CREDIT RISK ANALYSIS

A Thesis Submitted in partial

fulfilment ofthe requirements or the

degree of Bachelor of Technology

in

Data Science By

Aarya Chirag Mehta

(SAP ID: 70041019046)

under the guidance of

Dr. Rajni Aron

Department of Data Science

SVKM's

NMIMS

University

Mumbai

November 2022

CERTIFICATE

This is to certify that the thesis entitled **"Credit Risk Analysis"** is a Bonafede work of **"Aarya Chirag Mehta (SAP ID: 70041019046)"** submitted to the SVKM's NMIMS University in partial fulfilment of the requirement for the award of the degree of **"Bachelor of Technology"** in **"Data Science"**.

Dr Siba Panda

HoD Data Science
NMIMS University

ACKNOWLEDGEMENT

It is with a feeling of great pleasure that I would like to express my most sincere heartfelt gratitude to **Dr. Rajni Aron,** for their steady and able guidance throughout my thesis. I am greatly indebted to the university for providing me with an opportunity to showcase my learnings through this thesis.

I express my sincere thanks to **Prof. Sarada Samantaray, Prof. Jaykrishna Joshi, Dr. Siba Panda** for their guidance and for providing the facilities in the necessary department. I am also thankful to all the staff members of the department of Data Science, my family, and friends for inspiring me and being helpful.

Table of Contents

ABSTRACT

With the aid of a credit risk model, this project, Credit Risk Analysis, aims to advise the organization on how to enhance its operations. The issue statement for the proposed project is that lending institutions must assess and manage credit risk in a new and unfamiliar environment with limited visibility and access to trustworthy data as lockdowns are now being removed and businesses are returning to normal.
Approaches to collections and loss mitigation will also evolve. What's unique about this situation is that many impacted borrowers never thought they'd have trouble paying their bills. Corporate leverage has reached previously unheard-of heights in certain nations, particularly the US, in recent years. Bank workout departments, meanwhile, are only operating at a small portion of their potential.

The project assists us in determining who is eligible for a loan utilizing the estimated acceptance rates and a thorough decile analysis to ensure the accuracy of our conclusions. By examining the borrower's track record, payment history, etc., this takes the borrower's financial situation into account. We select the best fit among several machine learning models and train it. Even while it's impossible to predict who exactly may miss payments, correctly evaluating and managing credit risk can decrease the impact of a loss. Lenders and investors are compensated for taking on credit risk by receiving interest payments from borrowers or debt obligation issuers.

INTRODUCTION

Credit risk is the chance of suffering a loss as a result of a borrower's failure to make loan payments or fulfil contractual commitments. It alludes to the possibility that a lender won't get the principal and interest that is owed, which would cause a disruption in cash flows and raise collection expenses. Excess cash flows could be written to offer more protection against credit risk. A higher coupon rate, which generates more cash flows, can be used to reduce credit risk when it is present for a lender. There is a chance that the borrower won't be able to return the debt when lenders give mortgages, credit cards, or other sorts of loans. Similar to that, there is a chance that a consumer won't pay their invoices if a business extends credit to them. Credit risk also refers to the possibility that a bond issuer won't pay up when required or that an insurance provider won't be able to cover a claim. The borrower's general capacity to repay a loan in accordance with its original terms is used to determine credit risks. Lenders use the five Cs—credit history, ability to repay, capital, the terms of the loan, and any associated collateral—when determining the credit risk of a consumer loan.

What Factors are Used to Assess Credit Risk?

In order to assess the credit risk associated with any financial proposal, the project finance division of the firm first assesses a variety of risks relating to the borrower and the relevant industry.

The borrower credit risk is evaluated by considering:

- The financial position of the borrower, by analysing the quality of its financial statements, its past financial performance, its financial flexibility in terms of the ability to raise capital, and its capital adequacy
- The borrower’s relative market position and operating efficiency
- The quality of management, by analysing its track record, payment record, and financial conservatism

Industry-specific credit risk is evaluated by considering:

- Certain industry characteristics, such as the importance of the industry to the economic growth of the economy and government policies relating to the industry
- The competitiveness of the industry
- Certain industry financials, including return on capital employed, operating margins, and earnings stability

Challenges to Successful Credit Risk Management

- **Inefficient data management.** An inability to access the right data when it's needed causes problematic delays.
- **No groupwide risk modelling framework.** Without it, banks can't generate complex, meaningful risk measures and get a big picture of groupwide risk.
- **Constant rework.** Analysts can't change model parameters easily, which results in too much duplication of effort and negatively affects a bank's efficiency ratio.
- **Insufficient risk tools.** Without a robust risk solution, banks can't identify portfolio concentrations or re-grade portfolios often enough to effectively manage risk.
- **Cumbersome reporting.** Manual, spreadsheet-based reporting processes overburden analysts and IT.

Best Practices in Credit Risk Management

The first step in effective credit risk management is to gain a complete understanding of a bank's overall credit risk by viewing risk at the individual, customer and portfolio levels.

While banks strive for an integrated understanding of their risk profiles, much information is often scattered among business units. Without a thorough risk assessment, banks have no way of knowing if capital reserves accurately reflect risks or if loan loss reserves adequately cover potential short-term credit losses. Vulnerable banks are targets for close scrutiny by regulators and investors, as well as debilitating losses.

The key to reducing loan losses – and ensuring that capital reserves appropriately reflect the risk profile – is to implement an integrated, quantitative credit risk solution. This solution should get banks up and running quickly with simple portfolio measures. It should also accommodate a path to more sophisticated credit risk management measures as needs evolve. The solution should include:

- Better model management that spans the entire modelling life cycle.
- Real-time scoring and limits monitoring.
- Robust stress-testing capabilities.
- Data visualization capabilities and business intelligence tools that get important information into the hands of those who need it, when they need it.

LITERATURE REVIEW

The ongoing difficulty of rising credit risks and nonperforming loans to the global financial system has made this study necessary. The effectiveness of credit facilities provided to borrowers by commercial banks is of major interest to many stakeholders, including regulators. Better bank performance and acceptable levels of credit risk are indicators of a healthy banking industry and, eventually, a robust economy. Strong performance of commercial banks exposes the overall economy to significant economic and infrastructure advancements. Additionally, new employment opportunities are produced, and all of these factors increase policymakers' and regulators' interest in the financial system's performance. The study of the literature offers a thorough examination of earlier research that touch on the main factors that explain the connection between credit risk management and commercial banks' performance. It also makes an effort to analyze how macroeconomic factors and NPLs relate to one another. There are several ways to gauge a bank's performance, however this study suggests using the CAMELs financial rating model, which has also received recognition from many other studies. The study analyzes the various hypotheses that have been put up to account for the main factors that explain the proposed link. A thorough empirical literature evaluation of earlier research, paying particular attention to the important study factors, has been conducted. This aided in identifying the topics for future research and the gaps in current knowledge. The introduction of two significant factors that affect the relationship is really intriguing. These are macroeconomic factors as the moderating variables and non-performing loans as the intervening variable. These two significant variables have improved the research and explored the potential of the proposed link. The majority of the studies depended on secondary or gathered data, and interpretation was done utilizing logical and analytical thinking to identify patterns or trends.

The challenge of nonperforming loans (NPLs) in bank systems in many countries cannot be overemphasized. It is also a succinct that banks or financial institutions that need to manage and maintain acceptable levels of NPLs must invest in a robust and reliable credit risk management system. The implementation of robust and effective credit risk management has become a critical aspect that determines the performance of commercial banks on a global scale. Provision of credit facilities is one of the biggest sources of revenue for any commercial bank in any corner of the globe. Nevertheless, the likelihood of borrowers being unable to meet their loans obligations or commitments has lately been on increase and this is a major concern for banks especially those involved in unsecured lending. Sujeewa (2015) indicated that this is because the risks associated with borrower’s default could have huge impact on other related business. There is a clear interrelationship between the two main

variables i.e. credit risk management and the bank performance. It is generally believed that the robustness of the systems set to manage credit risks has a bearing on the levels of non-performing loans a bank would record and ultimately influences the level of profitability and by extension the composite bank performance.

Credit risk is a popular type of risk that both non-financial and financial institutions must deal with. Credit risk occurs when a debtor or borrower fails to fulfill his obligations to pay back the loans to the lender. In banking business it happens when payments can either be delayed or not made at all which can cause cash flow problems and affect a banks liquidity. (Greening&Bratanovic 2009, 161). Credit risk is by far the most significant risk faced by banks and the success of their business depends on accurate measurement and efficient management of this risk to a greater extent than any other risks. In our country the financial sector is still in the developing and many banks have not been able to establish a firm risk management framework, particularly credit risk management, in order to prevent unfavorable events. This is dangerous when banks customer services are still in their infancy and banks revenue depends heavily on lending activities and credit growth is central to any banking organizations profit. In addition, the control work from the central bank, though playing a growing role, has not been protective enough. Access to credit information and history is very limited. Small bank is file for bankruptcy due to bad credit assessment practices brought a big loss to the bank.

