0.8955	0.7122	0.0561	0.356	0.0968	0.0766	0.1147
	K Neighbors Classifier	0.9146	0.8945	0.5241	0.0093	0.113	0.0171	0.005	0.0097
	Decision Tree Classifier	0.8522	0.8355	0.5371	0.1617	0.139	0.1494	0.069	0.0693
	SVM - Linear Kernel	0.8508	0.8258	0	0.0774	0.075	0.0203	-9E-04	-0.001
SMOTE Balanced	Light Gradient Boosting Machine	0.9537	0.9541	0.9763	0.9092	0.998	0.9515	0.9073	0.9109
	Extra Trees Classifier	0.9527	0.9547	0.9828	0.9182	0.986	0.951	0.9054	0.9076
	Random Forest Classifier	0.9504	0.9513	0.9767	0.9064	0.994	0.9481	0.9008	0.9043
	Gradient Boosting Classifier	0.942	0.9244	0.9721	0.8915	0.992	0.9389	0.884	0.8886
	Ada Boost Classifier	0.9088	0.8915	0.9575	0.8677	0.945	0.9048	0.8175	0.8203
	Decision Tree Classifier	0.9036	0.8942	0.9036	0.91	0.898	0.9041	0.8071	0.8072
	Quadratic Discriminant Analysis	0.7054	0.6951	0.7747	0.7719	0.681	0.7237	0.4109	0.4146
	Ridge Classifier	0.6842	0.6724	0	0.6856	0.684	0.6846	0.3684	0.3684
	Linear Discriminant Analysis	0.6842	0.6725	0.7476	0.6856	0.684	0.6846	0.3684	0.3685
	K Neighbors Classifier	0.6681	0.6789	0.7263	0.7018	0.657	0.6789	0.3362	0.337
	Naive Bayes	0.5913	0.5844	0.6187	0.7552	0.569	0.6488	0.1827	0.1934
	Logistic Regression	0.5538	0.5499	0.5654	0.6689	0.544	0.5998	0.1076	0.1106
	SVM - Linear Kernel	0.513	0.5048	0	0.7342	0.524	0.5628	0.026	0.0373
ADASYN Balanced	Light Gradient Boosting Machine	0.9533	0.9528	0.9758	0.9071	0.998	0.9504	0.9065	0.9102
	Extra Trees Classifier	0.952	0.9526	0.9822	0.9148	0.987	0.9495	0.9039	0.9063
	Random Forest Classifier	0.9498	0.9494	0.9761	0.9036	0.994	0.9466	0.8994	0.903
	Gradient Boosting Classifier	0.9417	0.9348	0.9715	0.8894	0.991	0.9376	0.8832	0.8878
	Ada Boost Classifier	0.9088	0.8952	0.9569	0.8648	0.945	0.9033	0.8173	0.8203
	Decision Tree Classifier	0.903	0.8947	0.9031	0.909	0.896	0.9023	0.8061	0.8062
	Quadratic Discriminant Analysis	0.7061	0.6941	0.7751	0.7688	0.678	0.7204	0.4132	0.4168
	Ridge Classifier	0.684	0.6811	0	0.6748	0.681	0.6779	0.3678	0.3679
	Linear Discriminant Analysis	0.684	0.6742	0.7477	0.6748	0.681	0.6779	0.3678	0.3679
	K Neighbors Classifier	0.6637	0.6547	0.7215	0.6893	0.65	0.6688	0.3278	0.3284
	Naive Bayes	0.5871	0.5742	0.6162	0.7452	0.561	0.6401	0.178	0.188
	Logistic Regression	0.5493	0.5341	0.5612	0.6195	0.537	0.5753	0.1005	0.1017
	SVM - Linear Kernel	0.5091	0.4975	0	0.5855	0.422	0.4569	0.0202	0.0286
TOMEK Balanced	Logistic Regression	0.916	0.9168	0.526	0	0	0	0	0
	Ridge Classifier	0.916	0.9168	0	0	0	0	0	0
	Ada Boost Classifier	0.916	0.9169	0.7351	0.009	0.495	0.0177	0.0147	0.0575
	Light Gradient Boosting Machine	0.916	0.8924	0.7374	0.0061	0.497	0.012	0.01	0.0473
	Gradient Boosting Classifier	0.9159	0.8952	0.7375	0.0066	0.479	0.013	0.0108	0.0482
	Naive Bayes	0.9158	0.9051	0.5786	0.0003	0.125	0.0006	0.0001	0.0017
	Random Forest Classifier	0.9158	0.9041	0.7097	0.0098	0.442	0.0192	0.0157	0.0557
	Linear Discriminant Analysis	0.9156	0.9025	0.7271	0.0132	0.429	0.0256	0.0207	0.0633
	Extra Trees Classifier	0.9156	0.9075	0.7034	0.0083	0.396	0.0162	0.0128	0.047
	Quadratic Discriminant Analysis	0.9119	0.9065	0.7121	0.0605	0.357	0.1034	0.0811	0.1182
	K Neighbors Classifier	0.9105	0.8947	0.5229	0.0107	0.125	0.0198	0.0065	0.0125
	SVM - Linear Kernel	0.8913	0.8844	0	0.0275	0.05	0.0282	-8E-04	-9E-04
	Decision Tree Classifier	0.8469	0.8357	0.5379	0.1665	0.144	0.1546	0.0709	0.0711
ENN Balanced	Ada Boost Classifier	0.8993	0.8993	0.7422	0.0206	0.536	0.0396	0.0324	0.0898
	Gradient Boosting Classifier	0.8993	0.8993	0.746	0.0193	0.536	0.0372	0.0304	0.0869
	Light Gradient Boosting Machine	0.8991	0.8993	0.7458	0.0193	0.513	0.0371	0.0301	0.0843
	Logistic Regression	0.899	0.8824	0.535	0	0	0	0	0
	Ridge Classifier	0.899	0.8814	0	0	0	0	0	0
	Naive Bayes	0.8989	0.8835	0.588	0.0003	0.143	0.0007	0.0002	0.0026
	Random Forest Classifier	0.8988	0.8836	0.723	0.0229	0.48	0.0436	0.0347	0.0872
	Linear Discriminant Analysis	0.8985	0.8843	0.7332	0.0274	0.46	0.0517	0.0408	0.0926
	Extra Trees Classifier	0.8985	0.8817	0.7129	0.0176	0.44	0.0339	0.0263	0.0715
	Quadratic Discriminant Analysis	0.8931	0.8852	0.7182	0.1014	0.388	0.1607	0.1241	0.1568
	K Neighbors Classifier	0.8903	0.8814	0.5374	0.0212	0.164	0.0375	0.0147	0.0241
	Decision Tree Classifier	0.8252	0.8147	0.5489	0.2026	0.178	0.1897	0.0922	0.0925
	SVM - Linear Kernel	0.6762	0.6617	0	0.2884	0.076	0.0883	0.0053	0.0061
SMOTE + TOMEK	Light Gradient Boosting Machine	0.9537	0.9539	0.9762	0.9092	0.998	0.9515	0.9074	0.911
	Extra Trees Classifier	0.9529	0.9425	0.9828	0.9183	0.987	0.9512	0.9059	0.908
	Random Forest Classifier	0.95	0.9468	0.9764	0.9057	0.994	0.9477	0.9001	0.9037
	Gradient Boosting Classifier	0.9416	0.9357	0.9717	0.8911	0.991	0.9385	0.8832	0.8878
	Ada Boost Classifier	0.9081	0.8925	0.957	0.8632	0.948	0.9038	0.8162	0.8195
	Decision Tree Classifier	0.9029	0.8914	0.9029	0.9095	0.898	0.9035	0.8058	0.8059
	Quadratic Discriminant Analysis	0.7054	0.6941	0.7753	0.7707	0.682	0.7235	0.4108	0.4144
	Ridge Classifier	0.6851	0.6728	0	0.6868	0.684	0.6856	0.3702	0.3702
	Linear Discriminant Analysis	0.6851	0.6742	0.7483	0.6868	0.685	0.6857	0.3702	0.3702
	K Neighbors Classifier	0.669	0.6517	0.7281	0.7033	0.658	0.68	0.3381	0.3389
	Naive Bayes	0.5919	0.5817	0.6198	0.7562	0.569	0.6495	0.1838	0.1946
	Logistic Regression	0.5536	0.5434	0.5663	0.6678	0.544	0.5994	0.1072	0.1101
	SVM - Linear Kernel	0.5186	0.5072	0	0.4786	0.547	0.4409	0.0373	0.0507

