Bank Loan Default

Capstone Project Report

Submitted by:

Balajisriram Venkateshkumar

Bina Rajput

Narendra Kr. Gupta

Rakendu Sharma

Shailajha Deepak

Under the guidance of

Harshal Jawale

Batch: November 2020 A, Group BLD-2

Year of Completion: 2021

ACKNOWLEDGEMENT

We would like to express our gratitude and appreciation to our mentor Mr. Harshal Jawale and Professor Vinit Thakur for providing guidance and valuable feedback on the project, and also to the mentors who evaluated our interim report, presentation and shared constructive feedback.

Special thanks to Great Lakes Institute of Management for giving us this opportunity and our batch mates, program management and most importantly our family members who have helped us by providing us with the facilities and conducive conditions to successfully deliver the capstone project.

GLOSSARY OF TERMS / ABBREVIATIONS

Abbreviations	Descriptions
EDA	Exploratory Data Analysis
TP, TN, FP, FN	True Positive, True Negative, False Positive, False Negative
TPR, TNR	True Positive Rate, True Negative Rate
Specificity or TPR	True Negative / (False Positive + True Negative) Specificity is the complement to sensitivity, or the true negative rate, and summarises how well the negative class was predicted.
Sensitivity or TNR	True Positive / (True Positive + False Negative) Sensitivity refers to the true positive rate and summarizes how well the positive class was predicted.
G - mean	SQRT(Sensitivity * Specificity) Sensitivity and Specificity can be combined into a single score that balances both concerns, called the geometric mean or G-Mean.
Recall	True Positive / (True Positive + False Negative) Recall summarizes how well the positive class was predicted and is the same calculation as sensitivity.
Precision	True Positive / (True Positive + False Positive) Precision can be measured as of the total actual positive cases, how many positives were predicted correctly.
F-Measure	(2 * Precision * Recall) / (Precision + Recall) Precision and recall can be combined into a single score that seeks to balance both concerns, called the F-score or the F-measure. The F-Measure is a popular metric for imbalanced classification. F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.
AUC	The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.
SMOTE	Synthetic Minority Oversampling Technique (SMOTE) is an oversampling technique where the synthetic samples are generated for the minority class. This function used to bring highly imbalanced dependent variables to normalized dependent variables. The approach is effective because new synthetic examples from the minority class will be created relatively.
Apriori	Apriori is an algorithm for frequent item set mining and association rule learning over relational databases.
Text Analytics	A collection of features from Cognitive Service for Language that extract, classify and understand text within documents.
Anomaly Detection	Anomaly Detection is the technique of identifying rare events or observations which can raise suspicions by being statistically different from the rest of the observations.
Skewness	Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.
t-SNE	t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data.

Table 1 Glossary

EXECUTIVE SUMMARY

The purpose of this study is to build a machine learning model predicting bank loan defaults by undertaking comprehensive research on the information on consumer loans. The model will be applied on preliminary information provided by the customer while applying for the loan. The model is of paramount importance since bad loans are a critical risk faced by banks and financial institutions having far-reaching impact on the economy. Hence, the model will be used as a decision-making tool while sanctioning loans, which will help banks or financial institutions to lower the risk and maximize profitability.

Problem Statement

Loan default is a major risk faced by banks and financial institutions since it impacts profitability. Loan default risk is the risk that the borrowers would not be able to make the payments on their debt obligations. **The provisions for loan defaults reduce the total loan portfolio of banks, which lowers interest earnings on such assets. Since loan defaults impact profitability of banks; it also affects the dividend pay out to the shareholders.** Our goal is to help banks and financial institutions minimize defaults and improve bottom line.

Loans by banks are one of the key sources of capital in the economy. Lending growth is considered as important factor for inflation level and interest rate in an economy, reflecting economic condition and growth. Hence, financial health of banks and financial institutions becomes critical for the economy. **Higher defaults across the banking system can also have an impact on the growth of the economy.** Default prediction model would allow to determine whether the borrower would default in debt repayment.

Though the study is limited to consumer loans, we understand that the bigger problem of default is related to massive commercial loans. We have not included commercial lending in our study since the data is limited to consumer loans. For a commercial default prediction model, we would need information on commercial lending.

Main Results

To deal with the problem of default, we did a comprehensive study focussed on application of data analytics techniques and machine learning to build prediction model. Default prediction by the model would help in enhancing profitability of banks. Banks have vast information available on the borrowers and the same information has been used for the model building.

We used different model and techniques arriving at the following key results:

- Certain category of data such as debt-consolidation (purpose of loan) and State of California are more prone to defaults. More details have been covered in the later sections of this report.
- Classification model is appropriate for this study because the Target Variable which is ‘loan_status’ has two unique classes 1) Fully Paid, 2) Default.
- Two Class Boosted decision tree is the most appropriate model as it shows better performance while predicting both the classes.
- On analysis of results from the model, we arrived at business problems that the banks or financial institutions may face while deciding on approval of the loan. The two business problems identified are described in detail in the figure below:

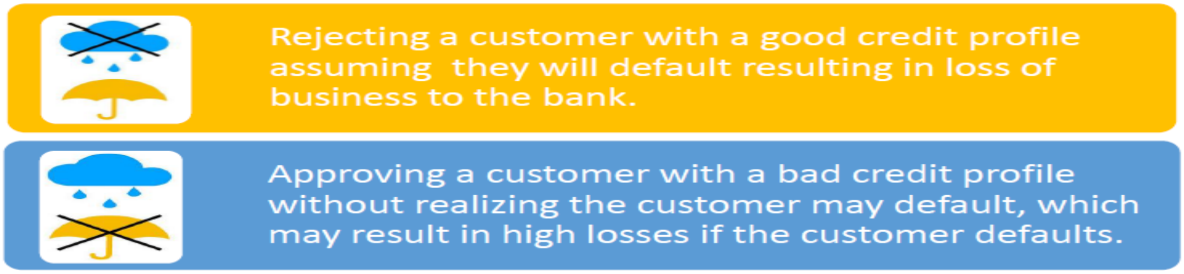


Figure 1 Types of Business Risks

Recommendations

According to the detailed analysis of the given data, our recommendations to the bank are:

- Our model indicates a good prediction of defaulters. Hence, banks should use our model based predictions to undertake proper risk benefit evaluation for customers predicted to default by our model. The variables 'funded_amnt_inv', 'loan_amnt', and 'installment' are considered as very critical by the model while predicting defaults, hence would have the most impact on the sanction of loans.
- Banks to explore growth opportunities through proper risk benefit evaluation of customers predicted to default by using our best selected model. For example: higher default categories include debt-consolidation, higher dti between 15 to 20, multiple credit lines, customers with property on mortgage, and State of California, and customers paying high installments.
- High collateral or guarantee can be undertaken for high risk borrowers (predicted to default by the model) to mitigate the losses.
- There are some risk free purposes of loans such as education, and renewable energy. Banks can expand into those areas so that the loans are less risky and defaults would decline. This would improve profitability of the bank.
- **Banks provide good, reliable data with lower missing values** to build a more robust prediction model.
- Banks can also **capture the additional demographic data from loan applicants such as age, gender, number of dependents, heuristics which will certainly enhance the model performance.**

To summarise, a better understanding of the customer behaviour and regional differences would help the banks make informed lending decisions. We recommend banks to use our machine learning prediction model which has Accuracy and Precision of 97% to predict default and improve their profitability.

Introduction

The increasing requirements of financial institutions to have robust risk management has led development of current methods of risk estimation. The implementation of **machine learning techniques could lead to better quantification of the financial risks that banks are exposed to**. There are different risk measures banks consider in order to estimate the potential loss they may carry in future.
Customers in default means that they did not meet their contractual obligations and might not be able to repay their loans. Thus, there is an interest in acquiring a model that can predict defaulted customers.
A **technique that is widely used** for estimating the probability of client default is **Supervised Machine Learning**.
In this report a set of machine learning methods will be investigated and studied in order to test if they can challenge the traditionally applied techniques.
Benefits of Supervised Machine learning bring about the following:

- **Banks can diagnose their current statuses based on prediction models and establish their strategies**.
- Executives can run their businesses more stably by managing **key indicators that affect default risk**.
- Investors can revise their strategies and improve their portfolios by examining the likelihood of fallacies and improve related financial regulations using default predictions. In these ways, default prediction models help in designing and improving the financial system.
- By employing machine learning algorithms and statistical models, **default predictions are at the cutting edge of advanced financial engineering**.
- The recent global financial crisis and the increase in credit risk highlights the importance of Machine Learning.

As depicted in below trend charts developed from the given dataset:

- Demand of loan increased significantly in last two years.
- In the same period, the percentage of loan defaulters also increased considerably.

So need for proper risk evaluation is essential, for balancing Risks and Rewards for the bank's loan business.

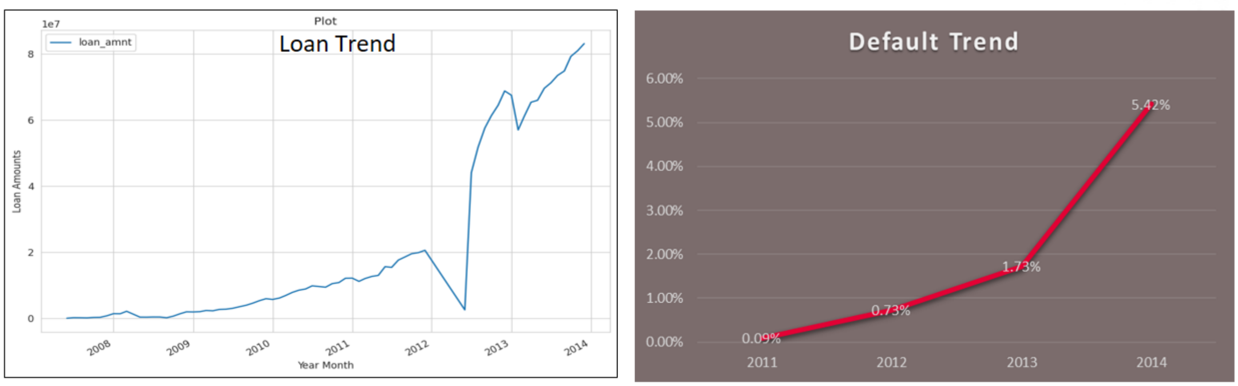


Figure 2 Loan Demand and Default trends

Data Source

Dataset named "Bank Loan Default" was received from Great Lakes University towards our capstone project.

Approach used and Limitations

The study is limited to consumer loans, we understand that the bigger problem of default is related to massive commercial loans. We have not included commercial lending in our study since the data is limited to consumer loans. For a commercial default prediction model, we would need information on commercial lending.

Literature Review

This section discusses in brief about some of the work that has already been done on creating Machine Learning models using various algorithms to improve the loan default prediction process and help the banking authorities and financial institutions to select a loan application based on the band of low to high risk. Based on the same, frame stringent contractual terms and conditions such as higher collateral and interest rate. Loan default prediction is a burning topic to talk about in Banking Sector.

Credit scoring is an important measure to gauge the consumers in this competitive financial world. Furthermore, the recent developments in Data Science and Machine learning has gained more attention and research interest on loan prediction and credit risk assessment. Due to the high demands of loans now, demand for further improvements in the models for loan default prediction is increasing significantly.

A multitude of techniques have been used to assign individuals a credit score and much research has been done over the years on the topic. Unlike previously, where experts were hired and the models depended on professional opinions for assessing the individual's creditworthiness, the focus has now shifted to an automated way of doing the same job. In recent years, the researchers and banking authorities have been focused on applying machine learning algorithms for Loan default predictions. Many noteworthy conclusions have been drawn in this regard which serve as stepping-stones for research and studies.

Methodology of the Study

The figure below elaborates clearly the phase wise methodological study of the business problem involving tools, techniques, machine learning models, model tuning to deliver an insightful observations and recommendations.