"Smoke cannot be released without a fire" There must have been something wrong that banks credit procedures. The main source of credit risk include limited institutional capacity, inappropriate credit policies, volatile interest rates, poor management, low capital and liquidity levels, direct lending, poor loan underwriting, laxity in credit assessment, poor lending practices, government interference and inadequate supervision by the central bank. For banks, managing credit risk is not a simple task since comprehensive considerations and practices are needed for identifying, measuring, controlling and minimizing credit risk

RESEARCH METHODOLOGY

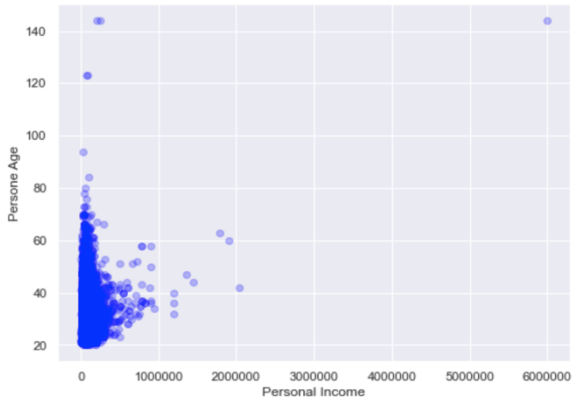
In order to determine the variables we need to validate the 10 various types of hypotheses in this project, I first ran exploratory data analysis (EDA) on the data set in order to gain a broad concept of what the data set contains and to remove some of the outliers. Then, many algorithms including Logistic Regression, RandomForestClassifier, XGBclassifier, and CatBoostClassifier were utilized in machine learning. The acceptance rate and predicted loss for each of the models utilized were calculated with the aid of the best fit model.

EDA:

	attributes	min	max	range	mean	median	std	skewness	kurtosis
0	person_age	20.000000	144.000000	124.000000	27.734600	26.000000	6.347981	2.581393	18.560825
1	person_income	4,000.000000	6,000,000.000000	5,996,000.000000	66,074.848470	55,000.000000	61,982.167945	32.865349	2,693.272776
2	person_emp_length	0.000000	123.000000	123.000000	4.789686	NaN	4.142565	2.614455	43.722338
3	loan_amnt	500.000000	35,000.000000	34,500.000000	9,589.371106	8,000.000000	6,321.989624	1.192477	1.423565
4	loan_int_rate	5.420000	23.220000	17.800000	11.011695	NaN	3.240404	0.208550	-0.671609
5	loan_status	0.000000	1.000000	1.000000	0.218164	0.000000	0.412999	1.364888	-0.137088
6	loan_percent_income	0.000000	0.830000	0.830000	0.170203	0.150000	0.106780	1.064669	1.223687
7	cb_person_cred_hist_length	2.000000	30.000000	28.000000	5.804211	4.000000	4.054939	1.661790	3.716194

As observed, we can notice the presence of outliers in `person_age` (`max = 144`) and `person_emp_length` (`max = 123`).

Age and income show a positive association when shown in a scatterplot, which may indicate that older recipients are further along in their careers and hence make more money. Additionally, it appears that the data contains outliers.

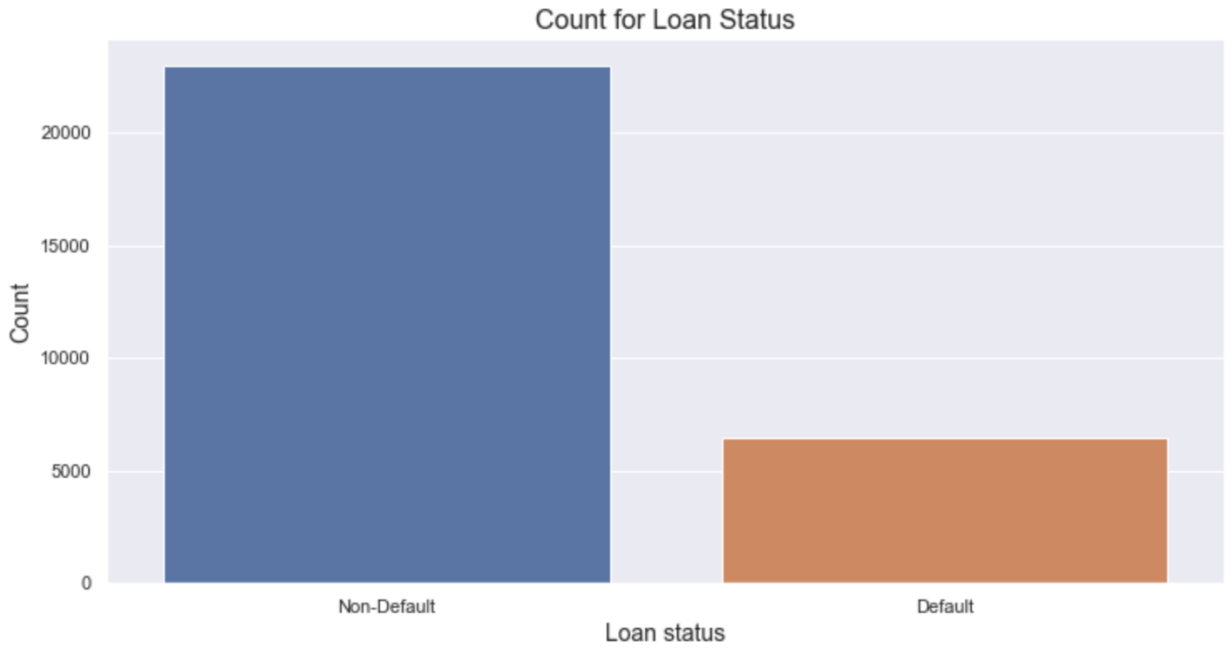


The next step is to locate and eliminate any outliers that may exist in the data. Cross tables and aggregate functions are available. The person emp length column can be seen. We may make use of methods like min and max to find outliers. Using this, we discovered two such entries in our dataset, which we removed.

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_inc
0	22	59000	RENT	123.000000	PERSONAL	D	35000	16.020000	1	0.59
210	21	192000	MORTGAGE	123.000000	VENTURE	A	20000	6.540000	0	0.10

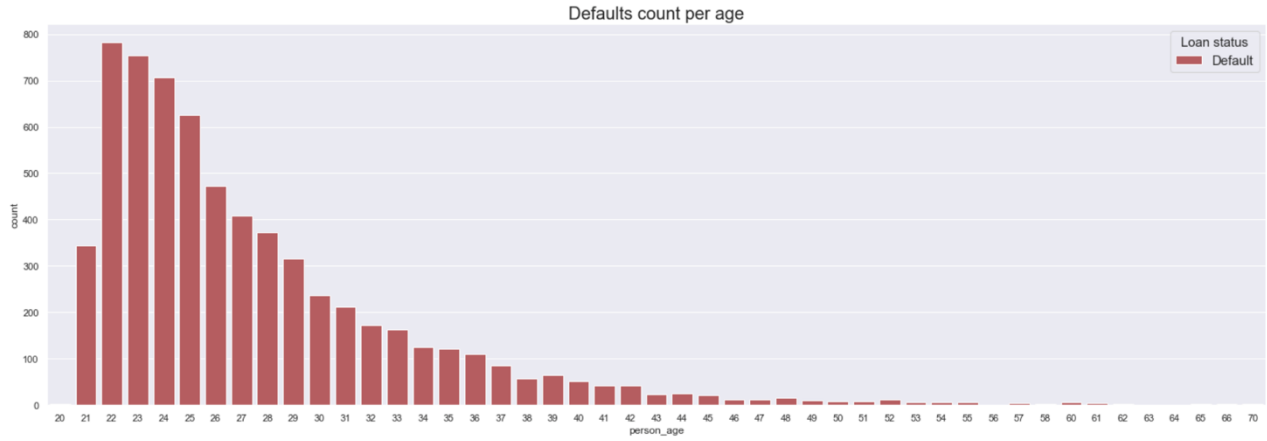
The hypothesis mind map is then made:
This mind map provides a general overview of credit risk and the numerous factors that contribute to it. We can use this mind map to generate our own hypotheses and then use UNIVARIATE and BIVARIATE analysis to determine whether they are true or untrue.

When we use Univariate Analysis to determine the number of loan default and non-default instances, the results show that there are significantly more non-default cases than default ones. As a result, the data set we're working with is unbalanced. As we continue with our analysis, we will take care of this.



Furthermore, we will be using Bivariate analysis to validate the hypotheses we came up.