Question 2: Which feature selection techniques work best on credit scoring datasets? What is the impact of feature selection techniques on credit scoring datasets?

The study used three feature selection techniques to generate three feature datasets each comprising of 30 features. The study generated three more combination or hybrid feature datasets by selecting features that are common to two of the above feature selection techniques.

Each of the datasets were tested using 13 base classifiers to ensure the performance does not include any classifier bias. The performance of the classifiers was evaluated on the basis of Accuracy, AUC, Precision, Recall, Kappa and F1 scores respectively.
The step forward feature selection wrapper method-based dataset was the best performing dataset across above mentioned classifiers

Question 3: What are the performance gains and falls of using base classifiers and ensembles on credit scoring datasets? What are the best performing models?
The study identified the best sampling technique and feature selection technique on the Home Credit dataset and generated the training and test datasets.
The dataset was subjected to 13 classifiers and the performance of the models was evaluated on the basis of Accuracy, Precision, Recall, AUC, F1 Score and Kappa. The top three classifiers were Light Gradient Boosting Machine, Extra Trees Classifier and Random Forest Classifiers. Hyper parameter Optimization was performed on the top three classifiers and LightGBM emerged as the best performing classifier post hyper parameter optimization.
The study created three heterogenous ensembles using bagging, blending and stacking classifiers and compared the performance of the ensembles to the LightGBM classifier. The ensemble scores were comparable to the top performing base classifiers of LightGBM and Extra Trees Classifier. The study concludes that application of feature engineering, feature selection and sampling techniques on the Home Credit dataset generated strong learners from the LightGBM and Extra Trees base classifiers. Ensembles are relevant and their efficacy stands out when they are run on weak learners which is not the scenario on the Home Credit dataset. Thus, the ensemble performance was comparable to base classifiers.

Question 4: How to explain the credit scoring models to various business stakeholders like loan officers?

The use of model interpretability techniques on the best performing base classifier allows for global and local explanations of the prediction value given by the model. Tree SHAP is used by this research to arrive at global explanations on feature significance and importance used by the model. The top features and their probability importance identified through global interpretability can be used to describe the rationale of the model used by Credit scoring business to regulatory and other stakeholders.
Due to Basel3 and other regulatory requirements, loan officers are required to explain the decision factors for a particular loan application. Local explanations provided by model

interpretability allow for explanations for that particular loan application. This study has used LIME as well as TREE SHAP to arrive at explanations for default loan application as well as non-default loan applications. The visualizations and feature explanations that can be derived through local explanations allow for explanations for how the model arrived at the prediction for that specific case.

5.6 Summary

The results of feature selection techniques used in this study are compared in section 5.2 and the step forward feature selection approach is determined to be the best feature selection approach. Using these features and the SMOTE+Tomek data balanced dataset, the performance of 13 classifiers on the dataset is compared using scoring metrics such as accuracy, precision, recall, F1 score, ROC AUC and Kappa, coefficient. On the top three performing classifiers, a fitting set of hyperparameters derived by the pycaret auto tuning approach is adjusted, with emphasis on improving the achievement of the developed models. For instance, the LightGBM model has shown a considerable improvement in its efficiency, while models such as Extra Trees Classifier and Random Forest Classifier failed to show any further improvement.

The study generated three heterogenous ensembles using Bagging, Blending and Stacking methods. Each of the heterogenous ensembles used the top performing base classifiers. An analysis of the performances of the ensembles compared to their category as well as to base classifiers results in the following conclusions:

- The performance metrics between bagging, blending and stacking ensembles are more or less comparable. There is no one ensemble method that outperforms its peer category
- The base classifier LightGBM out performs the ensemble approaches. On well prepared, feature engineered, feature selected and data balanced datasets, strong base classifiers prove themselves to be strong learners and such strong learners may outperform the ensembles.

Basis, the above performance comparison of base classifiers and ensembles, this study concludes that LightGBM is the top performing classifier for the Home Credit Dataset.

The LightGBM classifier is selected to model interpretability and the explanations generated with LIME and Tree Shap are evaluated with interpretation from different visualization plots. The global and local interpretation techniques available in SHAP are applied and the results explained. Local feature interpretations using LIME are detailed in the section.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This section presents a discussion on how this study achieved the goals and objectives defined in section 1. The conclusion and inferences to the research questions identified by this dissertation are explained. This chapter then goes on to summarize how this study contributes to the credit scoring or credit default risk prediction literatures. Finally, the limitations and constraints of the framework and approach of this study are discussed, with an objective to aid or influence future research direction.

6.2 Discussion and Conclusion

In the area of credit risk default, a wide range of exploratory research has been undertaken to make accurate financial decisions over time. The fundamental starting point of this dissertation is the problem of default on credit applications.

An extensive literature review identified how the credit risk prediction as a use case has been studied and evolved from 1970s to date and is still a relevant and very important use case to banks and financial institutions. The focal point of studies moved from generation of more sophisticated machine learning models as opposed to statistical methods to other areas like data balancing, feature engineering and feature selection techniques thereby rendering the dataset optimum and opportune to bring out the best performance from the models. There is no one model or approach that stands out as a best-in-class approach for credit scoring datasets.