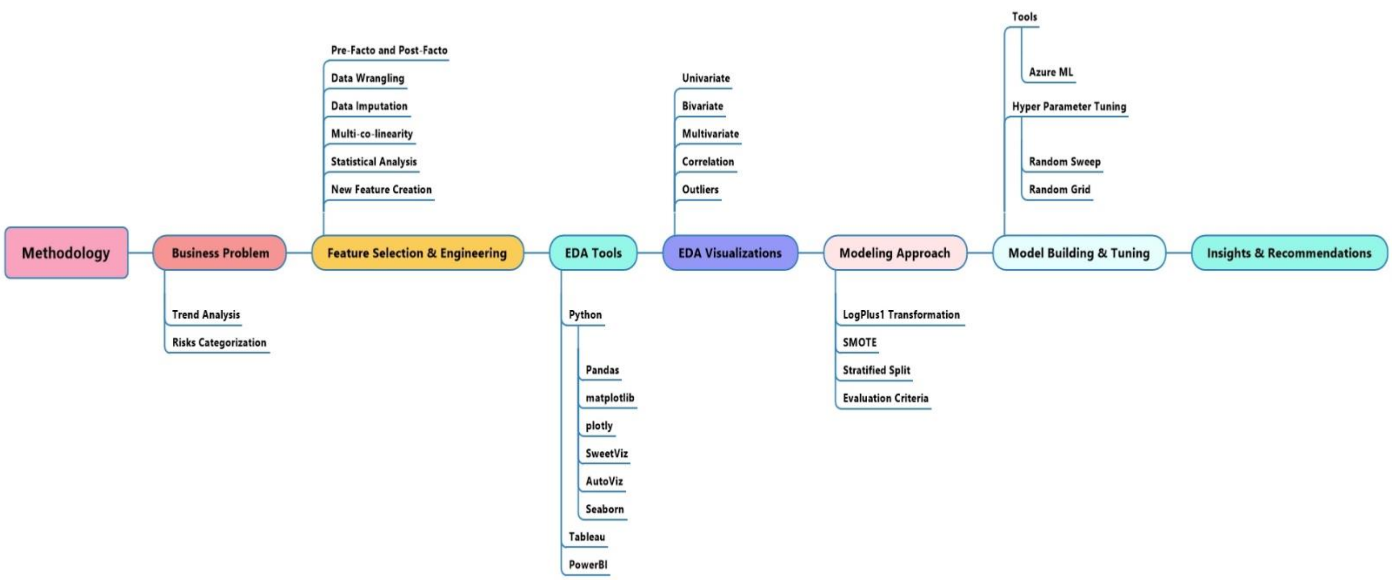


Figure 3 Methodology

Data Description and Feature Engineering

This section describes in detail the analysis of dataset statistically and visually to discover insightful inferences for model building and business recommendations.

Understanding the Dataset

There are 41 columns, 119145 rows having data types - integer, float, string, and date time. The data file size is 37.3+ MB. Non-null counts and a detailed description of each field is mentioned in the table below:

#	Fields	Description	Non-Null Count	Data Type
1	member_id	A unique Id for the borrower member.	119145	int64
2	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	119145	int64
3	funded_amnt	The total amount committed to that loan at that point in time.	119145	int64
4	funded_amnt_inv	The total amount committed by investors for that loan at that point in time.	119145	float64
5	term	The number of payments on the loan. Values are in months and can be either 36 or 60.	119145	object
6	int_rate	Interest Rate on the loan	119145	float64
7	installment	The monthly payment owed by the borrower if the loan originates.	119145	float64
8	grade	Assigned loan grade	119145	object
9	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	115306	object
10	home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.	119145	object
11	annual_inc	The self-reported annual income provided by the borrower during registration.	119145	float64
12	verification_status	Status of the verification done	119145	object
13	issue_d	The month which the loan was funded	119145	datetime64[ns]
14	pymnt_plan	Indicates if a payment plan has been put in place for the loan	119145	object
15	desc	Loan description provided by the borrower	61599	object
16	purpose	A category provided by the borrower for the loan request.	119145	object
17	addr_state	The state provided by the borrower in the loan application	119145	object
18	dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.	119145	float64
19	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years	119145	int64
20	earliest_cr_line	The month the borrower's earliest reported credit line was opened	119145	datetime64[ns]
21	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)	119145	int64
22	mths_since_last_delinq	The number of months since the borrower's last delinquency.	49916	float64
23	open_acc	The number of open credit lines in the borrower's credit file.	119145	int64
24	revol_bal	Total credit revolving balance	119145	int64

25	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	119053	float64
26	total_acc	The total number of credit lines currently in the borrower's credit file	119145	int64
27	out_prncp	Remaining outstanding principal for total amount funded	119145	float64
28	out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors	119145	float64
29	total_pymnt	Payments received to date for total amount funded	119145	float64
30	total_pymnt_inv	Payments received to date for portion of total amount funded by investors	119145	float64
31	total_rec_prncp	Principal received to date	119145	float64
32	total_rec_int	Interest received to date	119145	float64
33	total_rec_late_fee	Late fees received to date	119145	float64
34	recoveries	post charge off gross recovery	119145	int64
35	collection_recovery_fee	post charge off collection fee	119145	int64
36	last_pymnt_d	Last month payment was received	119145	datetime64[ns]
37	last_pymnt_amnt	Last total payment amount received	119145	float64
38	next_pymnt_d	Next scheduled payment date	3283	datetime64[ns]
39	last_credit_pull_d	The most recent month pulled credit for this loan	119137	datetime64[ns]
40	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	119145	object
41	loan_status	Current status of the loan	119145	object

Table 2 Dataset Described

Predictor Variables: There are 3 types of variables amongst 41 parameters: 10 categorical, 24 numerical, 5 Date Time

Feature Engineering

Data Pre-processing: In data pre-processing, we have cleaned the missing data by using **replacement techniques** and removed **unwanted columns**, which were not useful for prediction modelling. This step is critical because it helps in cleaning the data which enhances the performance of the model and avoids **Overfitting errors**.

Data wrangling: While reviewing the given dataset, it was observed that there are missing values and some columns are having datatype error. For example: ‘emp_length’ variable, the ideal data type should be ‘integer’ for this variable but it’s ‘String’. So, Data wrangling is applied here to convert the data into desired data type. Data wrangling is the process of transforming and mapping data from one "raw" data form into another format to make it more appropriate and valuable for a variety of downstream purposes such as analytics.

Data imputation: Data imputation is done to fill the missing values and logically replaced them either by 0 or by average of the available values in that particular variable. Data imputation is defined as the substitution of an estimated value that is as realistic as possible for a missing or problematic data item. The substituted value is intended to enable subsequent data analysis to proceed.

The following are the fields wherein Data wrangling and Data imputation are applied.

#	Fields	Description	Data Wrangling/ Data Imputation
9	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	Converted the data to numerical instead of string. Removed words like 'Years' and '+' Also replaced n/a with 0.
22	mths_since_last_delinq	The number of months since the borrower's last delinquency.	69229 Blanks so below is the strategy: - For 67510 'Fully Paid' Loan Status entries - missing values of this column were replaced by 0. - For 1719 'Default Loan Status missing values will be replaced by 'mean of available values' in this column corresponding to 'Default' Loan Status. Average: 34.03580563 Count: 1564 Sum: 53232 So, taking average as **34.036**
25	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	92 cells were blank/missing values, which were replaced by 0.

Table 3 Data Wrangling and Data Imputation

The following features are not considered for modelling due to the reasons mentioned below:

#	Fields	Description	Remarks / Observations and Data Cleaning
1	member_id	A unique Id for the borrower member.	This will not be useful for prediction models, but kept for **pivoting**.
15	desc	Loan description provided by the borrower	This was separated to do **Text Analytics**.
34	recoveries	post charge off gross recovery	All values are '0', hence removed from the data set.
35	collection_recovery_fee	post charge off collection fee	All values are '0', hence removed from the data set.
38	next_pymnt_d	Next scheduled payment date	115862 elements are blank. Removed from the data set.
39	last_credit_pull_d	The most recent month pulled credit for this loan	8 elements are blank. This would be removed in the Azure or Python or Tableau Modelling if the column is important from modelling perspective.
40	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	Only one type of data - this will not affect the prediction of models. Hence removed from the data set.

Table 4 Data Cleaning

Feature Selection

It can be inferred from the given dataset, Target or Dependent variable is 'loan_status'. Since dependent variable is known here, supervised method for machine learning has been applied.

Pre-Facto and **Post-Facto:** Data variables are split in two parts, which are termed as Pre-Facto and Post-Facto. **Only Pre-Facto variables are used for prediction modeling**. Pre-Facto variables are those which are available at the time of filing loan application. **Out of 41 variables we have considered only 17 variables as input variables for prediction modeling.**

Multi-collinearity: During data analysis, it is observed that two variables i.e. loan_amnt and funded_amnt are having the same values so only one variable is considered out of these two as input variable to address Multi-collinearity error in the dataset.

Final set of feature for modelling are [*loan_amnt, funded_amnt_inv, term, int_rate, installment,grade, emp_length, home_ownership, annual_inc, purpose, addr_state, dti, open_acc, revol_bal, revol_util, total_acc* , *loan_status*].

New Feature Creation

New feature creation from the existing features by using various techniques like Binning, Splitting, Date/Time Decomposition, Compound String Splitting, and One-Hot Encoding. These techniques are useful for insightful EDA.

One-Hot Encoding: technique is used to convert the Categorical Feature into Numerical Feature. Refer Appendix Python Code Snippet [1]

Numerical Binning: technique is used to create new features **by grouping** respective feature data based on the **value ranges** for the existing features. Following new features were created using Python code Refer Appendix Python Code Snippet [2]

#	Existing Feature	New Feature	Bins
1	loan_amnt	loan_amnt_range	['0-5000', '5000-10000', '10000-15000', '15000-20000', '20000-25000','25000+']
2	int_rate	int_rate_range	['0-7.5', '7.5-10', '10-12.5', '12.5-15', '15+']
3	annual_inc	annual_inc_range	['0-25000', '25000-50000', '50000-75000', '75000-100000', '100000+']
4	installment	installment	['low', 'medium', 'high', 'very high']
5	dti	dti_range	['0-5%', '5-10%', '10-15%', '15-20%', '20-25%', '25%+']

Table 5 New Feature Creation

Exploratory Data Analysis

An EDA is a thorough examination meant to uncover the underlying structure of a data set and is critical for data analytics because it exposes trends, patterns, and relationships that are not readily apparent. We have done Statistical, Univariate, Bivariate and Multivariate, Correlation analysis of the fields selected for model building.

Statistical Analysis of Dataset for EDA

In the tables below statistical data of finalized dataset for modelling is provided. This doesn't include new features created as these are created only for insightful EDA and will not be used for modelling.

Categorical columns	count	unique	top	freq
term	119145	2	36 months	98567
grade	119145	7	B	41450
home_ownership	119145	5	MORTGAGE	59846
purpose	119145	14	debt_consolidation	67115
addr_state	119145	50	CA	20891
loan_status	119145	2	Fully Paid	115862

Figure 4 Statistics of Categorical Features

Numerical columns	count	mean	std	min	25%	50%	75%	max
member_id	119145.0	4288577.055168073	3466993.5488595343	70699.0	1240242.0	2839634.0	7277043.0	12096968.0
loan_amnt	119145.0	12983.233245205422	7814.069451487893	500.0	7000.0	11300.0	18000.0	35000.0
funded_amnt	119145.0	12915.291451592597	7773.289945012495	500.0	7000.0	11200.0	17625.0	35000.0
funded_amnt_inv	119145.0	12768.857109709606	7801.0406259928495	0.0	6975.0	11000.0	17500.0	35000.0
int_rate	119145.0	13.293939989098295	4.255293900503904	5.42	10.16	13.11	15.96	26.06
installment	119145.0	405.6382267824621	240.12560518327885	15.69	226.07	357.03	532.35	1407.01
emp_length	119145.0	5.615283897771623	3.558067823911458	0.0	2.0	5.0	10.0	10.0
annual_inc	119145.0	72715.24845826512	61416.41564303386	4000.0	45000.0	62000.0	87000.0	7141778.0
dti	119145.0	15.74773964497029	7.5559785253060765	0.0	10.07	15.43	21.11	34.99
open_acc	119145.0	10.560073859582861	4.597445960677471	0.0	7.0	10.0	13.0	52.0
revol_bal	119145.0	15186.570909396114	17969.635467242526	0.0	5966.0	11135.0	19269.0	1743266.0
revol_util	119145.0	54.40327097234503	25.11520176071212	0.0	36.4	56.4	74.4	122.5
total_acc	119145.0	24.30443577153888	11.364550319999728	2.0	16.0	23.0	31.0	99.0

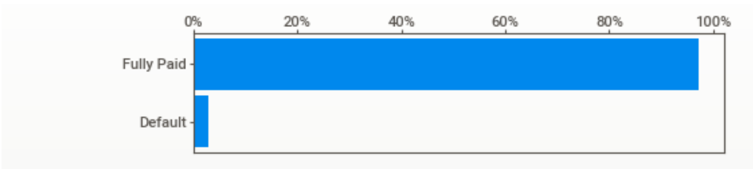
Figure 5 Statistics of Numerical Features

Looking at the descriptive statistics we could infer that:

- Variables namely employee length, dti have the 75 percentile values as influencers for defaults.
- The data is completely skewed, and the distribution is not normal