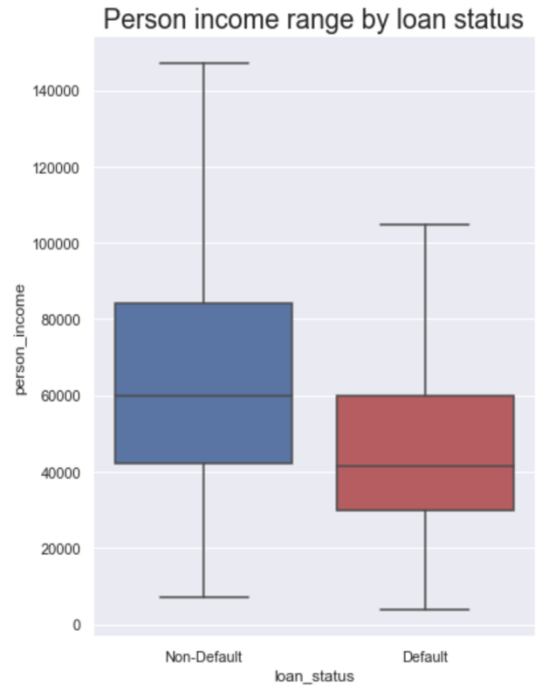
H1



As observed, there are more deault cases for people at young age (up to 40s).

> Thus, the hypothesis is **TRUE**.

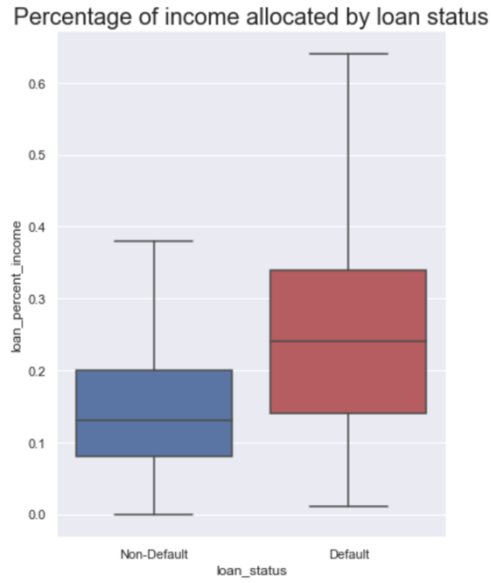
H2



As observed, the mean income for people who default is lower than for people who don't.

> Thus, the hypothesis is **TRUE**.

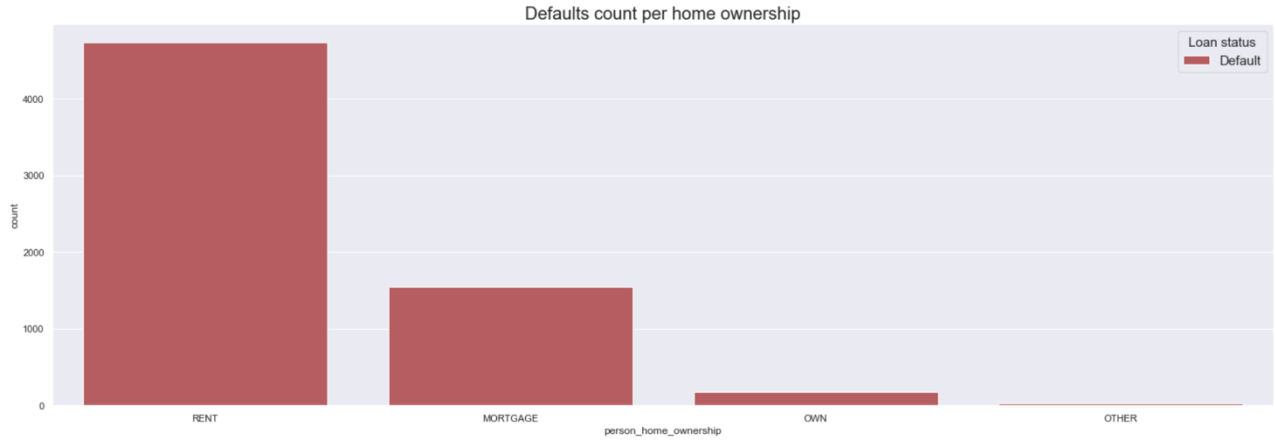
H3



As observed, the median of the percentage of income allocated is higher for people who default than for the people who don't.

> Thus, the hypothesis is **TRUE**.

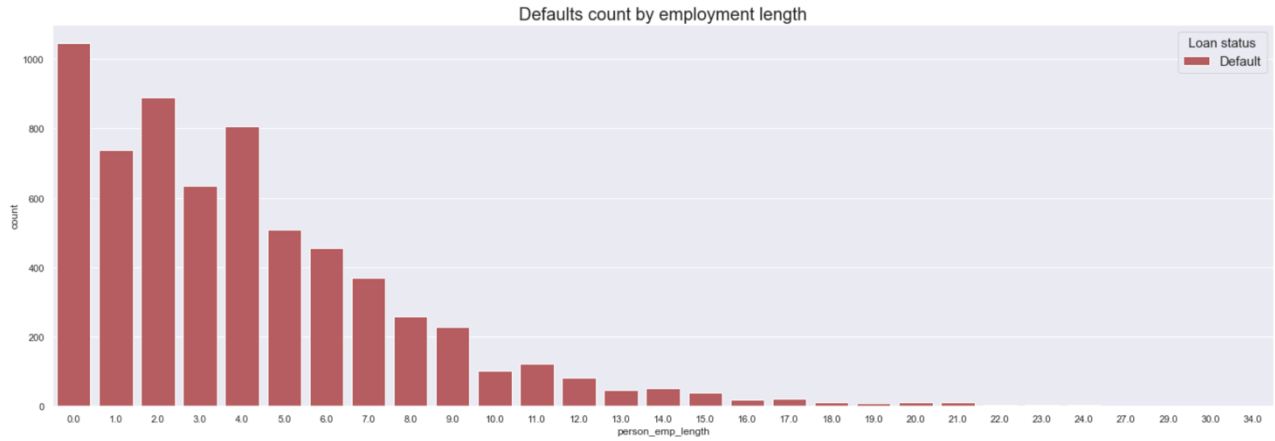
H4



As observed, `RENT` is the top home ownership type for people who default, not `MORTGAGE`.

> Thus, the hypothesis is **FALSE**.

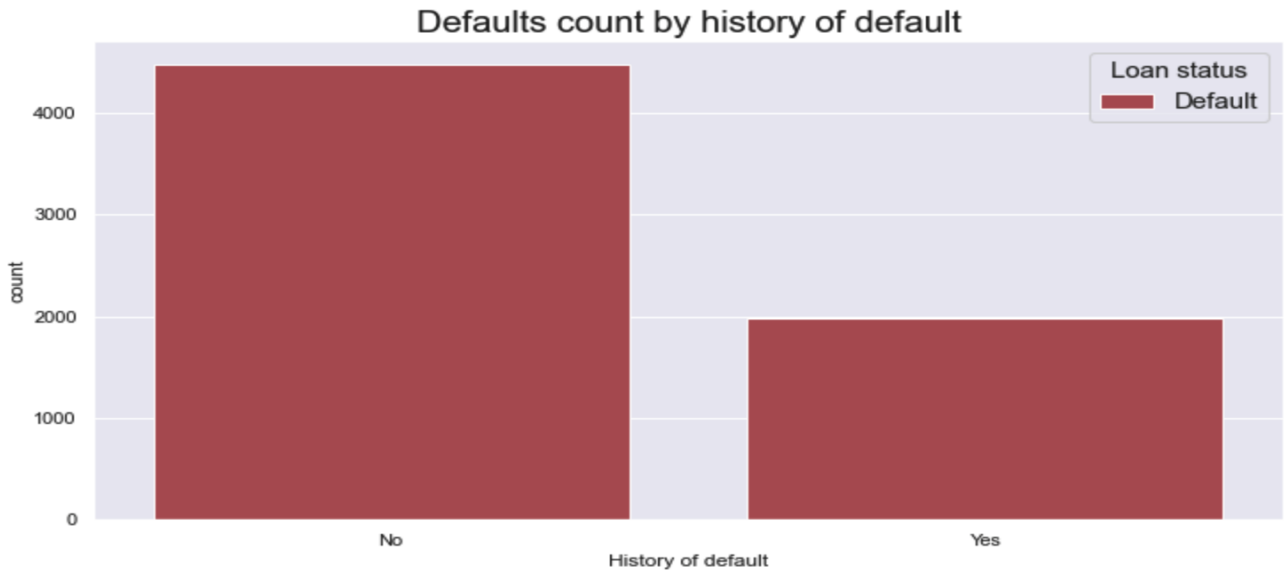
H5



As observed, the are fewer cases of default for people with long employment length that for people in the early years.

> Thus, the hypothesis is **TRUE**.