The Home Credit dataset provided seven datasets comprising of various facets of the credit application and comprised of a large number of features. This provided a unique opportunity to test out the best-in-class techniques on data preparation and EDA on credit scoring datasets. Feature elimination, missing value treatment, anomaly treatment techniques were applied on the datasets. Given that the number of features for Home Credit dataset were already high, feature encoding approaches needed to be more bespoke. The combination of approaches used in this study help strike a balance between respecting relevant features and at the same time not resulting in scarce and irrelevant features.

To improve the efficacy of models and their ability to predict credit default, this study employed manual feature engineering to generate polynomial features on related data points as well as generating domain-based aggregation features on the other datasets and merging the same with the Home Credit dataset.

The conclusion and model explainability results conclude that the features generated through polynomial feature generation and domain-based aggregation approaches feature at the top echelons of the feature selection and the model feature significance evaluation criteria. This is testament to the approach selected for the study.

The dataset is heavily imbalanced on the dependent or TARGET variable and this study analysed comparison of multiple data balancing techniques on the Home Credit scoring dataset and also highlighted the performance gains of applying data balancing on credit scoring datasets. The study unequivocally concludes that data sampling techniques outperform class imbalanced datasets. Furthermore, under-sampling techniques result in under performance if the underlying dataset does not have too many noisy or ambiguous datapoints. The best class imbalance treatment is to apply a hybrid approach of under-sampling and over-sampling techniques. This study concludes that SMOTE+Tomek resulted in the best performance for Home Credit dataset.

The Home Credit dataset comprises of 150+ features after feature engineering and the dissertation compared two levels of feature selection processes on credit scoring datasets. The step forward feature selection wrapper method outperformed two other industry methods as well as datasets comprised of features derived from combination of multiple feature selection techniques. The study concludes that feature selection techniques are important to bring out the best performance in classifiers. However, merely reducing features based on intersection approach of multiple feature selection techniques does not result in the best result.

The study compared the performance of 13 classifiers on the best dataset and identified top three base classifiers. Hyper parameter optimization was performed on each of the top three base classifiers. Optimization of hyperparameters significantly positively impacted the LightGBM model but underperformed on the Extra Trees and Random Forest Classifiers. The conclusion is thus that hyper parameter optimization is an integral an important part of machine learning use cases but may not always result in performance improvement on the models.

The literature review identified ensembles as a trend in recent studies on credit scoring datasets. This study implemented three heterogenous ensembles. The conclusion from performance comparison across the base classifiers and ensembles is that LightGBM base classifier emerged as the best performing classifier. The conclusion drawn is that ensembles work best when working with weak learners. If the dataset is well prepared, feature engineered and features optimally selected, base learners like LightGBM in the Home Credit dataset use-case can provide comparable or outperform the ensembles.

The literature review identified model interpretability as a relevant aspect to the adoption of more sophisticated models. This study provides local and global explanation of features from the top performing classifier (LightGBM) and compares the visualisation and feature significance approaches given by two explainable AI approaches.

6.3 Contribution and Importance of the study

In this dissertation, a literature review of papers in the field of credit default risk or credit scoring was conducted.

This paper presents a comparison of data sampling techniques and provides insights into which one suits the credit datasets and why others do not do better. This addresses the data sampling technique comparison gap from the literature review.

Datasets in this domain comprise of many features is a common theme from the papers but there is a gap on the feature engineering and selection techniques that suit the credit scoring domain. In this study, we generate new features using polynomial features and domain features. This study goes on to use three feature selection techniques to select the best features for model tuning.

There is a trend on credit scoring studies to build ensembles. This study identifies the top three out of thirteen base classifiers and tunes those models. The study goes on to build heterogenous ensembles and compares the performance of the classifiers to ensembles.

The conclusion from this study is that optimally tuned models on data balanced and feature engineered datasets perform equally well as ensembles.

This techniques for data balancing, feature engineering and model/ensembles in this study serves as a foundation for any future work in this area.

Finally, the study applies local and global feature explanations on credit scoring datasets using Tree SHAP and LIME. This approach of identifying feature importance and how those features influence credit default risk will help credit risk business groups and management make informed decisions and allow for acceptance of more complex models in this domain.

6.4 Future Recommendations

While data sampling, feature engineering and model generation on credit scoring datasets is covered in this paper, the study has also listed its limitations in section 1.5 of this paper. Those limitations spawn the below recommendations or ideas for future work in this domain.

- Given current trends on un-supervised learning algorithms and the changing financial market and people's economic and demographic status, this use-case could benefit from a comparison of un-supervised learning models and the impacts and performances on this use-case.
- The Home Credit use case provides data about previous loan applications and also about credit card history and point in sale credit history. This presents a use-case to identify credit default cases against the credit loan applications by years. The opportunity is to compare if the models identified showcase any data drift or model drift when compared to data points from different years.
- There is an opportunity to apply detailed domain acumen into feature engineering and generate newer features which may influence the model better. Usage of automated feature generation tools and an analysis on the performance and impact that those newly generated synthetic features are a good use case for future work.

REFERENCES

Abellán, J. and Castellano, J.G., (2017) A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, [online] 73, pp.1–10. Available at: https://doi.org/10.1016/j.eswa.2016.12.020.

Agarwal, S., Chomsisengphet, S. and Liu, C., (2008) *Determinants of small business default*. [online] *The Analytics of Risk Model Validation*. Woodhead Publishing Limited. Available at: http://dx.doi.org/10.1016/B978-075068158-2.50004-4.

Al-qerem, A., (2019) Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection. [online] pp.235–240. Available at: https://doi.org/10.1109/ACIT47987.2019.8991084.

Ala'raj, M. and Abbod, M.F., (2016) A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications*, [online] 64, pp.36–55. Available at: http://dx.doi.org/10.1016/j.eswa.2016.07.017.

Anon (n.d.) *Javapoint Website*.

Anon (n.d.) Sigmoid Curve.

Bahnsen, A.C., Aouada, D. and Ottersten, B., (2014) Example-dependent cost-sensitive logistic regression for credit scoring. *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*, [online] pp.263–269. Available at: https://doi.org/10.1109/ICMLA.2014.48.

Bellotti, T. and Crook, J., (2009) Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, [online] 362 PART 2, pp.3302–3308. Available at: http://dx.doi.org/10.1016/j.eswa.2008.01.005.

Brown, I. and Mues, C., (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, [online] 393, pp.3446–3453. Available at: http://dx.doi.org/10.1016/j.eswa.2011.09.033.

Chawla, N. V, Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., (2002) SMOTE : Synthetic Minority Over-sampling Technique. [online] 16, pp.321–357. Available at: https://doi.org/10.1613/jair.953.

Chen, F.L. and Li, F.C., (2010) Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, [online] 377, pp.4902–4909. Available at: http://dx.doi.org/10.1016/j.eswa.2009.12.025.

Crone, S.F. and Finlay, S., (2012) Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, [online] 281, pp.224–238. Available at: http://dx.doi.org/10.1016/j.ijforecast.2011.07.006.

