Interim Project Report
Bank Loan Default

Submitted By:

Balajisriram Venkateshkumar
Bina Rajput
Narendra Kr. Gupta
Rakendu Sharma
Shailajha Deepak

Under the mentorship of:
Harshal Jawale

1. Introduction

The goal of this study is to build a machine learning model that can predict if a customer will default on the loan based on the preliminary information provided by the customer while applying for the loan. The model is envisioned to be used as a decision-making tool for the financial institution to help make decisions on sanctioning loans, so that the risks can be lowered, and the profits can be maximized.

2. Problem Statement, Scope and Objective

Financial institutions play an important role in the growth of an economy. Since the global financial crisis, risk management has become critical in the bank's decision-making process. **One of the major risk management decisions that the banks face is whether or not to lend money to a borrower owing to default risk.**

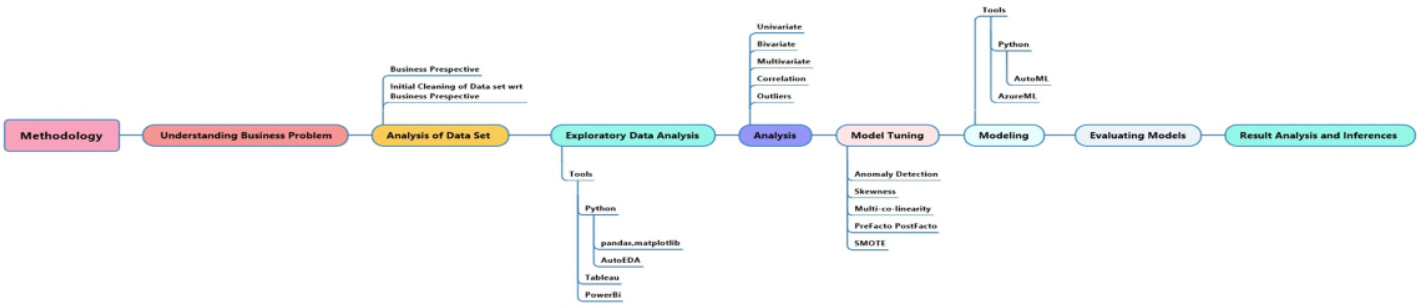
PROBLEM STATEMENT: Default risk is the risk that the borrower i.e., companies or individuals would not be able to make the payments on their debt obligations. **Loan default is a major concern faced by several banks and financial institutions since it directly impacts the bottom line of the banks and financial institutions**. 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 shareholders. Higher defaults across the banking system can also have an impact on the growth of the economy. Hence, default prediction is critical for a bank or financial institution. **Default prediction allows us to determine whether the borrower would default in the debt repayment.**

The **OBJECTIVE** of the study is below:

- To understand the behavioural pattern of customers before granting a loan
- To predict whether or not a borrower would default
- To help the bank / financial institution reduce the default rate and improve profitability

As a **SCOPE** of project, 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.

Methodology of the study



3. Data Description

Understanding the Dataset:

There are 41 columns, 119145 rows having data types - integer, float, string, and datetime. 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

4. 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.

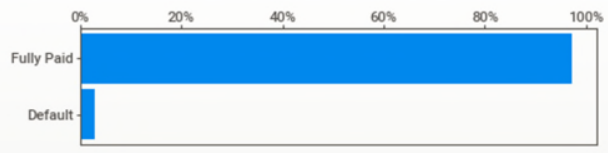
#	Fields	Description	Remarks / Observations and Data Cleaning
1	member_id	A unique Id for the borrower member.	This will not be useful for prediction model, but kept for pivoting.
9	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	Converted the data to numerical instead of string. Removed words like 'Years' and '+' Also replaced n/a with 0.
15	desc	Loan description provided by the borrower	This was separated to do text analysis.
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.
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.