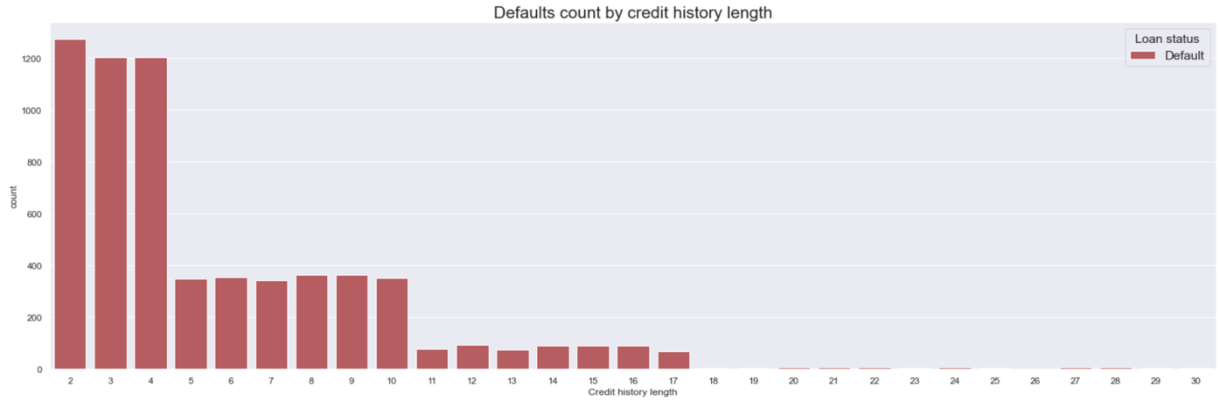
H6



As observed, there are more cases of default for people who don't have history of default.

> Thus, the hypothesis is **FALSE**.

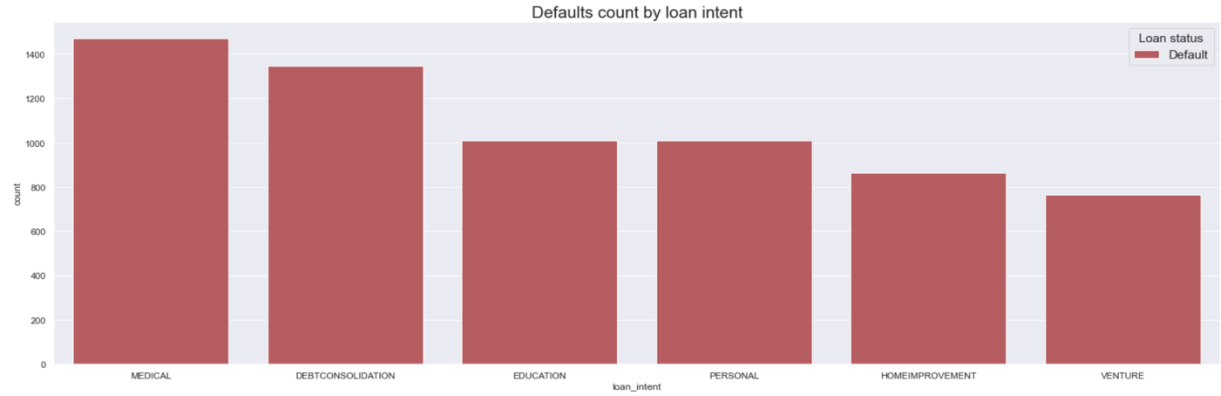
H7



As observed, there are fewer cases of default for people having longer credit history length.

> Thus, the hypothesis is **TRUE**.

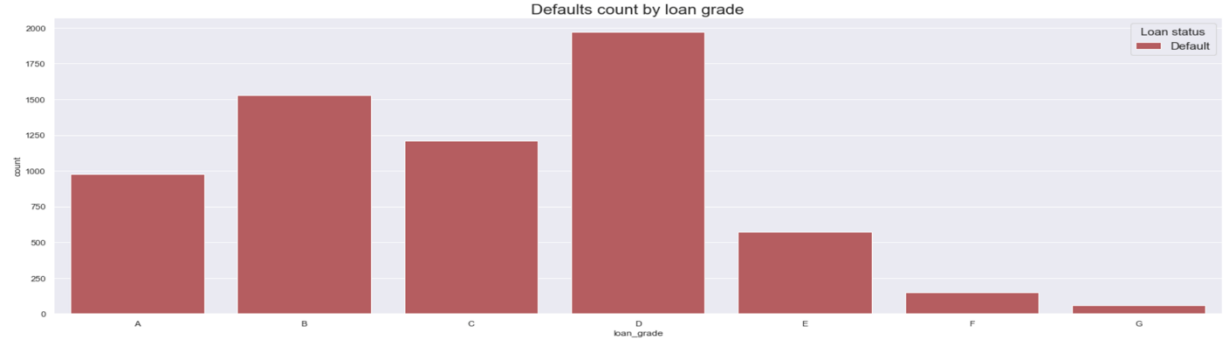
H9



As observed, `MEDICAL` holds the most number of default cases and `VENTURE` the least.

> Thus, the hypothesis is **FALSE**.

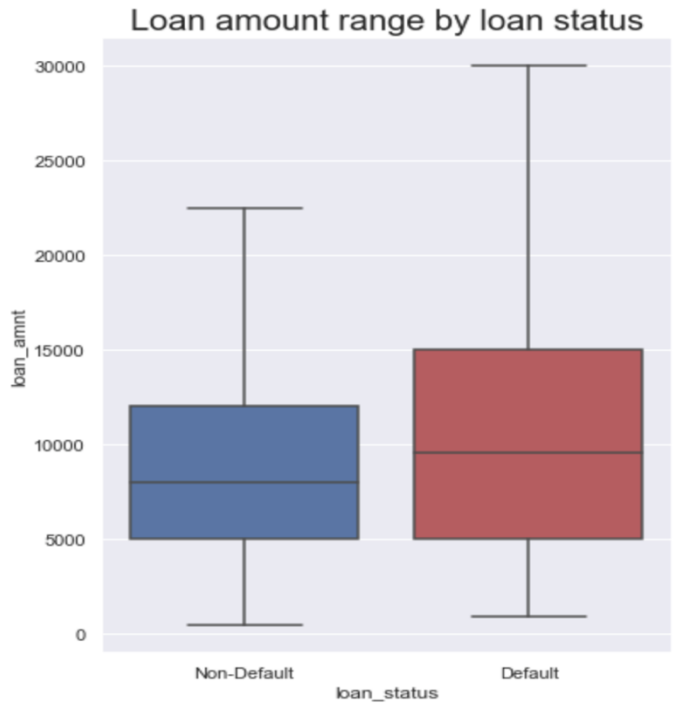
H9



As observed, despite the grade `D` that is right at the middle grade, the higher the loand grade, the higher is the number of default cases.

> Thus, the hypothesis is **FALSE**.

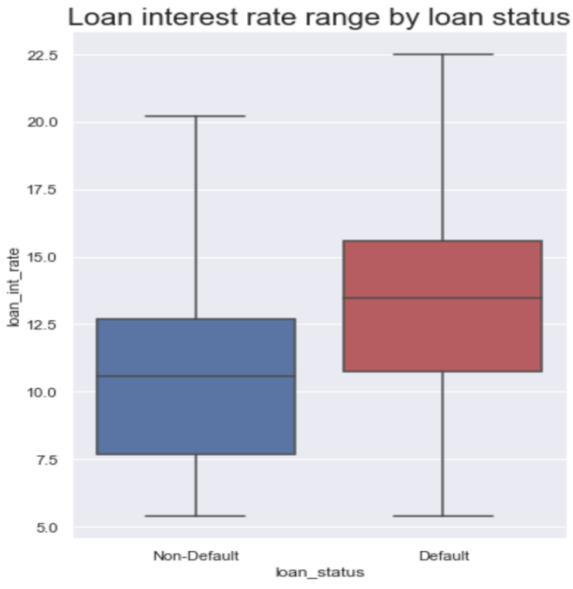
H10



As observed, the loan amount median for default cases is higher than for non-default.

> Thus, the hypothesis is **TRUE**.

H11



As observed, the interest rate median for default cases is higher than for non-default.

> Thus, the hypothesis is **TRUE**.