Univariate Analysis

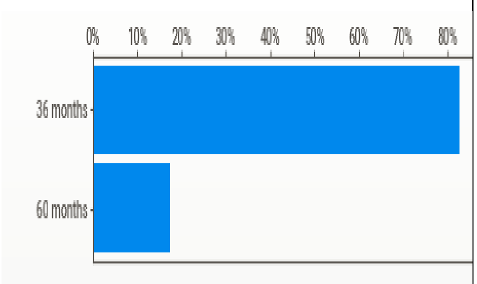
Target Variable - loan_status

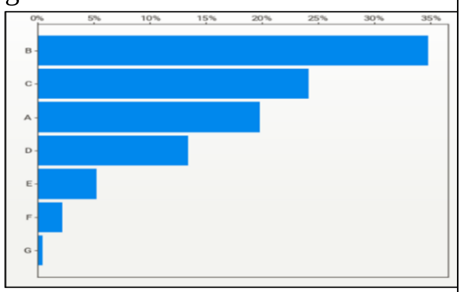
<table>
<tr><td>loan_status

</td><td>Fully Paid 97%
Default 3%

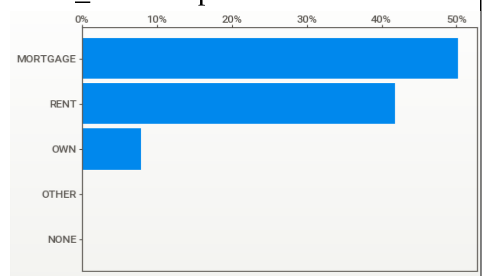
Considering the nature of Dependent variable having categorical data type and two unique values, Prediction model based on Two Class Classification Algorithms will be applicable.</td></tr>
</table>

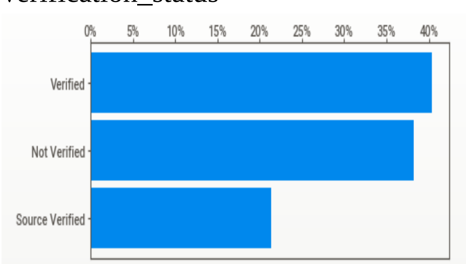
Table 6 Target Variable

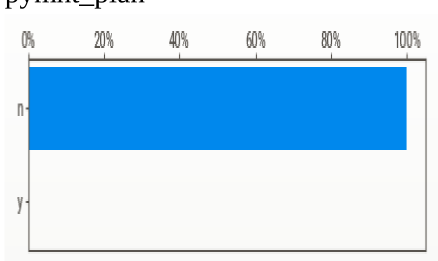
Key Categorical Features

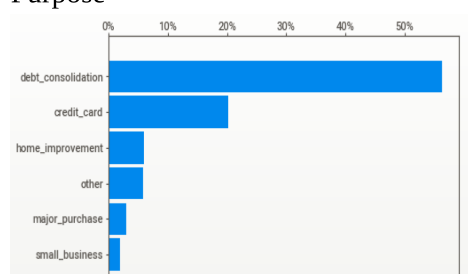
<table>
<tr><td>term


The loan term is either 36 months or 60 months.</td><td>grade


Customer with Grade B (35%) have taken most loans followed by
C (24%) and A (20%) least being G (<1%)</td><td>home_ownership


Highest percentage of loans are taken by customers who have Mortgage (50%) followed by
Rent (42%)</td></tr>
<tr><td>verification_status


Percentage of Verified customers is 40%, and Not Verified are 38%</td><td>pymnt_plan


For more than 99% of customers there is no payment plan.</td><td>Purpose


The top most purpose of loan is debt_consolidation 56% followed by credit_card 20%</td></tr>
</table>

Table 7 Univariate Analysis of Categorical Features

Key Numerical Features

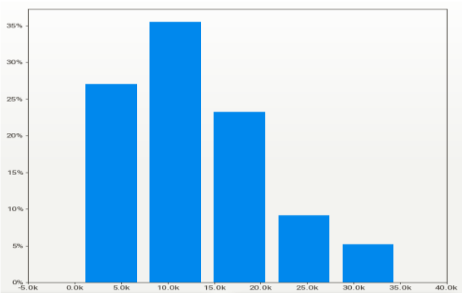
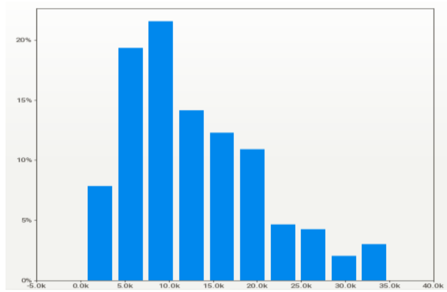
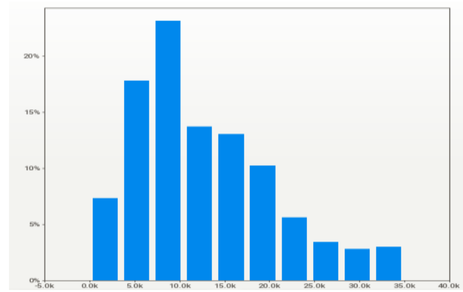
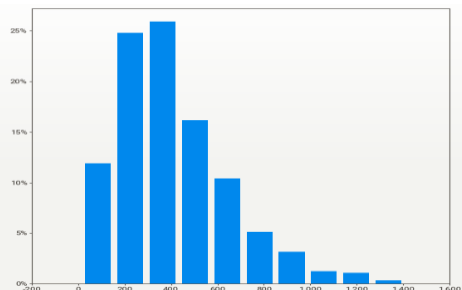
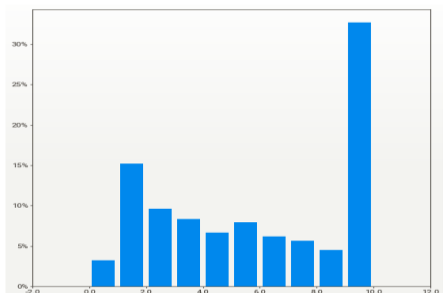
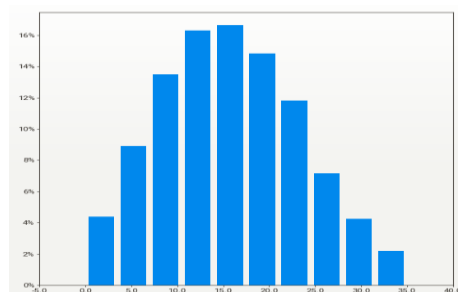
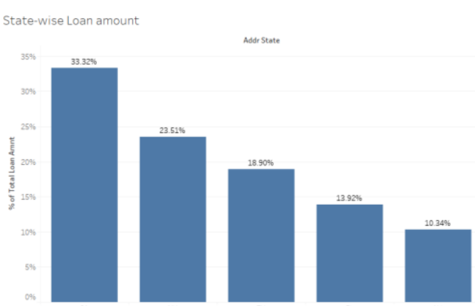
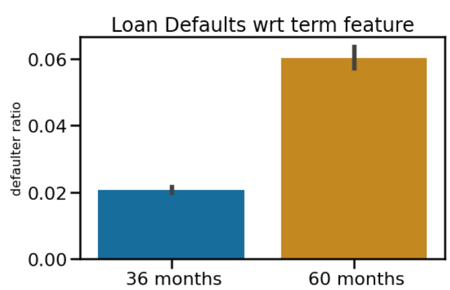
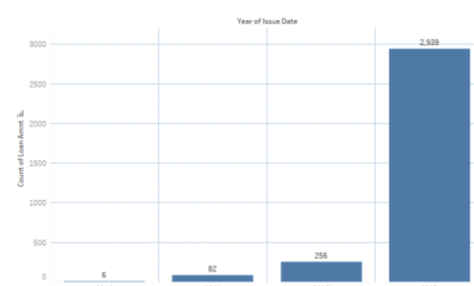
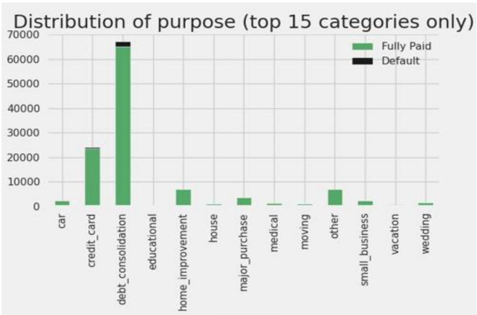
<table>
<tr>
<td>loan_amnt

Most frequent loan amounts $10000 (7.5%) and $12000 (5.9%) - skewed data</td>
<td>funded_amnt

Maximum loan amount is $35,000 whereas minimum loan amount in $500 - skewed data</td>
<td>int_rate

Maximum Interest rate is 26.1% while minimum Interest rate is 5.4% - skewed data</td>
</tr>
<tr>
<td>Instalment

Maximum Instalment paid by customers is $1407 while minimum being $16 - skewed data</td>
<td>emp_length

Most loans are disbursed to customers having 10 or 10+ years of experience followed by 1 year and 2 years - skewed data</td>
<td>dti

The maximum value of DTI for customers is 34.99 and least being 0.0</td>
</tr>
<tr>
<td>Address State

CA,NY,TX,FL & NJ are the top 5 states, having default</td>
<td>Term

Loan Defaults in 60 months term are thrice to 36 months.</td>
<td>Loan Defaults Trend

Loan defaulters increased exponentially in 2013 (2939 no’s)</td>
</tr>
</table>

Table 8 Univariate Analysis of Numerical Features

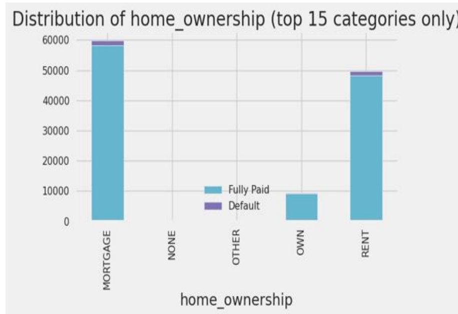
State wise Fully Paid and Default count (ID, IA, MS & ME states shows no Default) Refer Appendix Exploratory Data Analysis Plots [6]

Bivariate Analysis

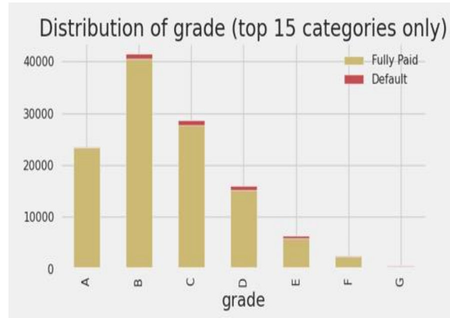
Categorical Independent Vs Target



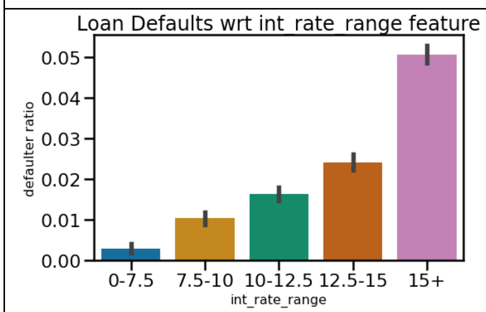
Highest number of default happen when the purpose of loan is 'debt_consolidation'.



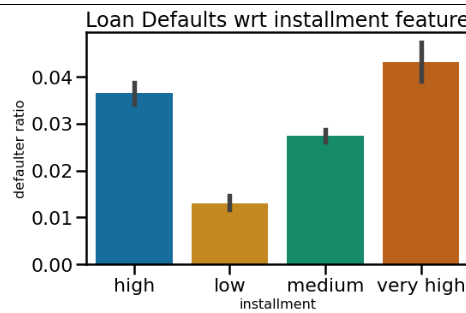
Highest number of defaulters have mortgaged their property or live on rent.