Dastile, X. and Celik, T., (2021) Making Deep Learning-Based Predictions for Credit Scoring Explainable. [online] 9. Available at: https://doi.org/10.1109/ACCESS.2021.3068854.

Dastile, X., Celik, T. and Potsane, M., (2020) Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, [online] 91, p.106263. Available at: https://doi.org/10.1016/j.asoc.2020.106263.

Demajo, L.M., Vella, V. and Dingli, A., (2020) Explainable AI for Interpretable Credit Scoring. *Expert Systems with Applications*, [online] pp.185–203. Available at: https://doi.org/10.5121/csit.2020.101516.

Fithria Siti Hanifah, Hari Wijayanto and Anang Kurnia, (2015) SMOTE bagging algorithm for imbalanced dataset in logistic regression analysis (case: Credit of bank X). *Applied Mathematical Sciences*, [online] 9137–140, pp.6857–6865. Available at: http://dx.doi.org/10.12988/ams.2015.58562.

Hand, D.J. and Henley, W.E., (1997) Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, [online] 1603, pp.523–541. Available at: https://doi.org/10.1111/j.1467-985X.1997.00078.x.

He, H., Zhang, W. and Zhang, S., (2018) A novel ensemble method for credit scoring : Adaption of different imbalance ratios. *Expert Systems With Applications*, [online] 98, pp.105–117. Available at: https://doi.org/10.1016/j.eswa.2018.01.012.

Hsu, C.F. and Hung, H.F., (2009) Classification methods of Credit Rating – A Comparative Analysis on SVM , MDA and RST. [online] pp.3–6. Available at: https://doi.org/10.1109/CISE.2009.5366068.

Huang, C., Chen, M. and Wang, C., (2007) Credit scoring with a data mining approach based on support vector machines. [online] 33, pp.847–856. Available at: https://doi.org/10.1016/j.eswa.2006.07.007.

LC, T., (2000) A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, [online] 16, p.149. Available at: https://doi.org/10.1016/S0169-2070(00)00034-0.

Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C., (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, [online] 2471, pp.124–136. Available at: http://dx.doi.org/10.1016/j.ejor.2015.05.030.

Louzada, F., Ara, A. and Fernandes, G.B., (2016) Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, [online] 212, pp.117–134. Available at: http://dx.doi.org/10.1016/j.sorms.2016.10.001.

Lundberg, S.M. and Lee, S.-I., (2019) A Unified Approach to Interpreting Model Predictions. *Expert Systems with Applications*, [online] 322, pp.1208–1217. Available at: https://doi.org/10.1016/j.inffus.2019.12.012%0Ahttps://doi.org/10.1016/j.ophtha.2018.11.016.

Maldonado, S., Pérez, J. and Bravo, C., (2017) Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research*, [online] 2612, pp.656–665. Available at: https://doi.org/10.1016/j.ejor.2017.02.037.

Marqués, A.I., García, V. and Sánchez, J.S., (2012a) Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, [online] 3911, pp.10244–10250. Available at: https://doi.org/10.1016/j.eswa.2012.02.092.

Marqués, A.I., García, V. and Sánchez, J.S., (2012b) Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, [online] 3912, pp.10916–10922. Available at: https://doi.org/10.1016/j.eswa.2012.03.033.

Mester, L.J., (1997) What's the Point of Credit Scoring ? *Business Review*, [online] 3January, pp.3–16. Available at: https://www.phil.frb.org/research-and-data/publications/business-review/1997/september-october/brso97lm.pdfhttps://doi.org/10.1016/j.eswa.2012.03.033.

Moscato, V., Picariello, A. and Sperlí, G., (2021) A benchmark of machine learning approaches for credit score prediction. *Expert Systems With Applications*, [online] 165May 2020, p.113986. Available at: https://doi.org/10.1016/j.eswa.2020.113986.

Qi, J., Yang, R. and Wang, P., (2021) Application of explainable machine learning based on Catboost in

credit scoring Application of explainable machine learning based on Catboost in credit scoring. [online] Available at: https://doi.org/10.1088/1742-6596/1955/1/012039.

Qiu, Z., Li, Y., Ni, P. and Li, G., (2019) Credit risk scoring analysis based on machine learning models. *Proceedings - 2019 6th International Conference on Information Science and Control Engineering, ICISCE 2019*, [online] pp.220–224. Available at: https://doi.org/10.1109/ICISCE48695.2019.00052.

Soares De Melo Junior, L., Nardini, F.M., Renso, C. and Fernandes De MacEdo, J.A., (2019) An empirical comparison of classification algorithms for imbalanced credit scoring datasets. *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, [online] pp.747–754. Available at: https://doi.org/10.1109/ICMLA.2019.00133.

Taha, A.A. and Malebary, S.J., (2020) An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access*, [online] 8, pp.25579–25587. Available at: https://doi.org/10.1109/ACCESS.2020.2971354.

Tounsi, Y., Anoun, H. and Hassouni, L., (2020) CSMAS: Improving Multi-Agent Credit Scoring System by Integrating Big Data and the new generation of Gradient Boosting Algorithms. *ACM International Conference Proceeding Series*. [online] Available at: https://doi.org/10.1145/3386723.3387851.

Wang, C., Han, D., Liu, Q. and Luo, S., (2019) A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. *IEEE Access*, [online] 7, pp.2161–2168. Available at: https://doi.org/10.1109/ACCESS.2018.2887138.

Wang, G., Hao, J., Ma, J. and Jiang, H., (2011) A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, [online] 381, pp.223–230. Available at: http://dx.doi.org/10.1016/j.eswa.2010.06.048.

Wang, G. and Ma, J., (2012) A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. *Expert Systems with Applications*, [online] 395, pp.5325–5331. Available at: http://dx.doi.org/10.1016/j.eswa.2011.11.003.

Xia, Y., Liu, C., Li, Y.Y. and Liu, N., (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, [online] 78, pp.225–241. Available at: http://dx.doi.org/10.1016/j.eswa.2017.02.017.

Zhang, D., Zhou, X., Leung, S.C.H. and Zheng, J., (2010) Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, [online] 3712, pp.7838–7843. Available at: http://dx.doi.org/10.1016/j.eswa.2010.04.054.

APPENDIX A - RESEARCH PROPOSAL:

1. Background

The need for Credit score analysis can be traced to the beginning of commerce probably around the same time as money lending and borrowing started. There have been numerous studies on credit scoring models and statistical and machine learning models exist which can measure the credit worthiness of the consumer. Many such studies exist but are limited to the particular dataset or the features therein. Since the Basel Committee on Banking Supervision released the Basel Accords, specially the second accord from 2004, the use of credit scoring has grown considerably, not only for credit granting decisions but also for risk management purposes. Basel III, released in 2013, render more accurate calculations of default risk, especially in the consideration of external rating agencies, which should have periodic, rigorous and formal comments that are independent of the business lines under review and that re-evaluates its methodologies and models and any significant changes made to them

Home Credit is a non-banking financial institution, founded in 1997 in the Czech Republic. The company operates in 14 countries (including United States, Russia, Kazakhstan, Belarus, China, India) and has over 29 million customers, total assets of 21 billion Euro, over 160 million loans, with the majority in Asia and almost half of them in China (as of 19-05-2018). Home credit group focusses on the demographic that has little to no credit history with the objective of leveraging credit scoring models that reduce the cost of error of approving a bad loan but at the same time allow for Home Credit group to expand their consumer base by approving loans to credit worthy consumers as correctly identified by the model. Towards this end, they have gathered a number of datapoints from their consumer base and the credit patterns (past and present) which will help in generation of credit scoring models.