ID	Hypothesis	Conclusion
H1	There are more defaults for young people	True
H2	People who default have lower income than people who not default	True
H3	The median percentage of income to the loan is higher for default than the median for non-default	True
H4	Mortgage have more cases of default, followed by rent and own	False
H5	The are fewer cases of default for people with long employment length	True
H6	There are more cases of default for people having history of default	False
H7	There are fewer cases of default for people having longer credit history length	True
H8	There are more cases of default for personal than any other intent	False
H9	The least cases of default are for venture	True
H10	The higher the grade, the fewer are the cases of default	False
H11	The loan amount median for default cases is higher than for non-default	True
H12	The interest rate median for default cases is higher than for non-default	True

DATA SET & PRE-PROCESSING

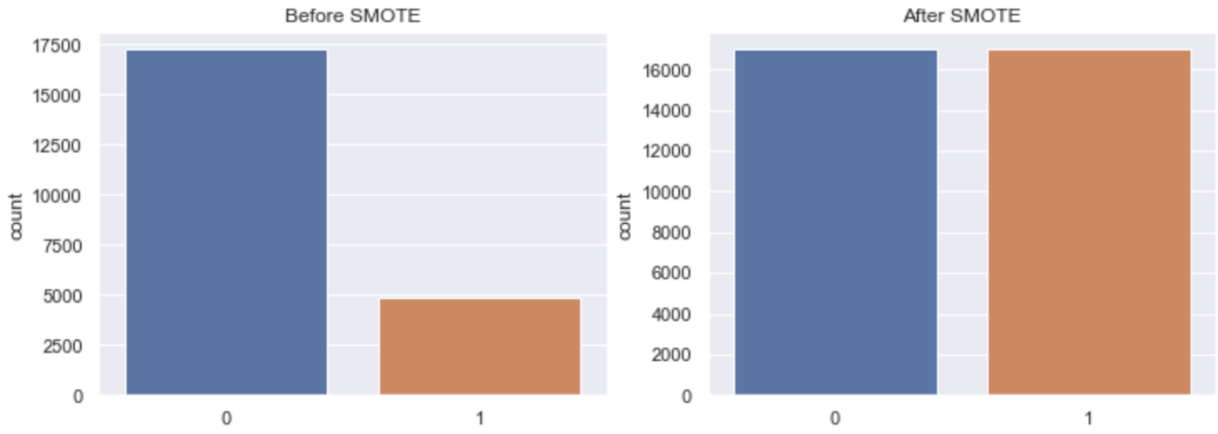
We start by examining the columns and data contained within each of the data set's rows.

Column	Description
person_age	The person's age in years
person_income	The person's anual income
person_home_ownership	The type of home ownership (RENT; OWN; MORTGAGE; OTHER)
person_emp_length	The person's employment length in years
loan_intent	The person's intent for the loan (PERSONAL, EDUCATION, MEDICAL, VENTURE, HOMEIMPROVEMENT, DEBTCONSOLIDATION)
loan_grade	The loan grade (A, B, C, D, E, F, G)
loan_amnt	The loan amount
loan_int_rate	The loan interest rate
loan_status	Shows whether the loan is currently in default with 1 being default and 0 being non-default
loan_percent_income	The percentage of person's income dedicated for the mortgage
cb_person_default_on_file	If the person has a default history (Yes; No)
cb_person_cred_hist_length	The person's credit history

The data must be scaled because the range of variables within them varies widely, allowing us to treat each feature equally when the model consumes them. After utilizing one-hot encoding to construct the new columns, we can concatenate them with the numeric columns to produce a new data frame that will be used for likelihood of default prediction for the rest of the course. Just one-hot encrypt the non-numeric columns, remember. This would result in a data set that was very large for the numeric columns. After all of this, we connect all of the data frames onto which our machine learning models will now be applied.

MACHINE LEARNING MODELS

After dividing the dataset into train and test sets, we use SMOTE to balance the sets and count the number of classes before and after oversampling. This is done because our univariate analysis revealed that the data is unbalanced.



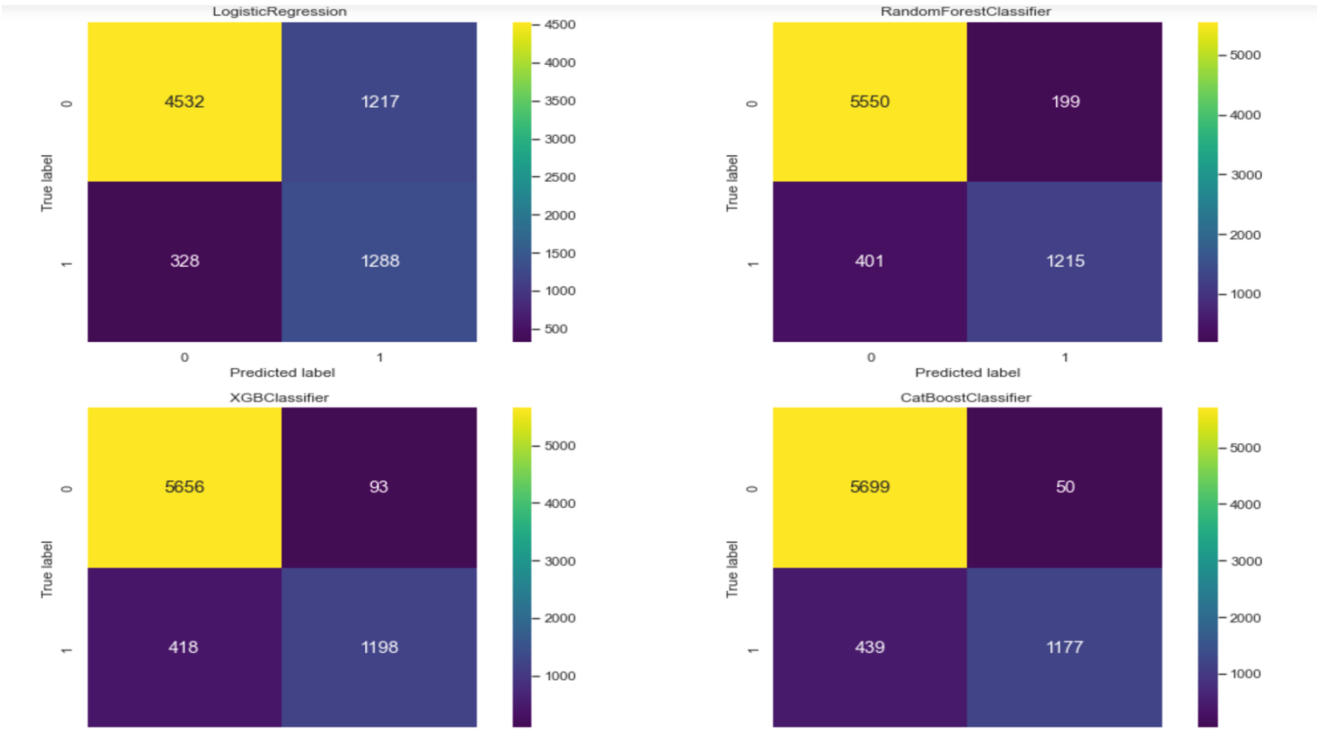
Next, we select four fundamental machine learning models—LogisticRegression, RandomForestClassifier, XGBclassifier, and CatBoostclassifier—set a common default threshold, and then use the test dataset to calculate each classifier's performance in order to determine which is the best fit for our analysis.

	model	precision	recall	f1-Score	ROC AUC	accuracy	cohen kappa
0	LogisticRegression	0.514172	0.797030	0.625091	0.872677	0.790224	0.488703
2	RandomForestClassifier	0.859264	0.751856	0.801980	0.930397	0.918534	0.750985
4	XGBClassifier	0.927963	0.741337	0.824217	0.945817	0.930618	0.781668
6	CatBoostClassifier	0.959250	0.728342	0.827999	0.942967	0.933605	0.787811

The maximum value for each column is shown by the highlighted cells. As you may have noticed, the goal of our project is recall. The highest recall score is provided via logistic regression. However, it has the lowest precision rating. Let's assume that our primary goals are memory and moderate to high precision. In this case, we can benefit from the f1-score, which is the harmonic average of recall and precision. CatBoostClassifier is the algorithm that meets this need. We computed a confusion matrix for all the models, along with a Probability Distribution graph comparing all the models, to further demonstrate that CatBoostClassifier is the best fit model for our research. As a last test, I utilized the Brier Score Loss test to confirm my choice. The maximum value for each