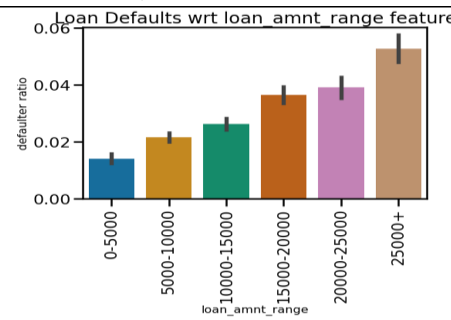
Defaulters are from Grade B, C and D, while A, F, and G grades have none. Maximum defaulters are in Grade C, followed by Grade B



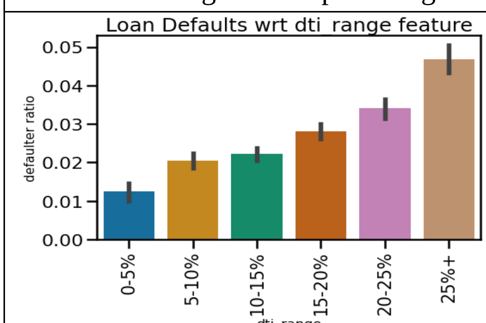
Maximum Loan Defaults happened for interest rate range of 15+ percentage



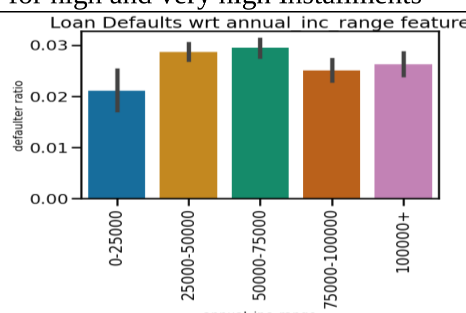
Maximum Loan Defaults happened for high and very high Installments



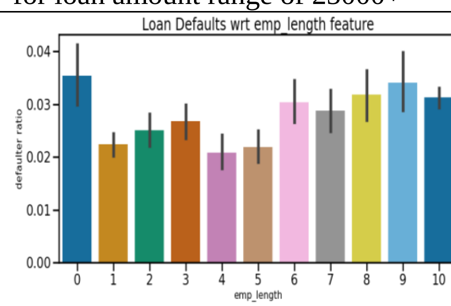
Maximum Loan Defaults happened for loan amount range of 25000+



Maximum Loan Defaults happened for DTI range of 25+ percentage



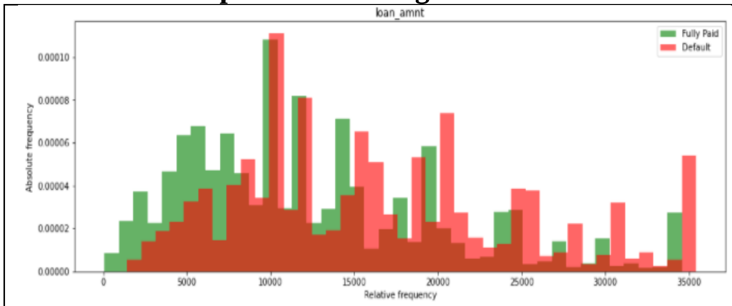
Loan default happen irrespective of annual income of customers.

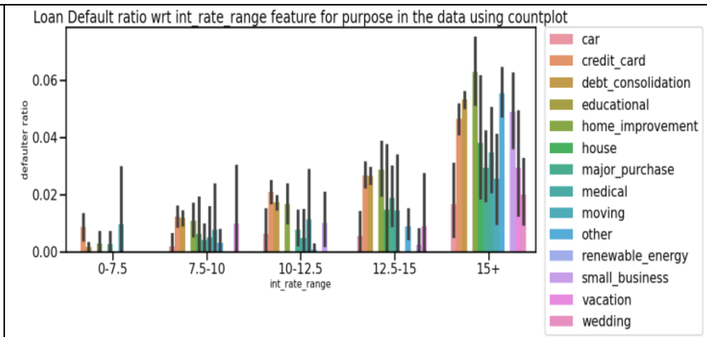


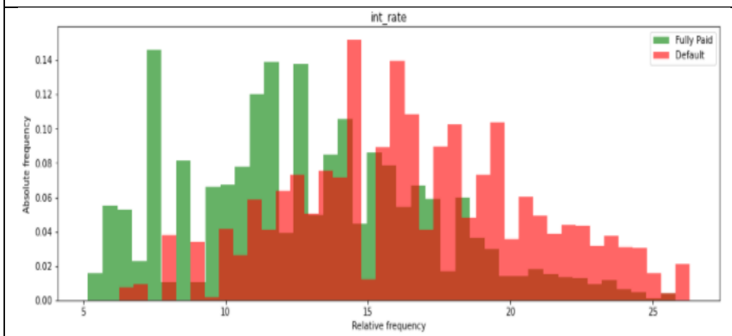
Loan default happen irrespective of employment length of customer.

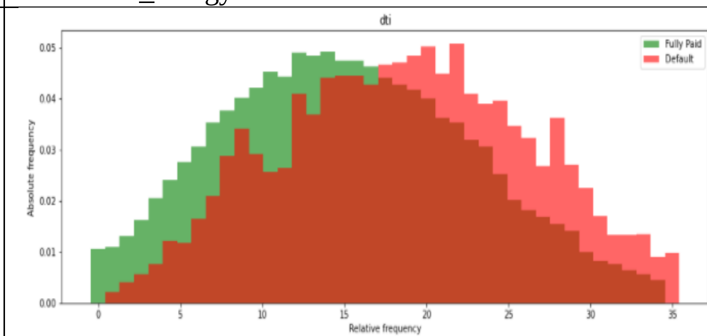
Table 9 Bivariate Analysis Categorical Features

Numerical Independent Vs Target

<table>
<tr>
<td>

Loan defaults happen irrespective of amount of loan applied.</td>
<td>

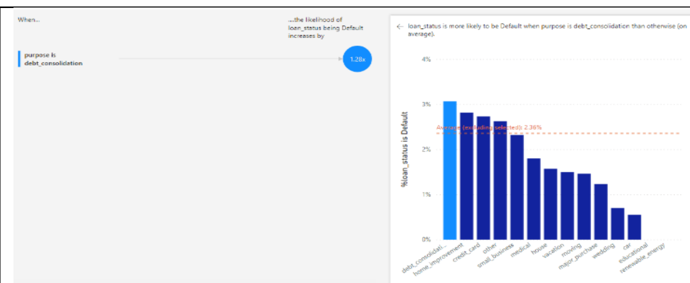
Default rate increases with interest rates irrespective of the purpose of the loan except education and renewable_energy</td>
</tr>
<tr>
<td>

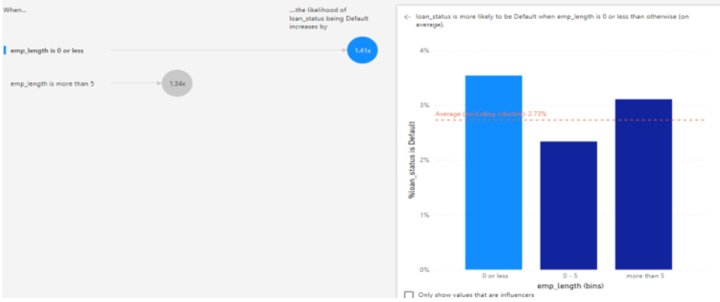
Maximum loan defaults happen for interest rates between 10 – 20 percent.</td>
<td>

Maximum loan defaults happen for dti ratios range 15 to 20.</td>
</tr>
</table>

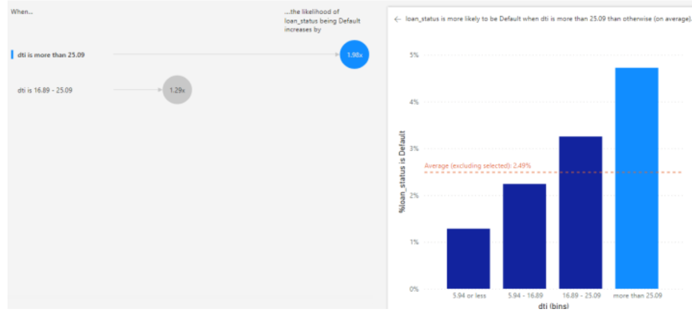
Table 10 Bivariate Analysis Numerical Features

Multivariate Analysis

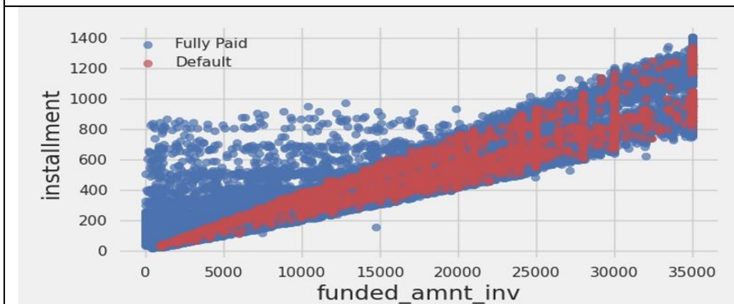
<table>
<tr>
<td>
When Loan amount is >26325, the default rate increases by 2.13x</td>
<td>
when purpose of loan is debt-consolidation, default rate increases by 1.28x</td>
</tr>
</table>



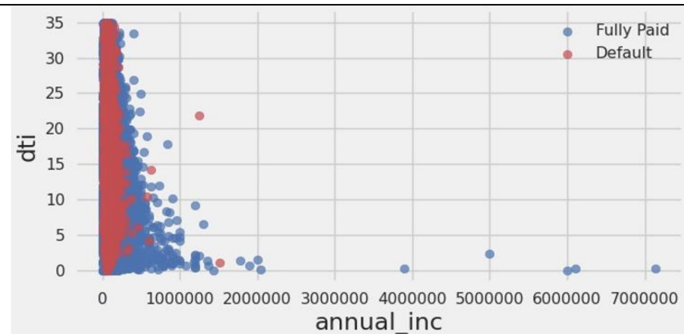
When employment length is less than 1 year the default increases by 1.41x.



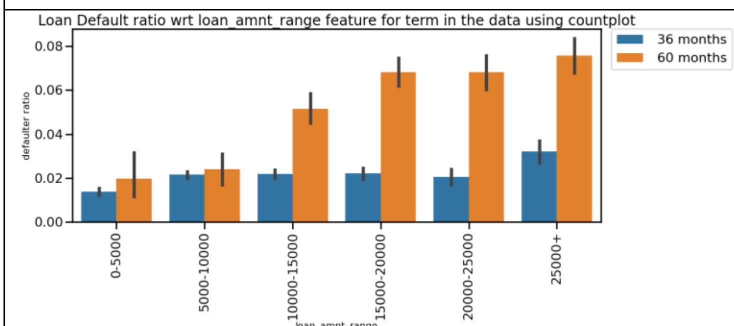
The likelihood of default increases by 1.98x when DTI >25.09



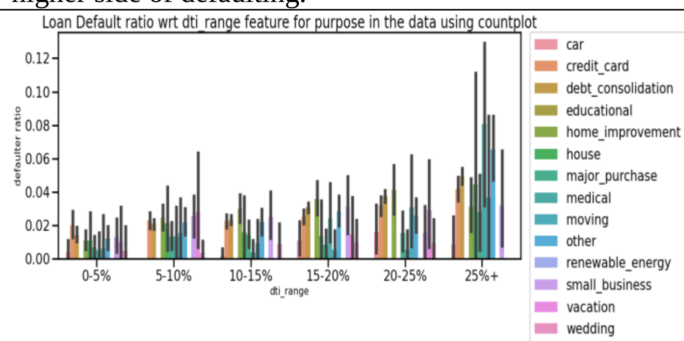
Clients with high installments and high funded_amnt_inv are on the higher side of defaulting.



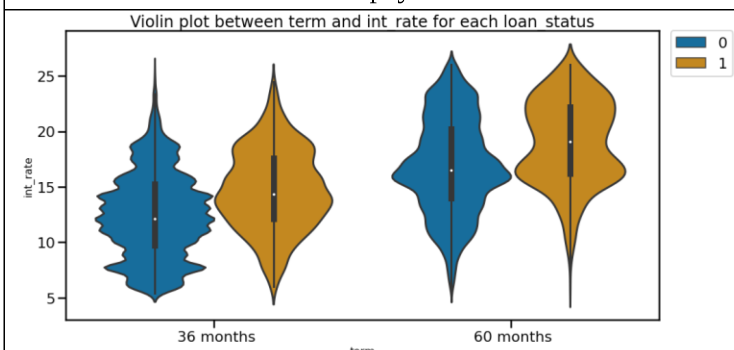
Clients with high dti and low annual_inc are on the higher side of defaulting.



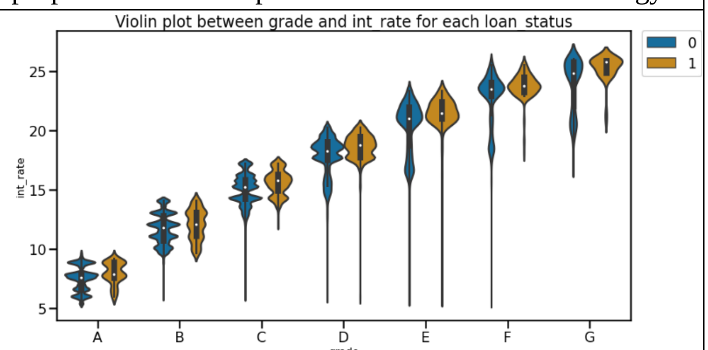
Higher number of defaults happened for the high loan amounts with 60 months of repayement term.



Higher defaults for higher DTI Range irrespective of purpose of loan except education and renewable energy.



int_rate increases with term on loan and the chances of default also increases



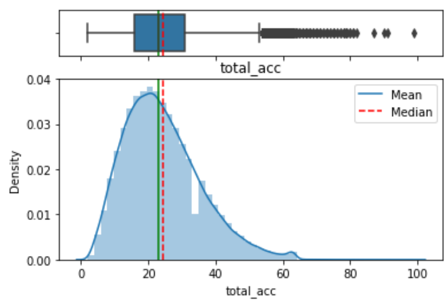
int_rate is increasing with every grade and also the defaulters for every grade are having their median near the non-defaulter.