The research will be to develop a hybrid ensemble model which will calculate the credit risk score for the Home Credit dataset and will compare the results of the new models with other classifiers that have been run on the dataset and attempt to identify best in class solution for the same. The study will also apply model interpretability techniques on the best-in-class model so that loan managers and regulators can interpret the decisions taken by the model.

2. Related Work

Credit scoring models are statistical decision models that evaluate the repayment ability of loan applicants. Credit scores help to depict the credit and repayment history, utilization of credit, tenures of previous debts. Credit scoring will never be able to predict with certainty the performance of an individual loan, but it does provide a method of quantifying the relative risks of different groups of borrowers. Scoring has the potential to be one of the factors that change small-business banking.(Mester, 1997)

The literature review by (Louzada et al., 2016) analyses and classifies 187 research literature on clustering and classification techniques applied to credit scoring published in scientific journals between 1992- 2015. The study asserts that credit scoring is a present and important financial area, and datasets used for modelling are mostly private due to the confidentiality of customer data. A trend that is observed among the papers in recent years is to propose a new model for credit score ratings, and the predictive performance of the hybrid techniques is almost similar. Moreover, comparison with statistical techniques is increasingly not visible in recent time periods. This highlights a trend that researchers are more interested in identifying generalist methods which have high predictive performance.

The literature by (Lessmann et al., 2015) updates Baesens et al. (2003) and explores the relative effectiveness of alternative classification algorithms in credit scoring. It compares 41 classifiers and concludes that heterogenous ensemble classifiers work better than the industry standard of Logistic Regression. The study highlights lack of acceptance of advanced scoring models due to the expertise required for the same and the insight that the decision taken by the model is not interpretable to people. This highlights an opportunity to implement model interpretable techniques

The authors propose Shapley (Lundberg and Lee, 2019) Additive exXplanations (SHAP), an unified framework for prediction interpretation. Shap comprises of values that represent features importance measures for prediction and are the result of 6 local feature attribution methods pooling their insights together. There are many positive indications that showcase that human intuition results are comparable to SHAP.

The study by (Marqués et al., 2012a)(Marqués et al., 2012b) has focussed on studying the behaviour of classifier ensembles. With this aim, seven classification methods and five ensemble approaches have been applied to six credit scoring problems. The study can be further extended using the Home Credit dataset and to compare the ensembles studied in the present

work with other methods that combine different classifiers (for example, stacking) (Abellán and Castellano, 2017) and (Zhang et al., 2010).

The paper by (Xia et al., 2017) involves performing data pre-processing and EDA followed by the implementation of feature selection techniques to identify redundant variables. Removal of the same showcases improvements in performance and computational costs. The XGBoost credit scoring model is selected as the base one and is further optimized using TPE hyper parameter tuning. The base model is applied over five datasets using five measures for evaluation. The best parameter tuning method for the study is showcased to be TPE. There is opportunity to further the study by improving EDA and managing class imbalance in the dataset.

The literature by (Brown and Mues, 2012) uses a private dataset but which has class imbalance and they have successfully used SMOTE to manage the class imbalance. Also, they have compared multiple classifiers and determined that XGBoost is the best performing amongst the same. This study provides insight into managing class imbalance and the opportunity to use stacking/ensemble approaches in credit scoring models to further the study.

The study by (Tounsi et al., 2020) compares gradient boosting methods (XGBoost, CatBoost and LightGBM) on the Home Credit dataset. However, there is opportunity for future research for better feature engineering and data pre-processing

The study by (Qiu et al., 2019) uses the same dataset as chosen by our study but uses only 4 out of 7 datasets and performs model training using Logistic Regression, Light GBM and Random Forest approaches. The study compared these classifiers and determine that Light GBM achieved the best AUC score of 78%.

3. Research Questions

A few gaps have been identified in the area of credit risk default prediction around using the optimal machine learning techniques together with handling class imbalance, feature selection and classification models. This forms the base for the following research questions

Question 1: How to achieve better model accuracy with limited set of credit default cases compared to non-default cases?

Question 2: How to improve the model's prediction capability by selecting the right input features?

Question 3: How does different ensemble methods compare to classification methods and/or industry standard Logistic Regression?

Question 4: How to explain the credit scoring models to various business stakeholders like loan officers?

4. Aims and Objective.

Overall, summary of literature review determines that credit scoring model evolution is still taking place and the current trend is to develop new models using ensemble approaches. The Home Credit dataset consists of data imbalance which when managed is ripe for the application of feature selection/extraction techniques. The aim of this objective is to use these features and train models using state of the art classification/ensemble approaches and showcase the results and compare those with previous studies. This will further the research carried thus far on the dataset and at the same time provide more insights on the Home Credit default dataset credit scoring best in class model by using model interpretability techniques. This will be useful to loan management and regulatory personnel in their ability to understand the decisions taken by the model and will also provide insights to future researchers in this area.

Following are the research objectives that have been put together on the basis of the aim of this study:

- To suggest a suitable class imbalance handling technique which can be applied on the imbalanced Home Credit Risk Default dataset.
- To improve the performance of classification/ensemble models for credit risk scoring using feature importance and selection techniques.
- To generate ensemble models and compare their performance vis-à-vis known better performing classification models to highlight the performance of the models when applied to the Home Credit Default Risk data set.
- To apply explainable artificial intelligence methodologies to the best performing model and to be able to explain the black box model better to business stakeholders.

5. Significance Of Study.

Credit scoring default risk model research is established as a clear and present requirement for banking and financial institutions in the world. The need for the same is to ensure that the cost of error is minimized. Another important factor is also the competitive banking space of these times where banks and financial institutions want to grab every opportunity to improve customer base and ensure all demographic that are capable of loan repayment have loans available, Though there have been many statistical models developed in the past and more machine learning models and now hybrid models, there still exists opportunity to improve on the model creation and comparison of its performance to classifiers, ensemble and statistical methods. Literature review over the years showcase that current trend in model identification is using hybrid /ensemble models though the comparison of performance of hybrid vs classifiers needs more study and results. This study will use the Home Credit Default Risk dataset which comprises of 7 datasets having class imbalance and on which there have been some studies where base classifiers have been compared. The significance of the study will be the ability to further the existing research papers by applying class imbalance techniques, apply feature selection algorithms and train classifiers and ensembles and compare the results of the same. This result will be useful to future research in the domain of credit scoring models as it will add cadence on performance of these classifier and ensemble models for further analysis. Another significant feature of the study will be to apply model interpretability techniques on the best classifier/ensemble model. This is an area of study hitherto not applied to the Home Credit dataset and the results will be useful not only to the machine learning community but also the business groups (loan management and regulators) who can use the study to apply and adopt more complex black box models and use model interpretability to make sense of the results. This will bridge the gap as identified in the study carried out by (Lessmann et al., 2015).