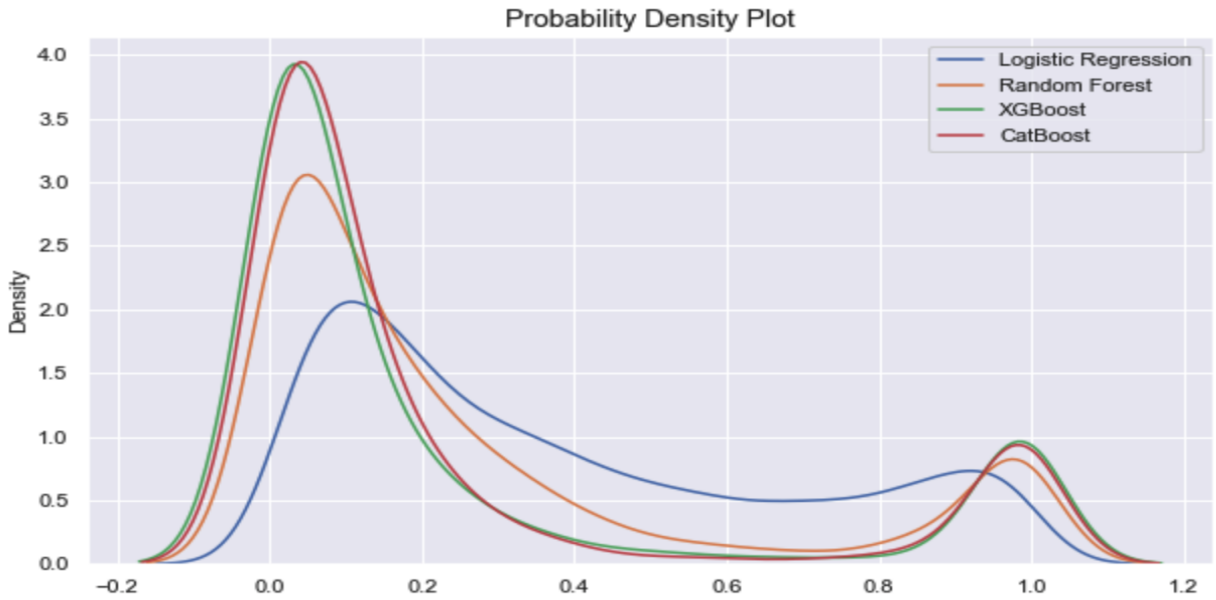
column is shown by the highlighted cells. As you may have noticed, the goal of our project is recall. The highest recall score is provided via logistic regression. However, it has the lowest precision rating. Let's assume that our primary goals are memory and moderate to high precision. In this case, we can benefit from the f1-score, which is the harmonic average of recall and precision. CatBoostClassifier is the algorithm that meets this need.
We computed a confusion matrix for all the models, along with a Probability Distribution graph comparing all the models, to further demonstrate that CatBoostClassifier is the best fit model for our research. As a last test, I utilized the Brier Score Loss test to confirm my choice.

CONFUSION MATRIX:



As previously stated, the CatBoostClassifier meets the requirements. However, why not choose Logistic Regression since it has the fewest FN? Yes, however compared to other techniques, the quantity of FP (1217) is far too high.

PROBABILITY DISTRIBUTTION:



As observed, in general all the algorithms present most probabilities concentrated around 0. However, XGBoostClassifier presents the largest number of probabilities concentrated
around 0 while LogisticRegression presenting the least with a moderate distributed probabilities around other value. RandomForest seems to be the mid-term between the other two, but its F1-Score precision score aren't satisfactory enough to consider it, as we compared with other algorithms. The mid-term between XGBoost and RandomForest is CatBoost which suits best for the context and challenge of our project.

BRIER SCORE LOSS:

The smaller the Brier score, the better, hence the naming with “loss”. Across all items in a set N predictions, the Brier score measures the mean squared difference between (1) the predicted probability assigned to the possible outcomes for item ‘I’, and (2) the actual outcome. Therefore, the lower the Brier score is for a set of predictions, the better the predictions are calibrated. Note that the Brier score always takes on a value between zero and one, since this is the largest possible difference between a predicted probability (which must be between zero and one) and the actual outcome (which can take on values of only 0 and 1). The Brier loss is composed of refinement loss and calibration loss."

```
# calculates the Brier Score Loss
bsl_lr = brier_score_loss(y_test, y_pred_lr_prob, pos_label=1)
bsl_rf = brier_score_loss(y_test, y_pred_rf_prob, pos_label=1)
bsl_xgb = brier_score_loss(y_test, y_pred_xgb_prob, pos_label=1)
bsl_catb = brier_score_loss(y_test, y_pred_catb_prob, pos_label=1)

# prints the calculated Brier Score Loss for each algorithm probability
print(f'Brier Score Loss (Logistic Regression): {bsl_lr}')
print(f'Brier Score Loss (Random Forest): {bsl_rf}')
print(f'Brier Score Loss (XGBoost): {bsl_xgb}')
print(f'Brier Score Loss (CatBoost): {bsl_catb}')
```

```
Brier Score Loss (Logistic Regression): 0.13724290910531756
Brier Score Loss (Random Forest): 0.06664486082824168
Brier Score Loss (XGBoost): 0.054907762949161124
Brier Score Loss (CatBoost): 0.05480004049550755
```

As observed, although the probability plots are quite similar, the Brier Score Loss are different. For this score, the closer to 0, the better. Thus, CatBoost suits best for our needs.

CatBoostClassifier:

Training and applying models for the classification problems. Provides compatibility with the scikit-learn tools.

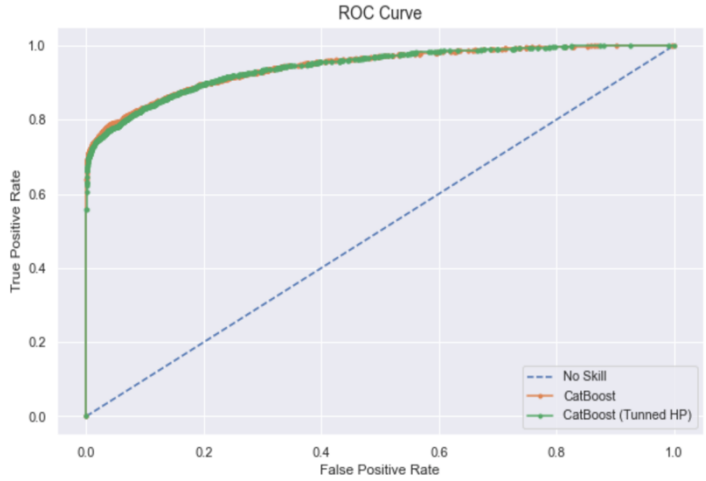
The default optimized objective depends on various conditions:

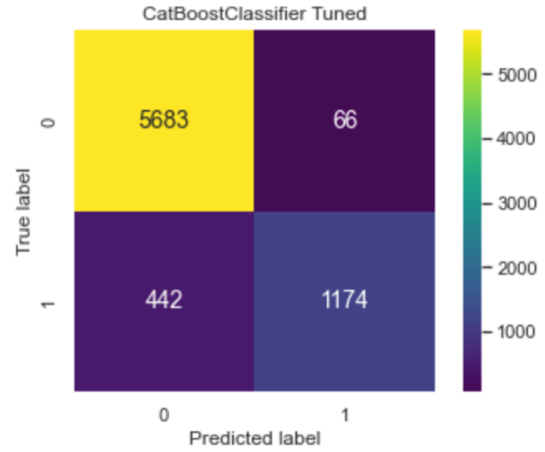
- Log loss — The target has only two different values or the target_border parameter is not None.
- MultiClass — The target has more than two different values and the border_count parameter is None.

A one-dimensional array of categorical columns indices (specified as integers) or names (specified as strings).This array can contain both indices and names for different elements. If any features in the cat_features parameter are specified as names instead of indices, feature names must be provided for the training dataset. Therefore, the type of the X parameter in the future calls of the fit function must be either catboost.pool with defined feature names data or pandas.DataFrame with defined column names.

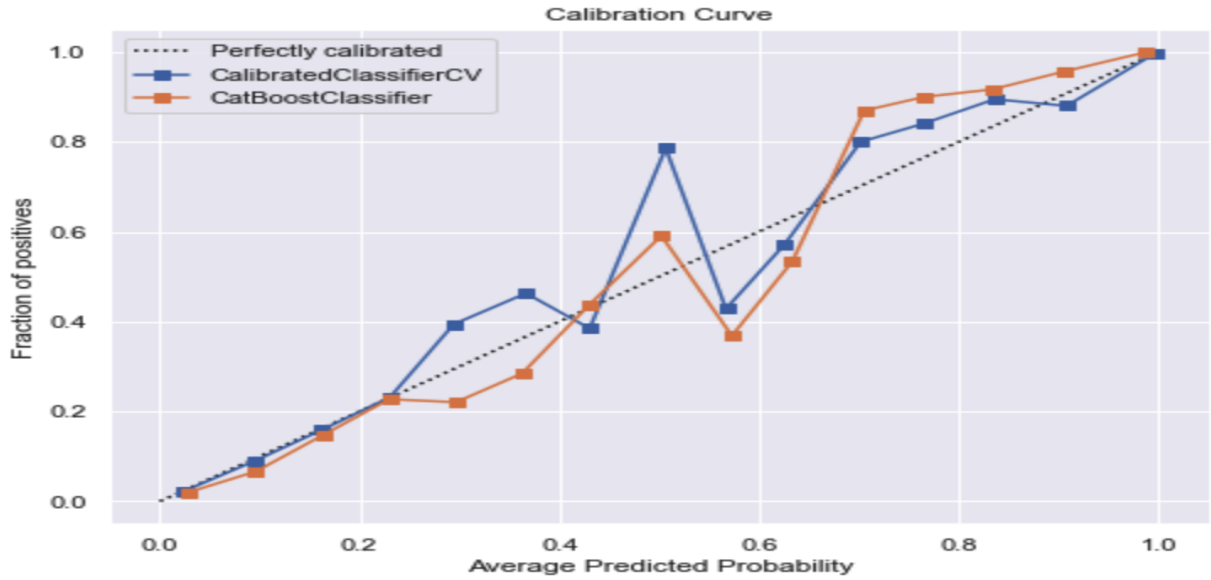
After carefully deciding that the CatBoostClassifier is the model we want to use, we proceed to fine-tune it so that we can examine the differences between the untuned and tuned CatBoostClassifiers.