Table 11 Multivariate Analysis

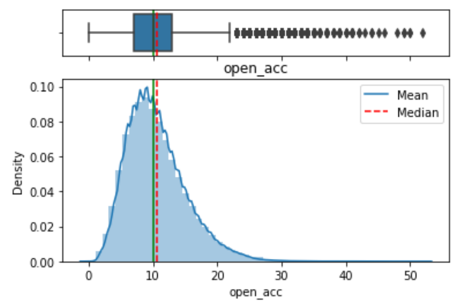
Outliers Detection

Box plot and Distribution plot were created for Numerical variable and high level of outliers were detected for features **total_acc, revol_bal, and open_acc**.

total_acc



open_acc



revol_bal

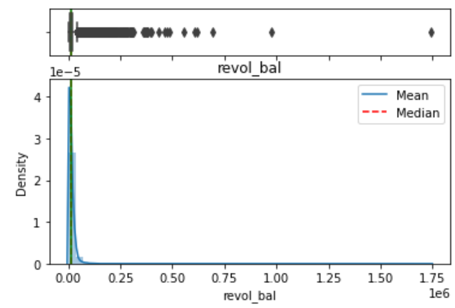


Table 12 Outliers Detection

Correlation Analysis

Heat map was created using **Pearson method** to visualize correlations amongst the features. Refer Appendix Exploratory Data Analysis Plots [5]

Highest correlated features are [**loan_amnt, funded_amnt, funded_amnt_inv, and instalment**]

Based on extensive EDA, following table enlists key inferences in various categories:

Category	High Default
Customer	● Grade C customers ● Customers with Debt consolidation loans ● Customers with mortgages ● Customer with employment length less than 1 year ● Customers with high debt to income ratio (dti) for dti>25.09% ● Customers with high annual income ● Customers with high number of credit lines ● Customers with high funded amount
Nature of Loan	● Loan with high installment ● Loan with longer tenure ● High Loan amount, especially those > $26,325
Macro economical influence	● Year 2014 has high % of loan defaults
Demographic influence	● California state has highest number of defaulters
Loan Disbursement status	● High loan disbursement to people with highest experience of 10 years, followed by 1 year and 2 years

Table 13 Insights from EDA

Model building

The dataset contains information about loan status of the applicants. This dataset will be used to build the various machine learning models and predict therein for deciding on from whom to accept or reject loan application.

After the Exploratory Data Analysis, Feature Engineering, Feature Selection and finally with Model Tuning, the patterns of fully paid or default applicants will be exposed by machine learning models.

Modelling Approach

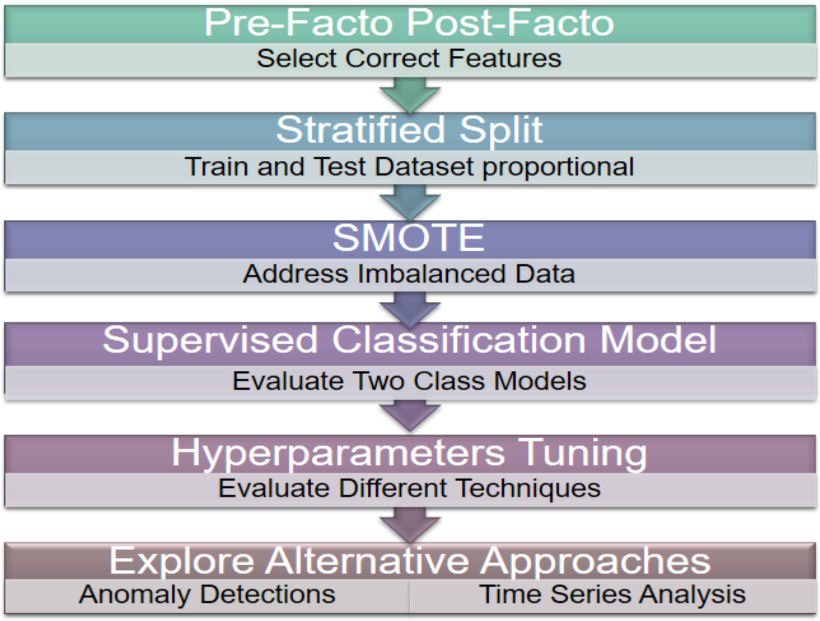


Figure 6 Modelling Approach

Pre-facto and Post-facto factors: The dataset contains broad features related to the applicant and their status of loan whether fully paid or default. For the purpose of model building and prediction, **we will be using pre-facto fields**.

Apply-Math Operation: This is performed on numerical data where **skewness** was observed. This helped to **remove skewness** and **normalize the data**. The technique followed for removing skewness was "**Logarithm PLUS1**" used [**Yeo-Johnson Algorithm**]. This is due to the presence of "zeros" in our dataset.

Stratified Split: Stratified split is a sampling technique where the samples are selected in the same proportion as they appear in the population. Generally, 70:30 is used for splitting data into train and test data. **Stratified splitting in cross-validation ensures the training and test sets have the same proportion of the dependent variable as in the original dataset.** By doing this with the target variable ensures that the cross-validation result is a close approximation.

SMOTE: Visualisation through Tableau and python programming showcased highly imbalanced dependent variable - Loan status the Fully Paid vs Default loan as 97% vs 3%. We used SMOTE to normalize the imbalance.

Classification Models: As we are dealing with a **Supervised Classification problem**, the goal is to train the best machine learning model to maximize the predictive capability, and to deeply understand the applicant's past profile for minimizing the risk of future loan repayments.

For building Classification Model, we will use Two-Class Classification algorithms. The reason for using two-class is due to our dependent variable has two-class values– Fully paid and Default. For the given dataset, following are the Two-class classification models that have been used.

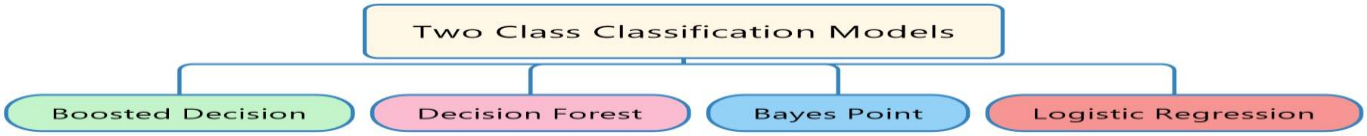


Figure 7 Two Class Classification Models

Choice of Model: The Two-Class Boosted Decision Tree is the best model in our study, because it has justifiable **Confusion Matrix and F1-score**.

Classification Models	Classification Report				Confusion Matrix			
	Accuracy	Precision	Recall	F1 Score	True Positive	False Positive	True negative	False Negative
Two-Class Boosted Decision	0.97	0.973	0.997	0.985	34610	961	24	149
Two-Class Decision Forest	0.968	0.975	0.993	0.984	34641	979	6	118
Two-Class Logistic Regression	0.972	0.972	1	0.986	34759	0	0	985
Two-Class Bayes Point	0.972	0.972	0.999	0.986	34740	984	1	19

Figure 8 Compare Models

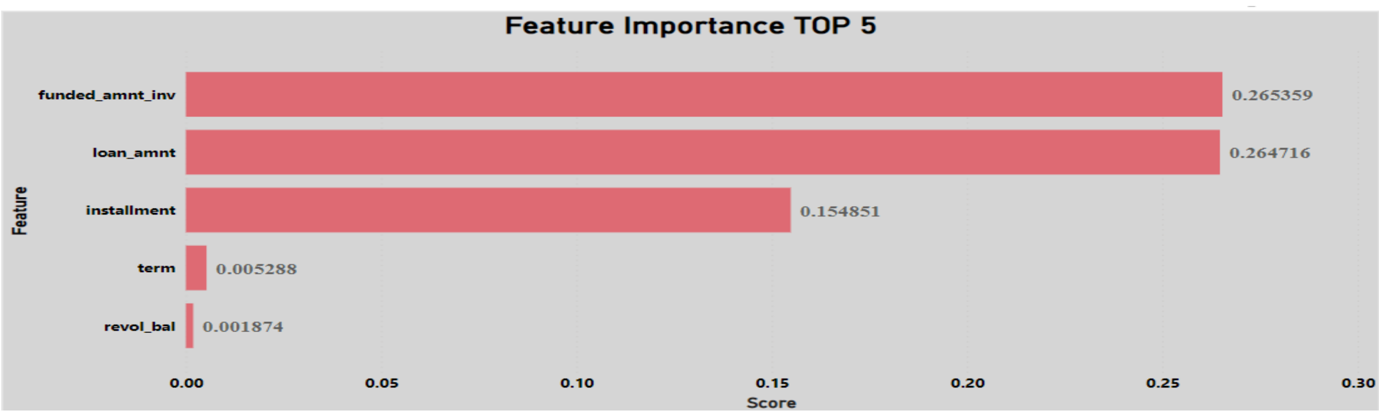


Figure 9 Feature Importance

Hyperparameters Tuning: A hyperparameter is a parameter whose value is set before training a machine learning or deep learning model. Different models require different Hyperparameters and some require none. Hyperparameters should not be confused with the parameters of the model because the parameters are estimated or learned from the data. Some keys points about the hyperparameters are:

- They are often used in processes to help estimate model parameters.
- They are often manually set.
- They are often tuned to tweak a model’s performance

Azure ML Studio Tune Model Hyperparameters function is used to tune the model. Refer Appendix Hyperparameter Tuning [7]

Stepwise manual execution of permutations between **Random Sweep and Random Grid** methods along with Evaluation Criteria were performed; and the resultant Confusion Matrix, and Classification Report were captured. These values were then compared with the Base model results, which showed that the base model is still the best choice, as Two Class Boosted Decision Tree by itself is a Tuned Ensemble model. Refer Appendix Hyperparameter Tuning [8]

Model Validation

The objective of model building is to predict future defaulters. So, we need a model which gives:

a) High Level of **Defaults prediction** from our test data **(True Negative)**

b) And Predicts **low** values of **Type 1 error (False Positive) and Type 2 error (False Negative)**

The other metrics that can be used for evaluation are the **FI Score** and **AUC**, as we are dealing with a highly **Imbalanced Data**.

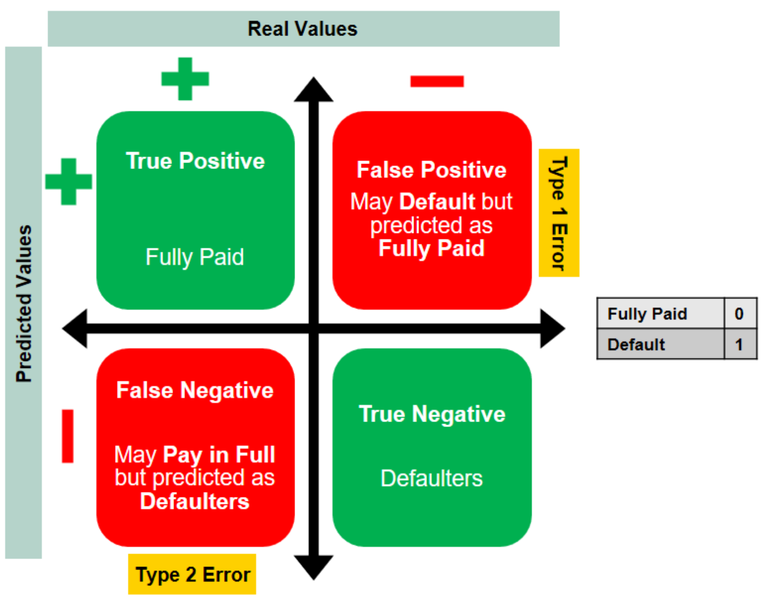


Figure 10 Confusion Matrix

Above figure clearly substantiate our business objectives here under:

- True Negative which is our default prediction should have justifiable values
- False Positive and False Negative needs to be of lower values

The evaluation model of Two-Class Boosted Decision Tree gives better result for **F1-score** and **Minority Classes** (False Positives and False Negatives), as this is **highly imbalanced data**, following is the ranking of applicable evaluation metrics in descending order.

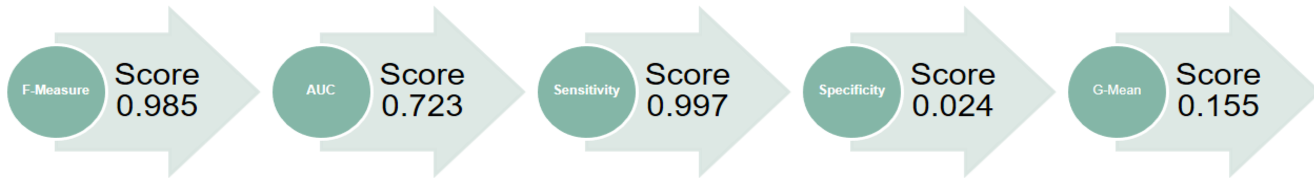


Figure 11 Imbalance Data Evaluation Metrics

Additional Machine Learning Models

Although the given problem statement requires a classification model machine learning approach, but when the given dataset was analysed further, it was discovered that other machine learning approaches that can also be applied for better understanding of problem areas and building a Heterogeneous Machine Learning model for prediction of defaults.