6. Scope Of Study.

6.1 Scope of study

The study will include comprehensive exploratory data analysis and feature selection analysis for different algorithms developed for credit risk prediction. SMOTE technique will be used for handling class imbalance. The study will make use of filter and wrapper based as well as embedded methods for feature selection.

Literature review on the credit scoring research papers indicate a current trend of identifying new model but highlights a dearth of comparison of its performance to other classifier models. This study will further the current research performed on Home Credit Default dataset and create and train new classification models using Extra Trees Classifier(et), Light Gradient Boosting (LightGBM), MLP (mlp), Ridge(ridge), Gradient Boosting(gbc), ADA Boost(ada), Naive Bayes(nb), Decision Tree(dt), Quadritic Discriminant Analysis(qda), Random Forest(rf), Logistic Regression(lr), K Neighbours(knn), SVM-Linear Kernel(svm), Linear Discriminant Analysis(lda). The study will also develop ensemble approaches using Bagging & Boosting. A comparison of the performance of the models will identify the best performing model and apply model interpretability techniques (SHAP, GIRP) on the same thereby making the results of the model more interpretable to the business (loan management & regulators) and the machine learning research community. In summary, the novelty of the study is the implementation of class imbalance technique, feature selection techniques, model creation and evaluation using classifier/ensemble approaches and application of model interpretability on the Home Credit Loan Default dataset.

6.2 Out of Scope of the study

The study will use the Home Credit Loan Default dataset provided by Home Credit and apply feature selection and class imbalance techniques on the same. Also, the model behaviour and performance comparison are limited to data from just this dataset. An opportunity exists to apply the same techniques on other globally available credit scoring datasets (Australian Credit scoring dataset, Taiwanese dataset, Austrian dataset) but the same is out of scope of this study. The processing will be done in python. Any equivalent implementations of the code in any other language are out of scope.

6.3 Limitation of study

While the study here includes a combination of state-of-the-art techniques on class imbalance, feature selection, comparing performance of classifier & ensembles and model interpretability, it is limited in its scope in the following manner.

The study only uses one technique for class imbalance management (SMOTE) and does not compare the behaviour of different other techniques like SMOTEEN, under sampling using IHT and random under sample.

The study is limited in its usage of model interpretability to either choose global or local variables and does not provide a comprehensible comparison of all explainable AI techniques like SHAP, GIRP, Anchor, LIME etc.

Only one dataset is used for the feature selection & model training and evaluation techniques used in this study. Identifying multiple datasets may provide insights for model drift which is not covered in this study.

7. Research Methodology.

7.1 Introduction

The primary approach of this study is to use a combination of class imbalance management techniques, feature selection/extraction techniques on the Home Credit Loan Default dataset and to train various classifier & ensemble models and to compare performance of the same to identify state of the art model. The study will also apply explainable AI techniques on the start

of the art model. The step-by-step approach to achieve the objective is depicted below.

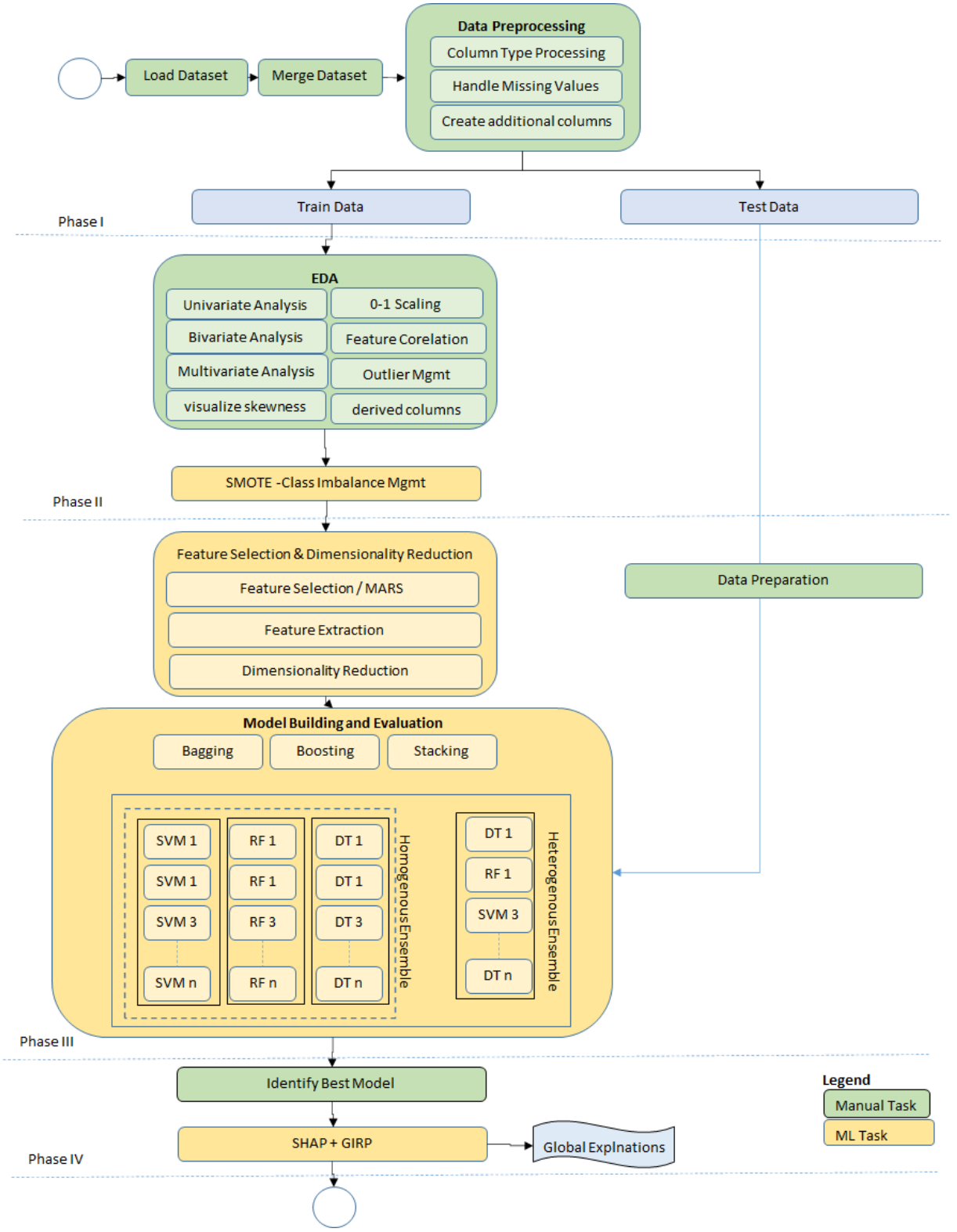


Figure 13: Flow chart of research methodology

7.2 Dataset Description

Home credit group focusses on the demographic that has little to no credit history with the objective of leveraging credit scoring models that reduce the cost of error of approving a bad

loan but at the same time allow for Home Credit group to expand their consumer base by approving loans to credit worthy consumers as correctly identified by the model. Towards this approach, Home Credit group has made available their dataset to the Kaggle community so that the machine learning community can build credit scoring models that can provide unique insights and with the hope of identifying a state-of-the-art model which addresses the unique combination of features in this dataset and achieves high accuracy.