The ROC curve and the use of a box-plot, which both demonstrate that the hypertuned catboost classifier performed better in both cases, help us to further support our conclusion. In addition, we checked the accuracy of the out-tuned model, which came out to be 0.9468 compared to the normal model's accuracy of 0.933605..





We need to check the calibration of two models (Calibrated and Non-Calibrated) to **see how stable the default prediction performance is across probabilities**. You can use a chart of each model's calibration to check this by calling the `calibration_curve()` function.



Visually speaking, the calibrated model seems to be better. However, we need to check a metric that will give us a number: Brier Score Loss.

```
Brier Score Loss (CatBoost Classifier (Tunned HP)): 0.05669827388618045
Brier Score Loss (CatBoost Classifier (Tunned HP + Calibrated)): 0.05422392485027883

Calibrated model has better calibration.
```

Saving this calibrated model we then implement this into our main aspect of our project, which is to how the use of our model or how the idea of not using any model affects our business in terms or loss incurred by the business.

To do this we first assess the financial impact using the default recall which is selected from the classification reporting using the function precision_recall_fscore_support(). For this, we will estimate the amount of unexpected loss using the default recall to find what proportion of defaults we did **not** catch with the new threshold. This will be a dollar amount which tells us how much in losses we would have if all the unfound defaults were to default all at once.
By this our estimate comes out to be **$3.3 million**

The next step is to determine the acceptance rates, which essentially involves determining the minimum percentage of new loans we are willing to take. The percentage of brand-new loans we wish to accept can be determined by setting an acceptance rate and figuring out the threshold for that rate. We're presuming that the test data consists of a brand-new batch of loans. To determine the threshold, we must utilize the numpy quantile() method. New loan status values should be assigned using the threshold. As can be observed from the dataset above, there are more Non-Defaults than Defaults, which causes our analysis to arrive at a 75% acceptance rate for our dataset.

Setting our threshold as 85% we check for how

the acceptance rate and threshold split up the data. Doing so we come to a conclusion that 6250 (0) people are not likely to default on their loan, whereas 1105(1) people will likely default on their obligation to fulfil the loan requirement.

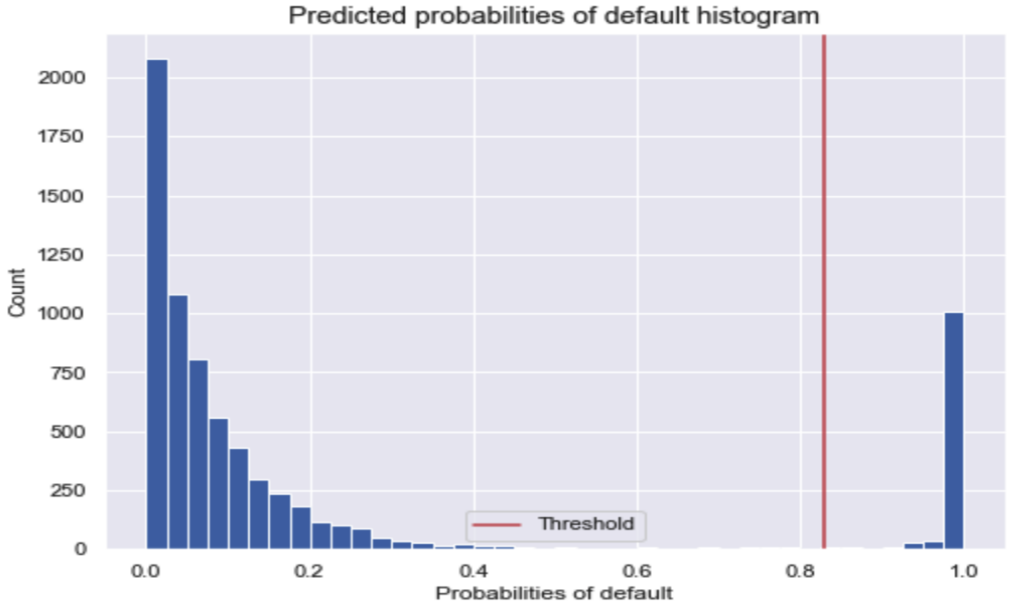
```
# calculates the threshold for a 85% acceptance rate
threshold_85 = np.quantile(test_pred_df['prob_default'], .85)

# applies acceptance rate threshold
test_pred_df['pred_loan_status'] = test_pred_df['prob_default'].apply(lambda x: 1 if x > threshold_85 else 0)

# prints the counts of loan status after the threshold
print(test_pred_df['pred_loan_status'].value_counts())
```

```
0    6260
1    1105
Name: pred_loan_status, dtype: int64
```

We know how quantile() works to compute a threshold, and we've seen an example of what it does to split the loans into accepted and rejected. What does this threshold look like for the test set, and how can we visualize it? To check this, we can create a histogram of the probabilities and add a reference line for the threshold. With this, we can visually show where the threshold exists in the distribution.



On the range of projected probabilities, we can see where the threshold is located here. We may also observe how many loans will be denied in addition to how many loans will be accepted (left side) (right side). To comprehend how this impacts the acceptance rate, we can run this code again with various threshold values. Now that we are aware of the acceptance rate, we can examine the bad rate within the accepted loans. We will be able to determine the proportion of defaults that have been accepted in this way. Consider the effects of the acceptance rate and the failure rate. Because defaults are more expensive,

we establish an acceptance rate to have fewer defaults in the portfolio. Will the bad rate be less than the percentage of defaults in the test data?

	true_loan_status	prob_default	pred_loan_status
0	1	0.040446	0
1	1	0.773904	0
2	1	1.000000	1
3	0	0.113816	0
4	0	0.000000	0

```
# creates a subset of only accepted loans
accepted_loans = test_pred_df[test_pred_df['pred_loan_status'] == 0]

# calculates the bad rate
print(np.sum(accepted_loans['true_loan_status']) / accepted_loans['true_loan_status'].count())
```

```
0.08306709265175719
```

This bad rate doesn't look half bad! The bad rate with the threshold set by the 85% quantile() is about 8.3%. This means that of all the loans we've decided to accept from the test set, only 8.3% were actual defaults! If we accepted all loans, the percentage of defaults would be around 22%.

In order to comprehend the influence on the portfolio for the acceptance rates, we also examine how this bad rate affects our acceptance rate, with a focus on the loan amt column of each loan. Cross tables with calculated values, like the new set of loans' average loan amount, are an option. We shall multiply the sum of each by the average loan amnt value for this. After that, these values are formatted as currencies.

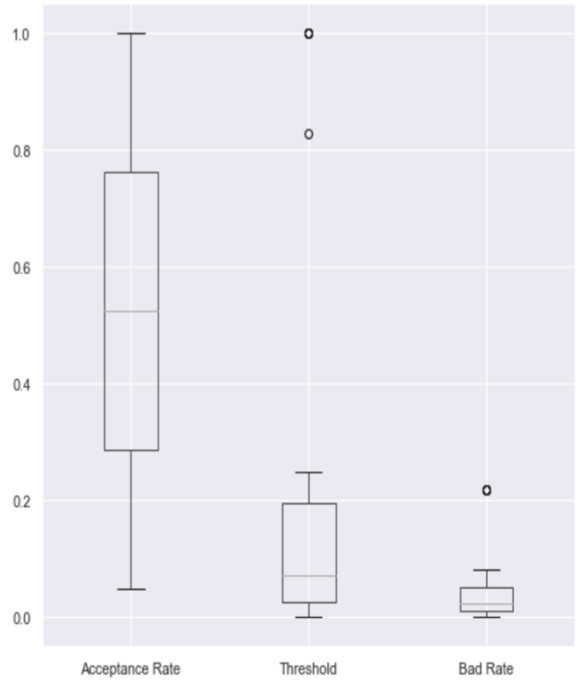
pred_loan_status	0	1
true_loan_status		
0	$55,171,473.25	$86,505.79
1	$4,998,112.56	$10,534,483.39

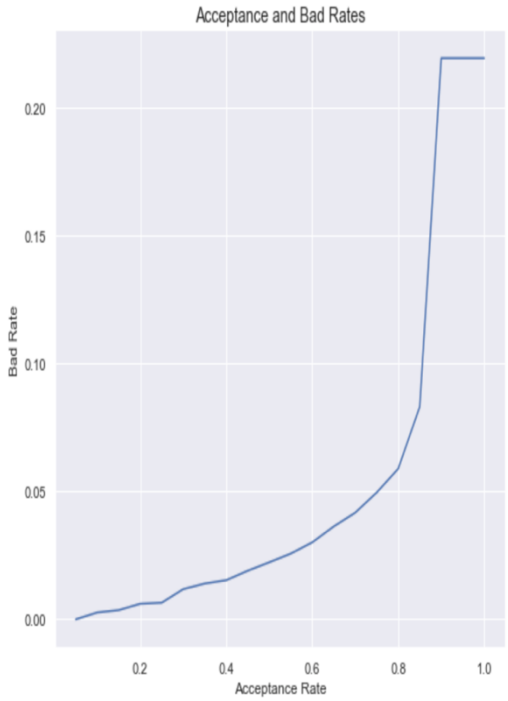
With this, we can see that our bad rate of about 8.3% ([1,0]) represents an estimated loan value of about **5.0 million dollars**. This may seem like a lot at first, but compare it to the total value of non-default loans ([0,0])! With this, we are ready to start talking about our acceptance strategy going forward.