<table>
<tr><th>Machine Learning Approach</th><th>Remarks and Findings</th></tr>
<tr><td>Text Analytics</td><td>The “desc” feature in the provided dataset has loan description provided by the borrower. So an exploratory experiment was done to analyse [‘desc’, ‘loan_status’] relationship. Refer Appendix Azure ML Studio Snippets [3]
Results were inconclusive.</td></tr>
<tr><td>Association Rules Mining</td><td>Automated association mining techniques were explored using Association Rules Mining with PyCaret Association Rules library and Apriori Libraries. Refer appendix Python code [18] and [19] .Results were inconclusive.</td></tr>
<tr><td>Time series Analysis</td><td>Based on periodical data available for each month from 2008 to 2014, we can explore further for forecasting Loan Demands and Default Rates. We could find trend charts for Loan Demand and Default Rate within this time-period.

Demand Trend graph depicted that demand for loan Increased significantly during 2012 to 2014. Refer Appendix Time series [9]
Default Trend graph depicted that Default Rate increased manifolds over 2011 to 2014. Refer Appendix Time series [10]
Refer Appendix Python Code Reference [13]
</td></tr>
<tr><td>Anomaly Detection</td><td>
Anomaly Detection algorithm ‘iforest’ was applied on 70% sampling of original data.
The resultant output of the model Anomalies and Anomalies Score was captured in Data Frame and then downloaded in .csv file. Refer Appendix Anomaly Detection [11]
The 3D tSNE model of dataset anomalies was created to visualize the anomalies in the dataset. Refer Appendix Anomaly Detection [12]

<table>
<tr><td>Total Data</td><td>83402</td></tr>
<tr><td>No of Defaulters</td><td>2339</td></tr>
<tr><td>No of Anomalies in Defaulters</td><td>929</td></tr>
<tr><td>%of Anomalies</td><td>40%</td></tr>
<tr><td>Average Anomaly Score</td><td>0.019</td></tr>
</table>
Result: ~ 50% of defaulters can be found by cross checking the anomaly score on the Pre-Facto Data.</td></tr>
</table>

	Another important result- if the Average Anomaly Score is greater than or equal to 0.019, then there are high chances that the customer would default. **Hence, Anomaly Detection can be used as an instrument to predict defaulters during loan application approval process.** Refer Appendix Python Code [17]

Table 14 Additional Machine Learning Approaches

Insights and Recommendations

Insights

Based on the research and analysis on the consumer loans information, we arrived at the following insights for the bank:

- There is evidence of outliers in the data. Further anomaly analysis indicates that the high percentage of outliers are loan defaulters.
- Certain categories in the data have more defaults compared to other categories. These categories are:
    - Debt-consolidation in the 'Purpose of Loan' variable: Debt consolidation is basically the act in which multiple loans of a borrower are combined into a single large loan. The multiple loans are then paid off with the cash inflow from this new loan. These may be customers who are unable to manage their loans, hence decide for debt consolidation but still end up defaulting.
    - State of California in 'States' variable: State of California is among richer states in the U.S. People here would like to have big houses and live extravagant lifestyles. Sometimes the borrower may be in financial trouble while many times it would be unwillingness to repay debt.
    - Debt to Income ratio ranging between 15 and 20: A higher dti would mean the borrower has low annual income.
    - Borrowers paying high installments: These are customers who have been charged higher interest based on higher risk.
    - Borrowers with a loan term of 60 months reflect more defaults.
    - Other categories which are also prone to default are customers with multiple credit lines, customers with mortgages, customers with employee length of less than one year, customers with 'Grade C' credit score
- Further we have also identified that certain categories of loans, such as education loans and renewable energy, are perceived to be risk-free i.e., have no defaults in the data.
- There are four US States - Idaho, Iowa, Mississippi, and Maine without any defaults.
- Anomaly analysis shows ~ 50% of defaulters can be found by cross checking the anomaly score on the Pre-Facto Data. Another important result - if the Average Score is greater than or equal to 0.019, then there are high chances that the customer would default.

Recommendations

- Our model indicates a good prediction of defaulters. Hence, banks should use our model based predictions to undertake proper risk benefit evaluation for customers predicted to default by our model. As discussed above, there are higher defaulters **in Debt-consolidation category**, **the State of California**, **dti ranging between 15 and 20**; so proper risk benefit evaluation has to be done of customers falling in these categories. Even customers who have **property on mortgage** or **multiple credit lines**, or have **high installments** are high risk customers who are shown to default more based on the given dataset.

- **Curated Risk Assessment -** Regionally tailored risk assessment and policies could potentially achieve more accurate default forecasts and reduce the inefficient allocation of resources to uncreditworthy borrowers. Risk assessment procedures could largely benefit from the application of ML methods. As we are aware, the State of California has more defaults. Banks could decide whether to grant a loan more conscious of the risk associated to each borrower type. Bank can also capture additional information from the borrower for State of California to understand the borrower's credit profile better.
- For loan amount above 25000+, term preference should be 36 months to reduce default risk, as evident in the analysis. [Refer Table 11 Multivariate Analysis]
- Banks can constantly monitor Grade C, which has shown more defaults despite not being a worst grade or look to review and improvise the grading system.
- If the bank decides to lend the borrower which has been predicted to default, it can ask the borrower for collateral or guarantee or both, so even if there is a default the bank has avenues to recover money and the negative impact on profitability would be low or insignificant.
- Banks can pursue growth opportunities in education loan and renewable energy areas. These loans are safer relative to others (debt consolidation), so banks can expand into those areas. These loans would be less risky and would improve profitability of the bank.
- Another growth opportunity for the banks would be exploring loan portfolios in the four states - Idaho, Iowa, Mississippi, and Maine, which have no defaults.
- We also recommend that the banks provide good, reliable data with lower missing values to build a more robust prediction model.
- Bank can also provide us with additional data point such as age, gender, number of dependents, heuristics which can help build a better model.
- Heuristics must be a part of risk profile because most often borrower's inability to pay due to income shocks, such as a job loss, or due to an adverse change in economic conditions directly affects the loan payment.

To summarise, a better understanding of the default behaviour and of the regional differences in these credit markets with the help of default prediction model could help policy makers to undertake more effective risk-mitigating actions. Based on the data the features 'funded_amnt_inv', 'loan_amnt', and 'installment' are considered as very critical by the model while predicting defaults, hence would have the most impact on the sanction of loan. Alternatively, default prediction can be validated using Anomaly detection technique.

Bibliography

Python Libraries References

[1] PyCaret Classification: Link
[2] PyCaret Anomaly Detection: Link
[3] Matplotlib Documentation: Link
[4] Seaborn Documentation: Link
[5] Plotly Documentation: Link
[6] SweetViz Documentation: Link
[7] AutoViz Documentation: Link
[8] Pandas Documentation: Link
[9] Numpy Documentation: Link
[10] PyCaret Association Rules: Link
[11] Apriori Association Rules: Link

Machine Learning References

[12] Azure ML Hyperparameter tuning: Link
[13] Feature Engineering: Link
[14] Machine Learning Glossary: Link
[15] Reference Case Study A: Link
[16] Reference Case Study B: Link

Appendix

Python Code Snippets

[1]

One-Hot Encoding Feature Engineering

```
[ ]  Exp_Df = clean_df.copy()
     Exp_Df.loc[(Exp_Df.loan_status == 'Fully Paid'),'loan_status']=0
     Exp_Df.loc[(Exp_Df.loan_status == 'Default'),'loan_status']=1
     Exp_Df['loan_status']= pd.to_numeric(Exp_Df['loan_status'])
     Exp_Df.info()
```

Figure 12 One-Hot Encoding

[2]

Numerical Binning Feature Engineering

```
# create bins for loan_amnt range
bins = [0, 5000, 10000, 15000, 20000, 25000, 36000]
bucket_l = ['0-5000', '5000-10000', '10000-15000', '15000-20000', '20000-25000','25000+']
EDA_Df['loan_amnt_range'] = pd.cut(EDA_Df['loan_amnt'], bins, labels=bucket_l)

# create bins for int_rate range
bins = [0, 7.5, 10, 12.5, 15, 100]
bucket_l = ['0-7.5', '7.5-10', '10-12.5', '12.5-15', '15+']
EDA_Df['int_rate_range'] = pd.cut(EDA_Df['int_rate'], bins, labels=bucket_l)
# create bins for annual_inc range
bins = [0, 25000, 50000, 75000, 100000, 1000000]
bucket_l = ['0-25000', '25000-50000', '50000-75000', '75000-100000', '100000+']
EDA_Df['annual_inc_range'] = pd.cut(EDA_Df['annual_inc'], bins, labels=bucket_l)
# create bins for installment range
def installment(n):
    if n <= 200:
        return 'low'
    elif n > 200 and n <=500:
        return 'medium'
    elif n > 500 and n <=800:
        return 'high'
    else:
        return 'very high'

EDA_Df['installment'] = EDA_Df['installment'].apply(lambda x: installment(x))
# create bins for dti range
bins = [-1, 5.00, 10.00, 15.00, 20.00, 25.00, 50.00]
bucket_l = ['0-5%', '5-10%', '10-15%', '15-20%', '20-25%', '25%+']
EDA_Df['dti_range'] = pd.cut(EDA_Df['dti'], bins, labels=bucket_l)
```

Figure 13 Numerical Binning

Azure ML Studio

[3]

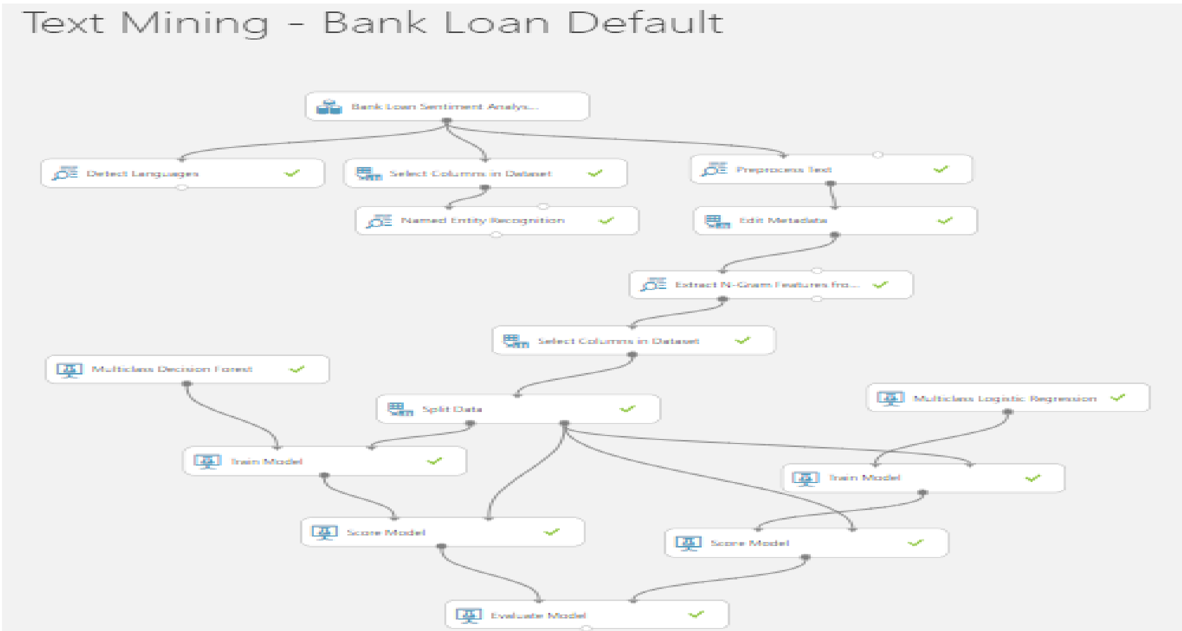


Figure 14 Text Analytics

[4]