5. Exploratory Data Analysis (EDA)

An EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for data analytics because it exposes trends, patterns, and relationships that are not readily apparent. **We have done Univariate, Bivariate and Multivariate analysis to understand the correlation with target variable.**

Univariate Analysis

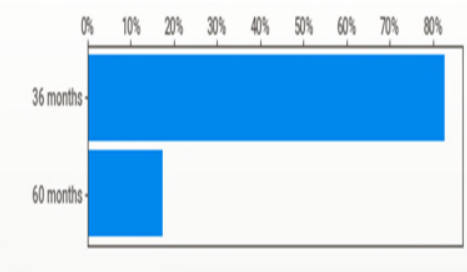
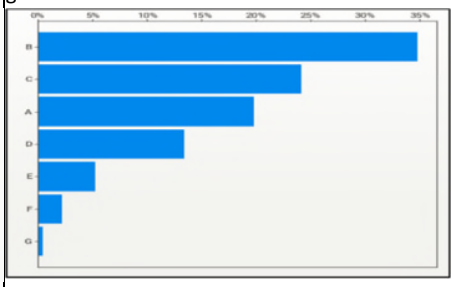
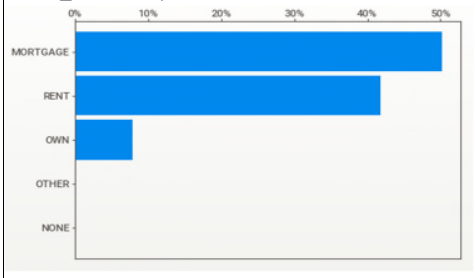
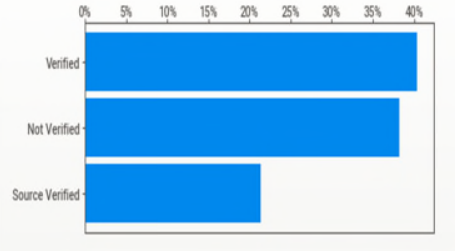
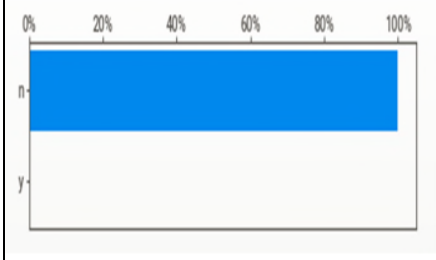
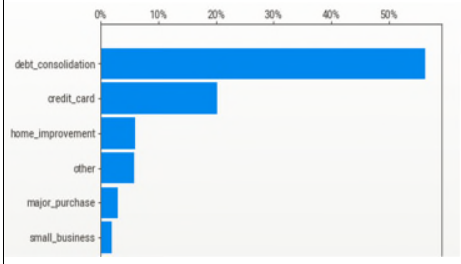
Target Variable - loan_status

loan_status 	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.

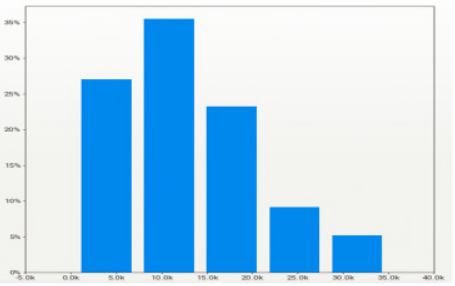
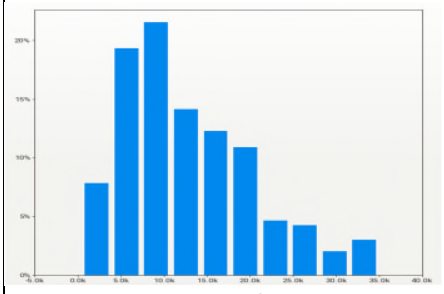
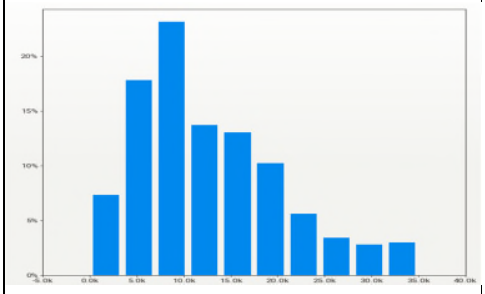
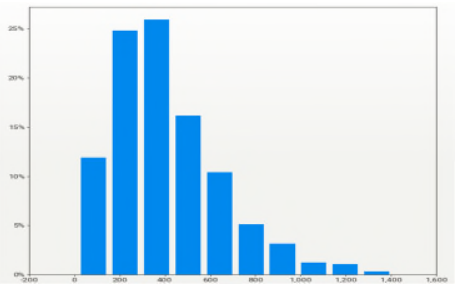
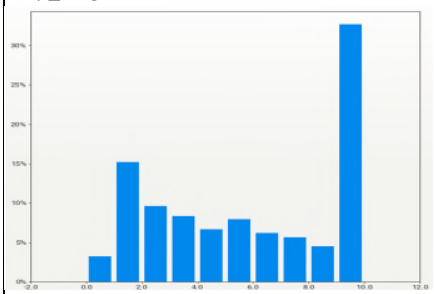
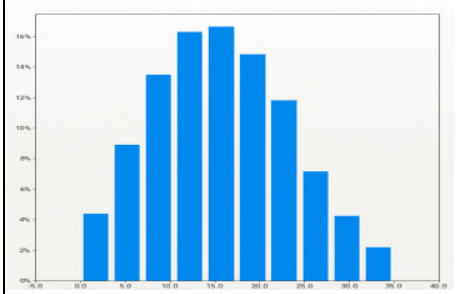
Predictor Variables