The dataset provided by Home Credit comprises of seven sources of data.

Application train: Information about each credit application at Home Credit. This becomes the main training and testing dataset. Each row in the dataset represents one loan application.

Bureau: This dataset contains information on previous loans that the customer has taken from other different financial institutions.

Bureau balance: This dataset contains monthly balances of credits in Credit Bureau. Each row represents one month of every previous credit.

Previous application: historical applications for credits at Home Credit of customers who have credits in the application data.

Pos cash balance: monthly data about previous point of sale or cash credits customers have had with Home Credit.

Credit card balance: monthly data about previous credit cards customers have had with Home Credit.

Instalment's payment: historical payment lineage for previous credits at Home Credit.

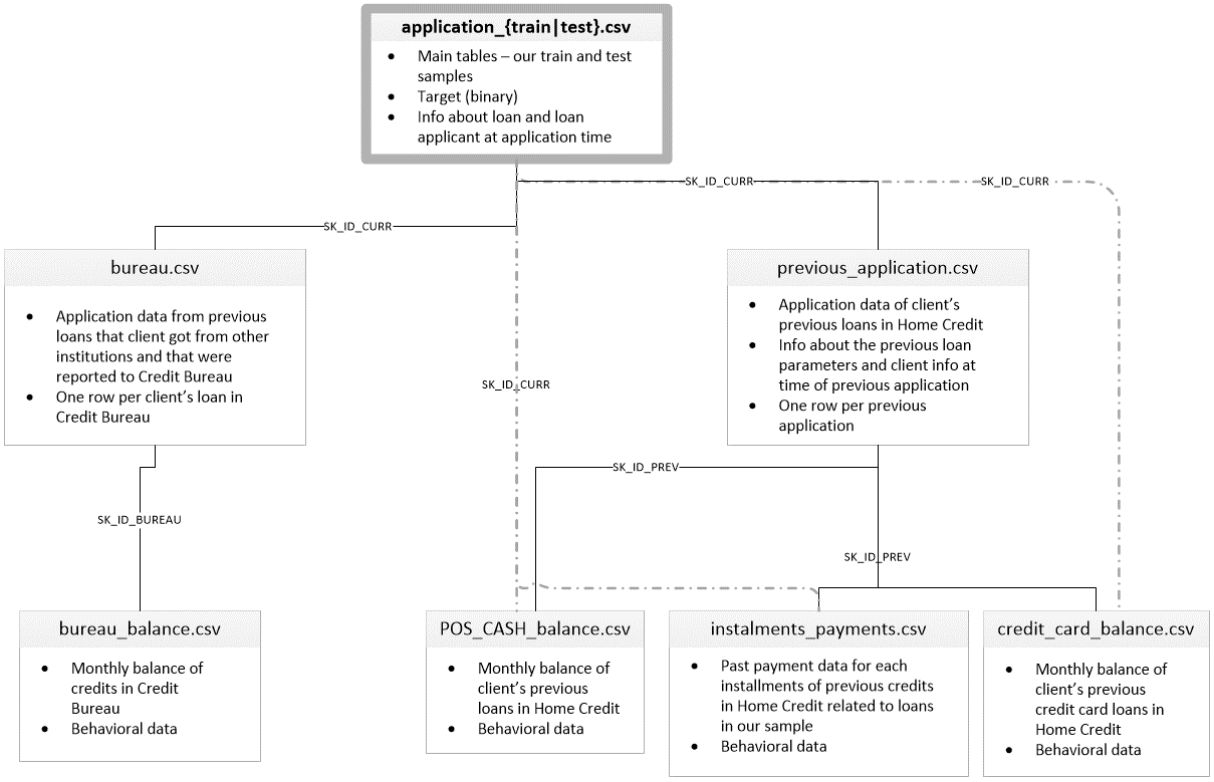


Figure 14: The relational data model - Home Credit Default Risk

7.3 Data Pre-processing and Transformations

The current application dataset will be loaded and merged with credit bureau as well as data from previous loan applications. New derived columns will be added to state those values. The dataset will be introspected to check the column sizes, types, mean, average values etc to get some understanding of the dataset before further pre-processing. Post the dataset inspection, the pre-processing steps will be performed equally on the train and the test files. The following steps will be performed separately for both train and test datasets for data pre-processing.

- Missing and Duplicate data handling
- Univariate Analysis to handle outliers
- Bivariate analysis using different types of plots
- Multi-collinearity inspection between features
- Feature scaling

The pre-processing will be done on train and test datasets separately to avoid data leakage issues. As a part of the Exploratory Data Analysis (EDA) on the big dataset, various methods like univariate and multivariate analysis, scaling will be performed to analyse the financial ratios as autonomous features. The EDA helps establish approximate corelation between the features with the target variable. At this stage the study will be innately able to describe the importance and effect of different features on credit default.

7.4 Class Imbalance Handling

There are various types of sampling methods which can be used to handle this class imbalance. Under-sampling methods reduce the instances of the majority cases while over-sampling methods either duplicate or introduce new instances of the minority cases to balance the data. There are also hybrid techniques that are actually amalgamation of over and under sampling methods. Certain algorithm-based sampling techniques are also available to balance the imbalanced class dataset. SMOTE is a Synthetic Minority Oversampling Technique which is used to handle the class imbalance in the dataset. Hence the study will use the SMOTE to handle class imbalance.

7.5 Feature Selection

Feature subset selection algorithms can be classified into two categories: the filter approach and the wrapper approach. The filter approach first selects important features subsets. It separates

features from classifier that are independent of any learning algorithm. The filter approach relies on various measures of the general characteristics of the training data such as distance, information, dependency and consistency. The wrapper model usually uses the predictive accuracy of a pre-determined learning algorithm to determine the accomplishment of the selected sub- sets. These methods of learning algorithm are computationally expensive for data with a large number of features. Generally, filters are faster and can be used as a pre-processing step to reduce space dimensionality and over fitting. On the other hand, wrapper approach may perform better in finding useful subsets of relevant variables. (Chen and Li, 2010)

The study will be targeted towards making use of filter based, wrapper based and embedded methods for feature selection to achieve better accuracy with the models.