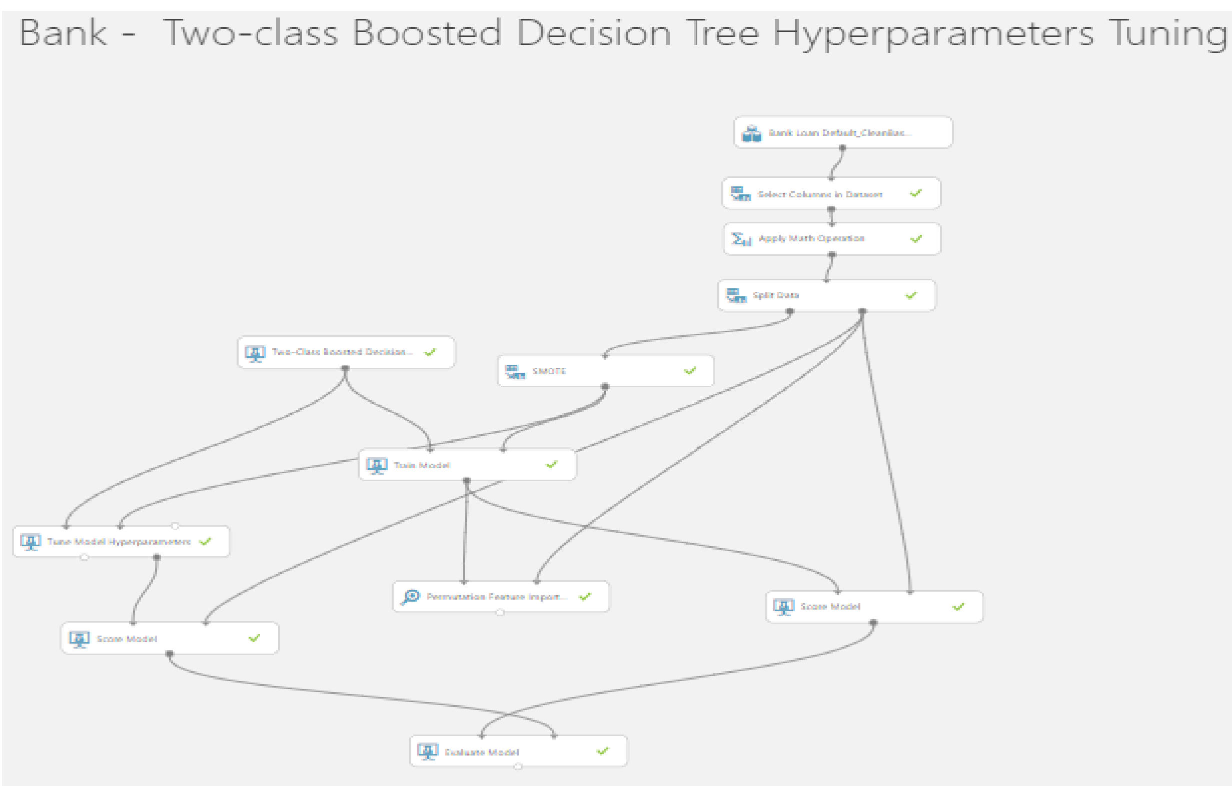


Figure 15 Hyperparameters Tuning

Exploratory Data Analysis Plots

[5]

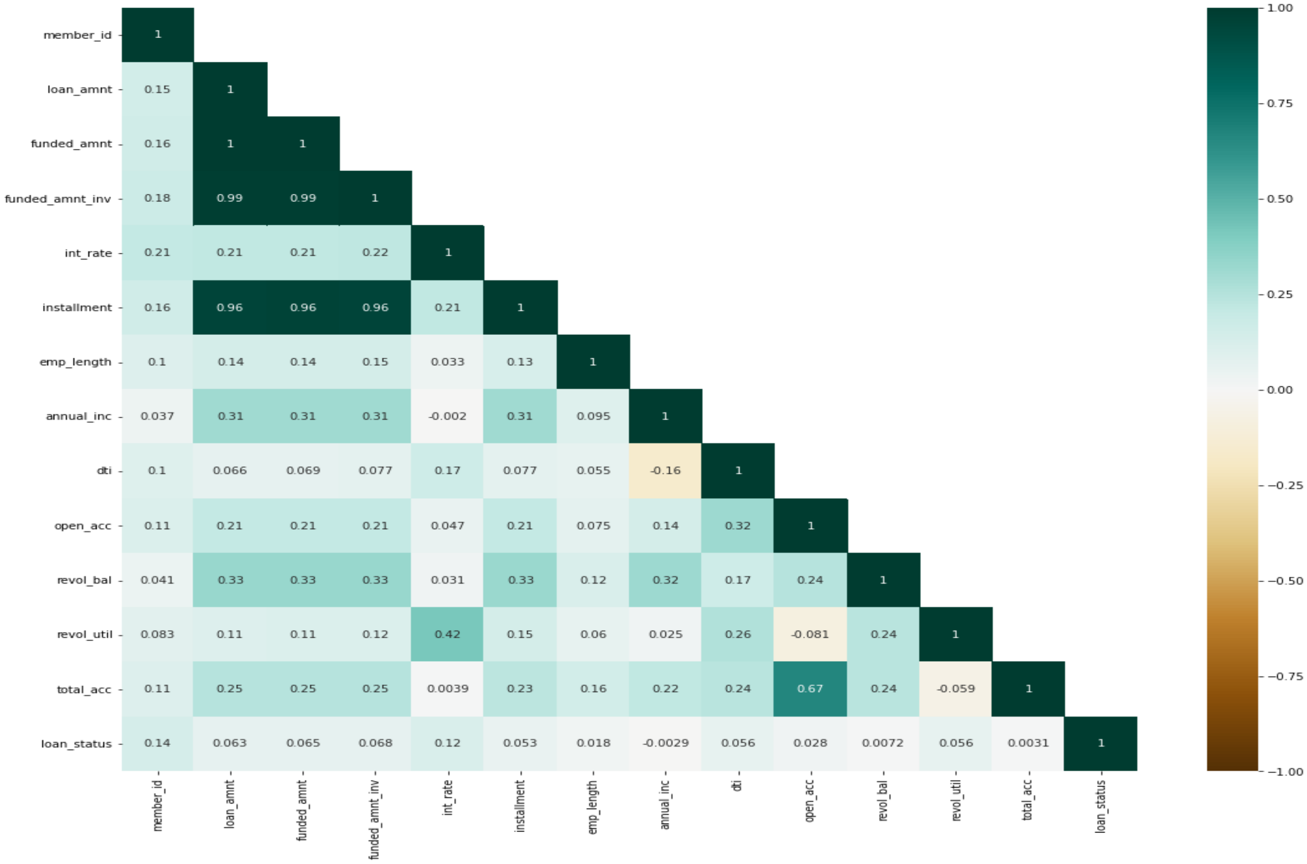


Figure 16 Pearson’s Heatmap

[6]

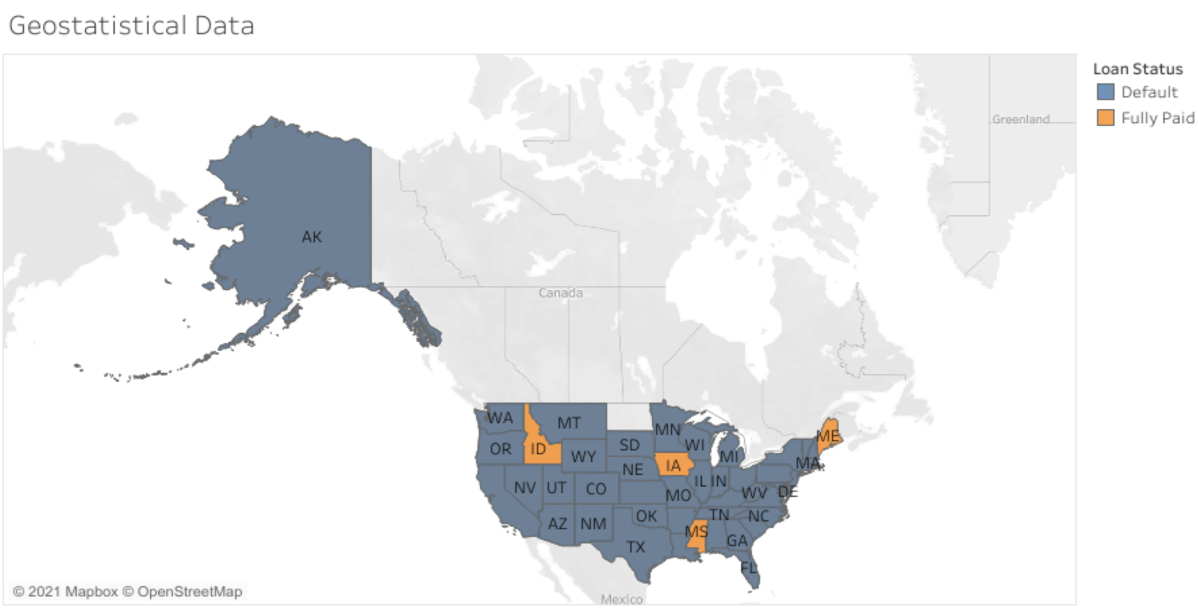


Figure 17 US States with No Defaults

Hyperparameters Tuning

[7]

Evaluation Matrix for Two Class Boosted Decision Tree												
	Accuracy	Precision	Recall	F1 Score	AUC	TP	TN	FP	FN	TPR	TNR	G-Mean
Base Model	0.969	0.973	0.996	0.984	0.723	34610	24	961	149	0.995713	0.024365	0.155759545
Hyper parameter Tuning												
Random Sweep; Accuracy	0.971	0.973	0.998	0.985	0.726	34696	15	970	63	0.998188	0.015228	0.123291627
Random Sweep; F-score	0.971	0.973	0.998	0.985	0.726	34696	15	970	63	0.998188	0.015228	0.123291627
Random Sweep; Precision	0.965	0.973	0.992	0.982	0.72	34465	41	944	294	0.991542	0.041624	0.203155843
Random Sweep; Recall	0.972	0.972	0.999	0.986	0.716	34734	1	984	25	0.999281	0.001015	0.031851189
Random Sweep; AUC	0.965	0.973	0.992	0.982	0.72	34465	41	944	294	0.991542	0.041624	0.203155843
Random Sweep; Average Log Loss	0.971	0.973	0.998	0.985	0.726	34696	15	970	63	0.998188	0.015228	0.123291627
Random Sweep; Train Log Loss	0.972	0.972	1	0.986	0.712	34759	0	985	0	1	0	0
Random Grid; Accuracy	0.964	0.973	0.99	0.982	0.727	34423	44	941	336	0.990333	0.04467	0.210328897
Random Grid; F-score	0.964	0.973	0.99	0.982	0.727	34423	44	941	336	0.990333	0.04467	0.210328897
Random Grid; Precision	0.964	0.973	0.99	0.982	0.727	34423	44	941	336	0.990333	0.04467	0.210328897
Random Grid; Recall	0.972	0.972	1	0.986	0.718	34751	0	985	8	0.99977	0	0
Random Grid; AUC	0.964	0.973	0.99	0.982	0.727	34423	44	941	336	0.990333	0.04467	0.210328897
Random Grid; Average Log Loss	0.964	0.973	0.99	0.982	0.727	34423	44	941	336	0.990333	0.04467	0.210328897
Random Grid; Train Log Loss	0.972	0.972	1	0.986	0.71	34759	0	985	0	1	0	0

Figure 18 Hyperparameters Tuning Compare Sheet

Best <--->Worst

[8]

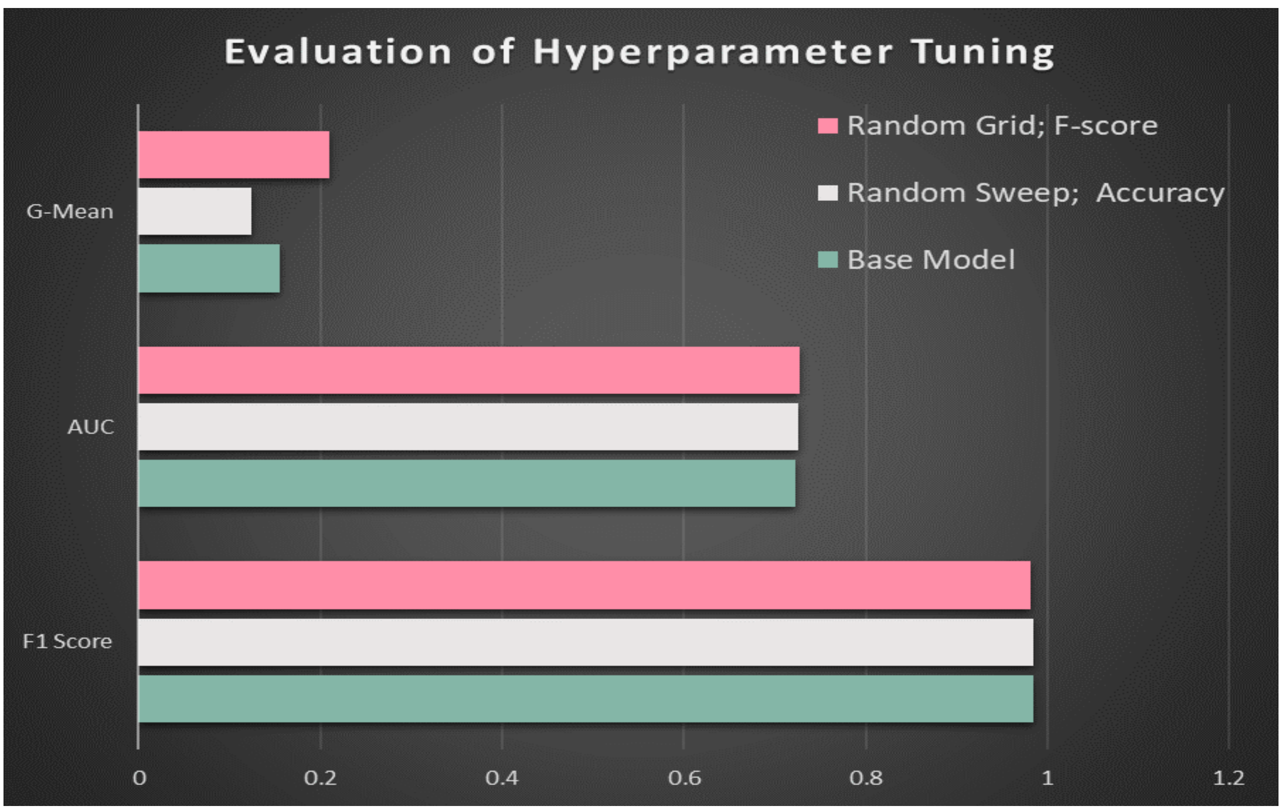


Figure 19 Hyperparameters Tuning comparisons with Base Model

Time-Series Analysis

[9]