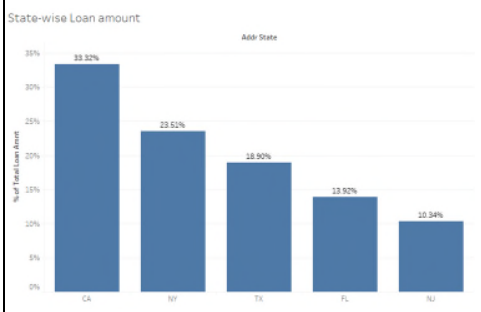
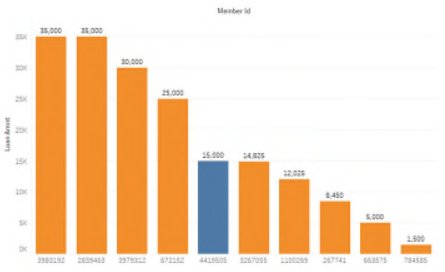
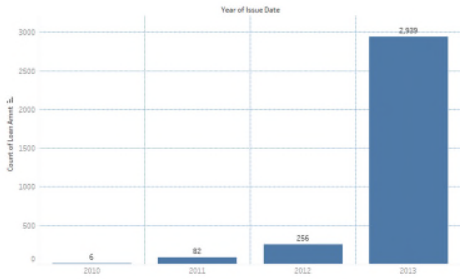
There are 3 types of variables amongst 41 parameters: 11 categorical, 25 numerical, 5 Date Time

Key Categorical Features

term  The loan term is either 36 months or 60 months.	grade  Customer with Grade B (35%) have taken most loans followed by C (24%) and A (20%) least being G (<1%)	home_ownership  Highest percentage of loans are taken by customers who have Mortgage (50%) followed by Rent (42%)
verification_status  Percentage of Verified customers is 40%, and Not Verified are 38%	pymnt_plan  For more than 99% of customers there is no payment plan.	Purpose  The top most purpose of loan is debt_consolidation 56% followed by credit_card 20%

Key Numerical Features

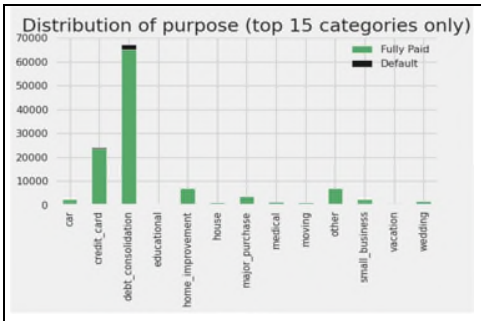
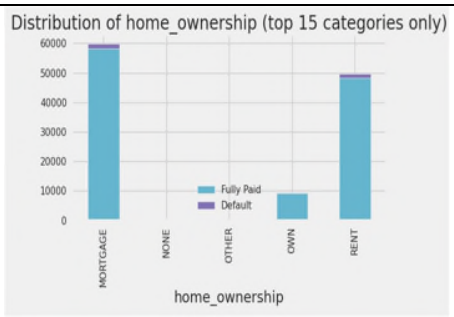
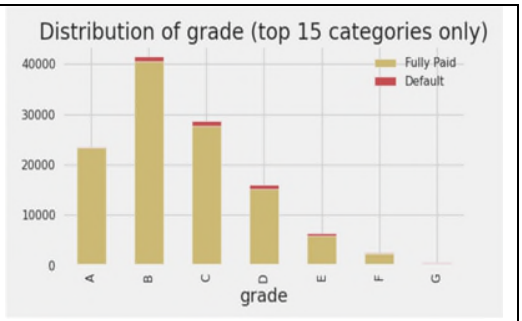
loan_amnt  Most frequent loan amounts $10000 (7.5%) and $12000 (5.9%) - **skewed data**	funded_amnt  Maximum loan amount is $35,000 whereas minimum loan amount in $500 - **skewed data**	int_rate  Maximum Interest rate is 26.1% while minimum Interest rate is 5.4% - **skewed data**
Instalment  Maximum Instalment paid by customers is $1407 while minimum being $16 - **skewed data**	emp_length  Most loans are disbursed to customers having 10 or 10+ years of experience followed by 1 year and 2 years - **skewed data**	dti  The maximum value of DTI for customers is 34.99 and least being 0.0