7.6 Model Development and comparison

The literature review shows that over the years multiple statistical and machine learning models have been created to predict credit score. The current trend in research is to identify hybrid models but highlights a lack of comparison of predictive accuracy between the new models and classifiers. In this study, Home Credit Loan Default dataset and apply class imbalance techniques and feature selection techniques on the dataset which has not yet been applied on the same dataset in research papers. Using this dataset, the study will create a combination of classifiers using *Extra Trees Classifier(et), Light Gradient Boosting (LightGBM), MLP (mlp), Ridge classifier (ridge), Gradient Boosting(gbc), ADA Boost(ada), Naive Bayes(nb), Decision Tree(dt), Quadritic Discriminant Analysis(qda), Random Forest(rf), Logistic Regression(lr), K Neighbours(knn), SVM-Linear Kernel(svm) and Linear Discriminant Analysis(lda). The study* will also create ensemble models using either bagging or boosting and compare the performance of the models to arrive at the best-in-class model on which model interpretability models will be applied. The novelty of the study is in the combination of class imbalance techniques, feature selection techniques, classifier vs ensemble performance comparison and application of model interpretability on the state of art model, a combination which is novel for the Home Credit Loanda Default dataset.

7.7 Evaluation Metrics

Since the study is about leveraging supervised classification techniques for predicting credit default, the Confusion matrix will play a crucial part in the model evaluation. All the models which are a part of this study, will be evaluated based on below metrics

Accuracy: Accuracy indicates the ratio of the correctly predicted cases (positive and negative) to the total number of cases from the dataset.

Precision: Precision indicates the ratio of the correctly projected positive cases to overall predicted positive instances. This indicates the rate of observation that is precisely determined by the system against all the positives determined by the system.

Recall: Recall is also known as Sensitivity and it indicates the ratio of the projected positive cases to the actual overall positive instances in the dataset.

F1: F1 score is the weighted average of the precision and recall. It is considered as the measure of test's accuracy.

Area under the curve (AUC): AUC is used along with ROC curve as a measurement for performance of classification techniques using various cut-off points for the prediction probability. The model with higher AUC is better.

7.8 Model Interpretability

Credit scoring comprises of machine learning models that use regression, classifiers, clustering or a combination of hybrid or ensemble models that help define the probability of credit worthiness of a customer to the bank or financial institution running the model.

While many classifications and ensemble machine learning models have been applied to the area, the models' decisions are not interpretable by human beings.

Model interpretability comprises of explainable AI techniques which provide global or local explanations and use various techniques for the same. A description of how the classification model works overall from a global understanding is provided by global form of explanations. Business teams comprising of loan managers and regulators prefer this approach over individual explanations. These explanations aid the business teams in their primary responsibility of ensuring the model is fair, correct and provides compliant predictions. This study will employ explainable AI techniques for global explanations using GIRP and SHAP methodologies.

A model employs rules by priority for its predictions. GIRP can extract these rules and interpret them. Recent studies showcase that this methodology has achieved success in being agnostic of models in interpretation.

SHAP (SHapley Additive exPlanations) (slundberg/shap GitHub, n.d.) on the other hand, uses the classic Shapley values from game theory and their related extensions.

8. Expected Outcome.

The outcome of the study compares the performance of classifiers vs ensemble models that have been built on the Home Credit Loan Default dataset; that has been subjected to class imbalance treatment and feature selection techniques. This comparison of models will identify a state-of-the-art model on which model interpretability using explainable AI has been applied.

- The study will showcase the efficacy of feature selection techniques and class imbalance techniques for model building and creation. This will help future research conducted on this dataset.
- The performance comparison of classifiers vs ensemble models is in line with current trend of research in credit scoring and will form the basis for future research and trends.
- The use of model interpretability using explainable AI will help the business (loan management and regulators) adopt more black box models and be able to understand the decision behind the same. This will address the gap identified by (Lessmann et al., 2015) towards the lack of adoption of black box models

Overall, the novelty of the study is in the combination of class imbalance techniques, feature selection techniques, classifier vs ensemble performance comparison and application of model interpretability on the state of art model, a combination which is novel for the Home Credit Loan Default dataset. This should inspire adoption of new models by the business community and provide as inspiration for further research.

9. Required Resources.

9.1 Hardware requirement

The implementation will be done on a laptop with below prerequisites:

Processor:	Intel(R) Core (TM) i5-1035G1 CPU @ 1.00GHz 1.19 GHz
Memory:	16 GB
System Type:	64-bit operating system, x64-based processor
GPU:	Intel(R) UHD Graphics.
Operating System:	Windows - 10 and 64-bit OS

9.2 Software requirement

The prediction model will be implemented in Python 3. Various Python libraries like pandas, NumPy, matplotlib, seaborn will be used for data analysis. Sklearn will also be utilised for data pre-processing, model development & feature engineering and selection.

The study will make use of the open-source python-based machine learning library “pycaret”. It will also use 14 models from pycaret viz Extra Trees Classifier(et), Light Gradient Boosting (LightGBM), MLP (mlp), Ridge(ridge), Gradient Boosting(gbc), ADA Boost(ada), Naive Bayes(nb), Decision Tree(dt), Quadritic Discriminant Analysis(qda), Random Forest(rf), Logistic Regression(lr), K Neighbours(knn), SVM-Linear Kernel(svm) and Linear Discriminant Analysis(lda).

Sklearn metric libraries (ROC, AUC, Confusion matrix, Precision, Recall, etc.) will be used for model evaluation. Any other library in python will be used as and when required.

The python code will be maintained and executed using Jupyter Notebook IDE which is packaged as a part of Anaconda installation.

Risk and Mitigation Plan.

The below table summarizes the risks involved and contingency plans.

Table1: Risk and Mitigation plan

Sr No.	Risk	Contingency	Severity
1	Literature availability during research process	Literature review has been done prior to the submission of proposal. The same will continue during the thesis preparation. A good amount of literature has been covered during proposal incase there is any challenge during next stages.	Medium
2	Data quality not good for the research	The dataset has been used in multiple research activities so far. The study will also be using extensive EDA and class imbalance as well as feature selection techniques before using it for modelling.	Low
3	Hardware / software crash	Maintain the document and code backups on wekely basis.	Low
4	Unforeseen professional or personal commitments	The research plan has a buffer factored in to cover for any unplanned loss of time	Low
5	Research and thesis quality below expectation of thesis supervisor and the univercity	Regular interactions on research progress and regular review of incremental versions of research thesis has been planed with the thesis supervisor	Low

Research Plan

The research plan is prepared considering different phases mentioned in the research methodology (*Figure 1: Flow chart of research methodology*) and the milestones in the research thesis process. This plan spans across 21 weeks from 1st July till 6th December.