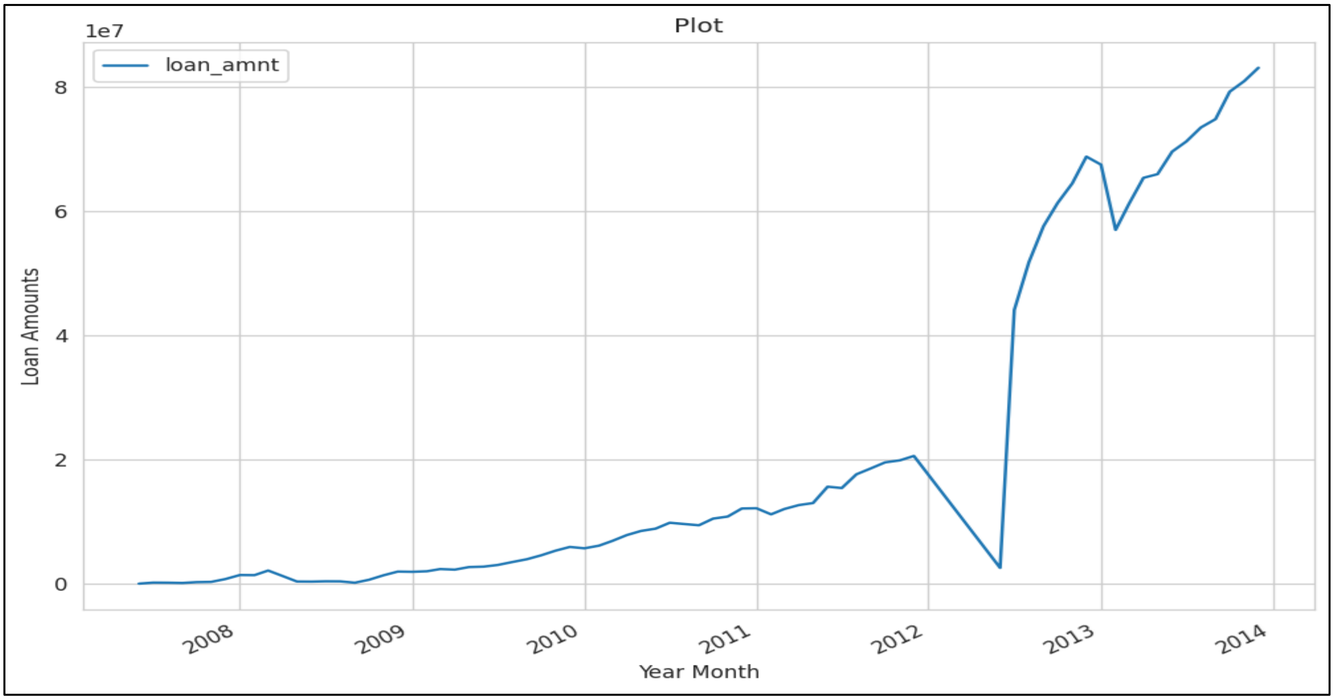


Figure 20 Loan Demand Trend

[10]

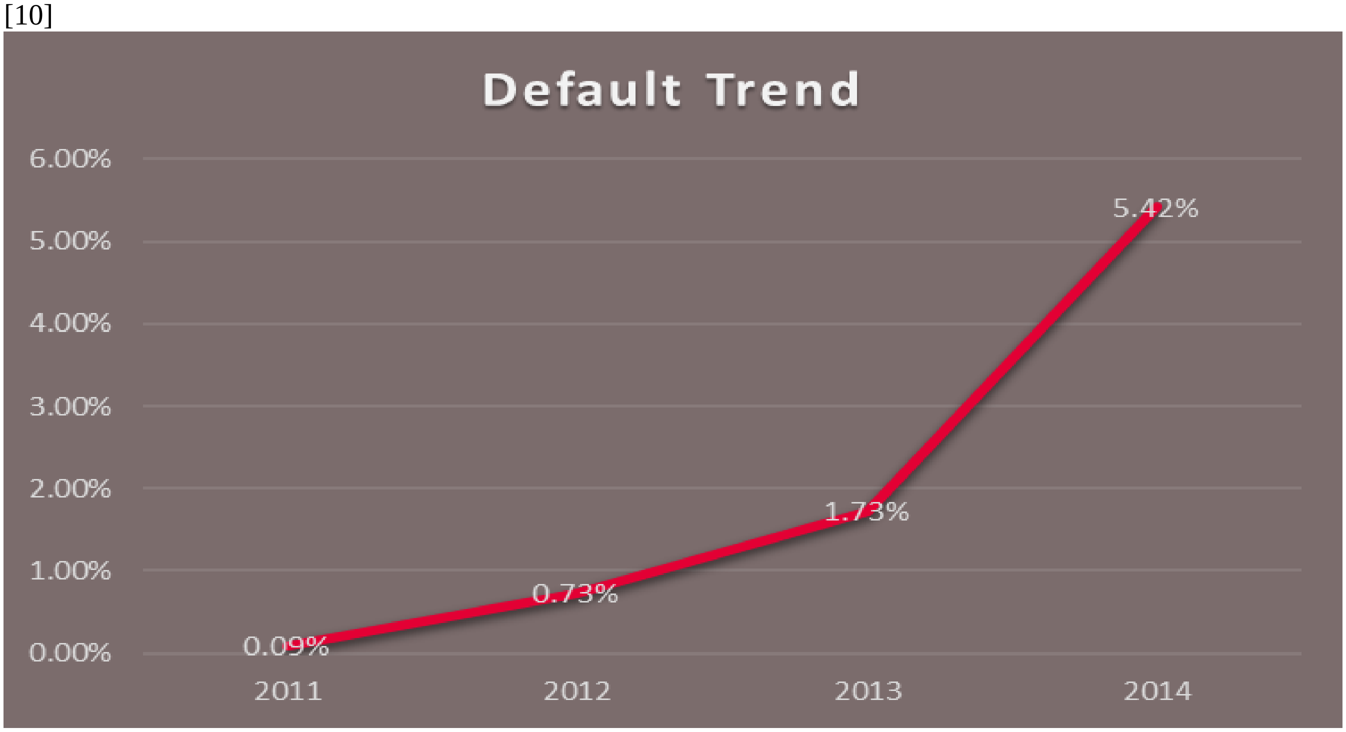


Figure 21 Loan Default Trend

Anomaly Detection

[11]

Assign Model and Capture the Results of Anomaly Analysis

```
iforest_results = assign_model(iforestmodel)
iforest_results.head()
```

inded_amnt_inv	term	int_rate	installment	grade	emp_length	home_ownership	annual_inc	purpose	addr_state	dti	open_acc	revol_bal	revol_util	total_acc	loan_status	Anomaly	Anomaly_Score
15975.0	36 months	12.12	532.35	B	9	OWN	40000.0	credit_card	NC	30.60	21	17324	56.4	29	Fully Paid	0	-0.022195
12000.0	36 months	13.61	407.87	C	1	RENT	52000.0	credit_card	WI	10.32	11	16733	45.6	20	Fully Paid	0	-0.053363
10000.0	36 months	11.99	332.10	B	0	MORTGAGE	74563.0	credit_card	IL	19.11	11	6688	53.5	21	Fully Paid	0	-0.054333
14400.0	36 months	7.66	448.99	A	1	MORTGAGE	82000.0	home_improvement	FL	13.10	13	12845	36.2	18	Fully Paid	0	-0.052986
18250.0	36 months	18.25	662.08	D	4	RENT	65883.0	debt_consolidation	WA	8.54	8	18112	93.4	10	Fully Paid	0	-0.038387

Figure 22 Anomaly Analysis Report using “iForest” Algorithm

[12]

Plot the tSNE model of Anomales in 3D

```
plot_model(iforestmodel)
```

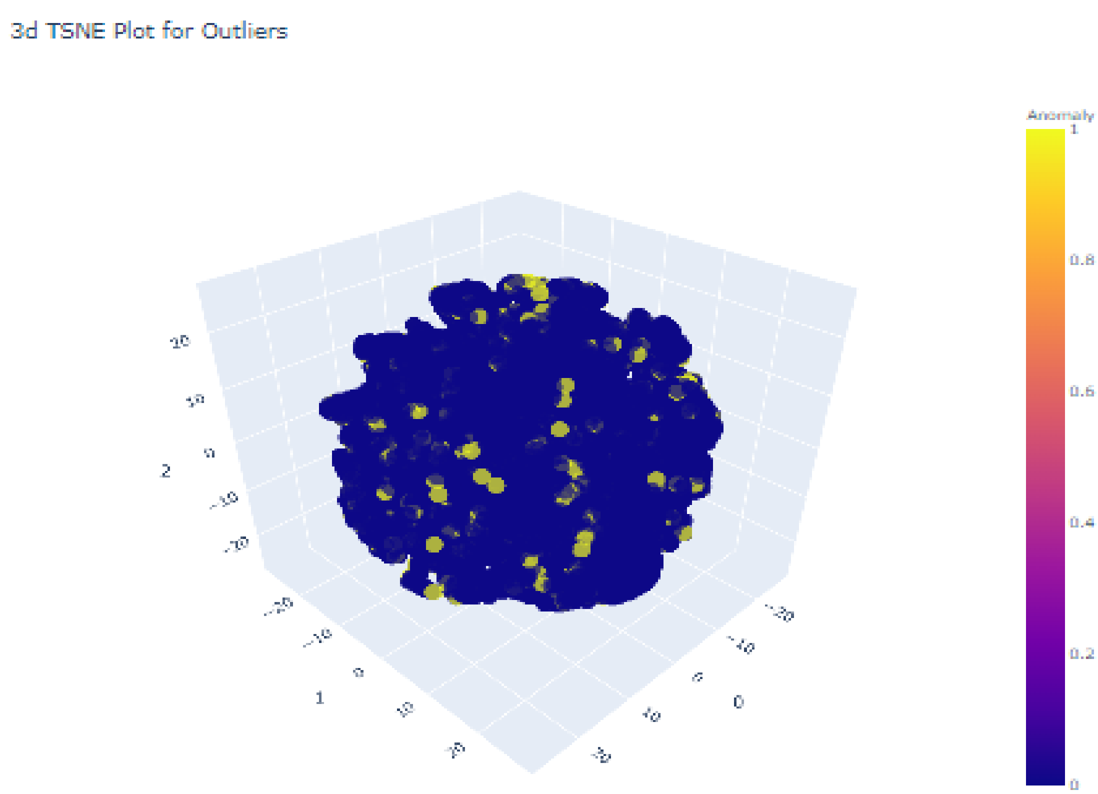


Figure 23 3D tSNE graph

Python Code Google Colaboratory Notebooks Links

[13] Trend Analysis from Dataset: Link
[14] Feature Engineering and Exploratory Data Analysis: Link
[15] Auto EDA using SweetViz and AutoViz: Link
[16] Auto ML using PyCaret Classification: Link
[17] Anomaly Detection using PyCaret Anomaly Detection: Link
[18] Association Rules Mining using PyCaret Association Rules: Link
[19] Association Rules Mining using Apriori: Link

EDA Graphs Links

[20] SweetViz Output: Link
[21] AutoViz Output: Link

Tableau Public Links

[22] Bank Default EDA Dashboard: Link

Azure ML Studio Links

[23] Bank Default Machine Learning Modelling: Link
[24] Bank Default Text Analytics / Sentiment Analysis: Link
[25] Bank Default Machine Learning Hyper parameter Tuning: Link

««End of Report»»