Before we implement a strategy, we should first create a strategy table containing **all the possible acceptance rates you wish to look at along with their associated bad rates and threshold values**. This way, we can begin to see each part of our strategy and how it affects our portfolio.

This uses our specific predictions on the credit data, and can be used to see the acceptance rates, bad rates, and financial impact all at once. One of these values has the highest estimated value. Next we visualise our strategy table using the strat_df function.

We should check at the distribution of each column with a box plot. If the distribution of Acceptance Rate looks the same as the Bad Rate column, that could be a problem. That means that the model's calibration is likely much worse than we thought. We can also visualize the strategy curve with a line plot. The Acceptance Rate would be the independent variable with the Bad Rate as the dependent variable.





The boxplot shows us the distribution for each column. Look at the strategy curve. The bad rates are very low up until the acceptance rate 0.6 where they suddenly increase. This suggests that many of the accepted defaults may have a prob_default value between 0.6 and 0.8.

The strategy table, strat_df, can be used to maximize the estimated portfolio value and minimize expected loss. Extending this table and creating some plots can be very helpful to this end. Using this we find out that our maximum estimated value for our portfolio could be $50169400.649437.

With our credit data and our estimated average loan value, we clearly see that the acceptance rate 0.85 has the highest potential estimated value. Normally, the allowable bad rate is set, but we can use analyses like this to explore other options. For the final part of analysis and project we find out the total expected loss.
We've looked at some scoring and have seen samples of the predictions, but what is the overall effect on portfolio performance? Try using expected loss as a scenario to express the importance of testing different models.

It's time to estimate the total expected loss given all our decisions. The data frame test_pred_df has the probability of default for each loan and that loan's value. We'll use these two values to calculate the expected loss for each loan. Then, we can sum those values and get the total expected loss.

- Probabilities of default (PD)
- The loss given default (LGD)
- The loan_amnt which will be assumed to be the exposure at default (EAD).

$$Total\ Expected\ Loss = \sum_{x=1}^{n} PD_x * LGD_x * EAD_x.$$

expected_loss = prob_default * LGD * loan_amnt

We'll assume that the exposure is the full value of the loan, and the loss given default is 100%. This means that a default on each loan is a loss of the entire amount**.**
The total expected loss comes out to be $17,519,110.78, this is the total expected loss for the entire portfolio using the CatBoost (HP + Calibrated).
Had we chosen other models that we tested or didn't chose a model all together our losses would be as shown below.

Used algorithm	Calibrated?	Boruta?	Algorithm that was used by Boruta	Number of remaining features	Total expected loss
No algorithm, just human guessing	N/A	N/A	N/A	26	35,395,287.50
XGBoost Classifier	No	Yes	Random Forest Classifier	7	29,276,107.72
XGBoost Classifier	Yes	Yes	Random Forest Classifier	7	28,948,507.08
XGBoost Classifier	Yes	Yes	XGBoost Classifier	19	27,821,760.79
XGBoost Classifier	Yes	No	N/A	26	19,596,497.74
Stacking Classifier XGB	Yes	No	N/A	26	20,431,941.38
CatBoost Classifier (HP)	Yes	No	N/A	26	17,547,088.71

This is the total expected loss for the entire portfolio using the CatBoost (HP + Calibrated).
$17.55 million may seem like a lot, but the total expected loss would have been over $29 million with the XGBoost non-calibrated and **$35,395,287.50 without the any algorithm!** Some losses are unavoidable, but our work here might have saved the company about $17.8 million dollars!

CONCLUSION

Only through its measurement is credit risk management possible. Models are the most efficient instruments for measuring how exposed different financial organizations are to risk. The management of credit risk will improve in effectiveness and efficiency with the right measurement.

This research focuses on creating a method to assess the credit risks connected to different bank borrowers. The primary evaluation criteria for the bank are used as the predictor variables in this. There are numerous methods for creating credit risk models, many of which have already been covered in the interim report. Since each strategy has advantages and disadvantages, it is difficult to claim with certainty which one can anticipate default the best. The decision is based on the unique business conditions and portfolio characteristics of each bank. Depending on the situation, it may occasionally be wise to combine several different approaches to improve the bank's credit decision system.

Data accessibility is a significant barrier to such studies, and with the availability of more reliable data, the results may be even more beneficial to banks.

The internal risk management at banking institutions may in fact improve as a result of the credit risk modelling. Before models can be employed in the process of determining regulatory capital requirements, however, significant obstacles must be overcome, most notably those relating to data constraints and model validation.

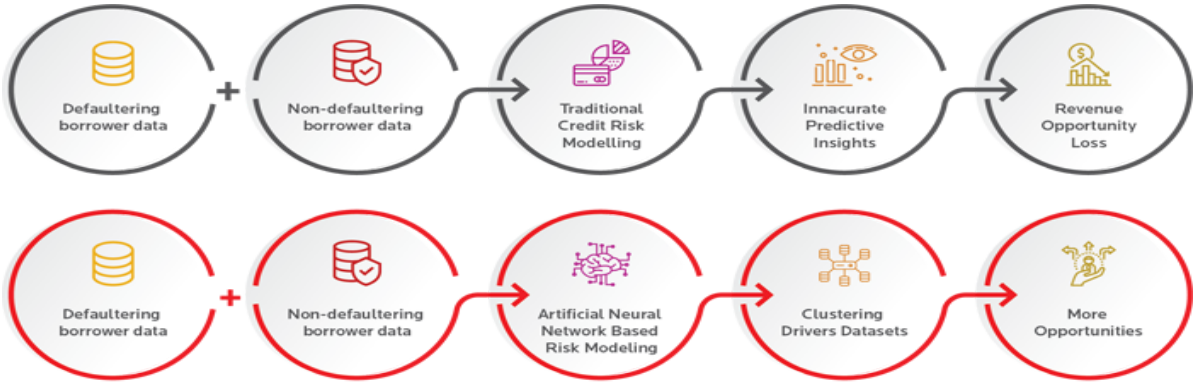
FUTURE SCENARIO

Risk management is at the top of the list of possible use cases for AI & ML technology in the BFSI industry. The number of businesses adopting AI more than doubled between 2017 and 2018, and 40% of financial services companies are using it to manage risk. This is due to the possibility that AI and ML could offer real value at every stage of the credit value chain, from early underwriting to risk measurement and analysis to determining the final maximum exposure. The following are some of the main use cases that would be covered:

- Underwriting
- Risk Measurement
- Exposure Decision

For instance, the vast majority of applicants in each borrower group will be non-defaulters, with defaulters making up a very small percentage of the total. Therefore, if banks are utilizing conventional analytical methods, a sample of undesirable consumers entering the credit dataset could lead to imbalance and skew results. The predictive insights are wrong as a result of the performance degradation brought on by this gap, and the bank misses out on good commercial prospects.

To determine whether a certain consumer should be offered a loan, an ML model, such the Artificial Neural Network, would build discrete clusters of datasets and use merging procedures. ML builds majority and minority clusters and merges them to create a varied dataset that reflects the actual situation on the ground, as opposed to only looking at the mean values.



How machine learning models lead to better revenue opportunities

Historically, credit risk approaches have relied on statistical techniques like Logistic Regression and Linear Discriminant Analysis. Large datasets cannot be handled by these methods, though. This is where AI steps in.

This has occurred because of AI's true ability to add value throughout the entire credit value chain. The majority of AI adoptions center on managing and analyzing credit risk. Currently, the field of credit risk management serves as a testing ground, and both demonstrates and exposes the advantages of utilizing AI. Credit AI opens the door for more applications of the technology, such as wholesale banking, retail banking, insurance, wealth management, and capital markets, through this iterative process.

REFRENCES

- https://www.lexingtonlaw.com/credit/length-of-credit-history
- https://www.investopedia.com/terms/c/creditrisk.asp
- https://www.youtube.com/watch?v=bx_LWm6_6tA - The Crisis of Credit Visualized
- https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-risk/
- https://www.sas.com/en_us/insights/risk-management/credit-risk-management.html
- https://www.federalreserve.gov/econres/notes/feds-notes/the-pandemics-impact-on-credit-risk-averted-or-delayed-20210730.html
- https://www.wallstreetmojo.com/credit-risk/
- https://www.birlasoft.com/articles/ai-machine-learning-and-future-credit-risk-management