Address State  CA,NY,TX,FL & NJ are the top 5 states, having default	Member id  Member ID '4419505' is the Only defaulter among the top 10 Borrowers	Issue date  Loan defaulters increased exponentially in 2013 (2939 nos)

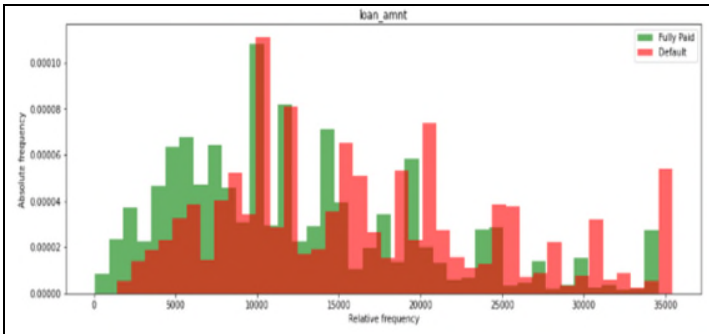
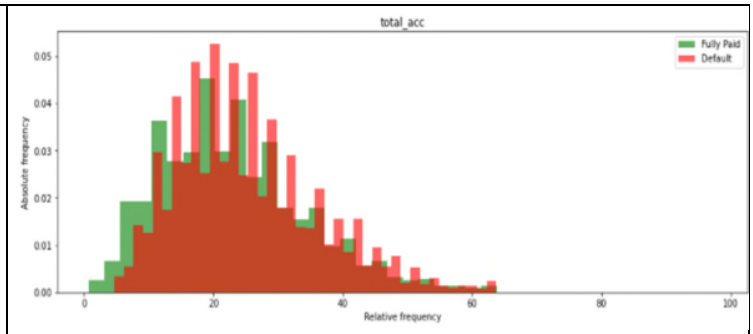
Bivariate Analysis

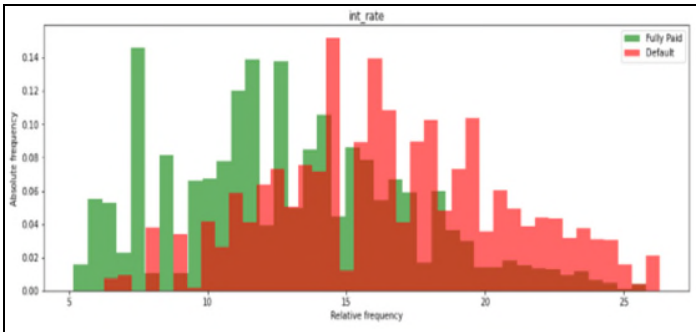
Bivariate analysis helps finding empirical relationships between two variables. Specifically, the Dependent vs Independent Variables.

Categorical Independent Vs Target

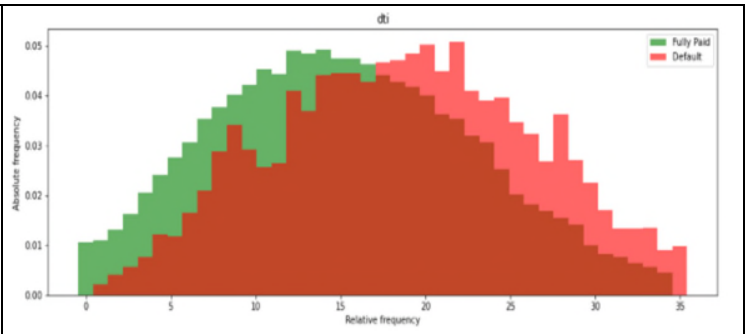
 Highest number of default happen when purpose of loan is 'debt_consolidation'.	 Highest number of defaulters have mortgaged their property or live on rent.	 Defaulters are from Grade B,C and D, while A, F, and G grades have none. Maximum defaulters are in Grade C, followed by Grade B

Numerical Independent Vs Target

 Loan defaults happen irrespective of amount of loan applied.	 Higher the number of credit lines, higher the defaults



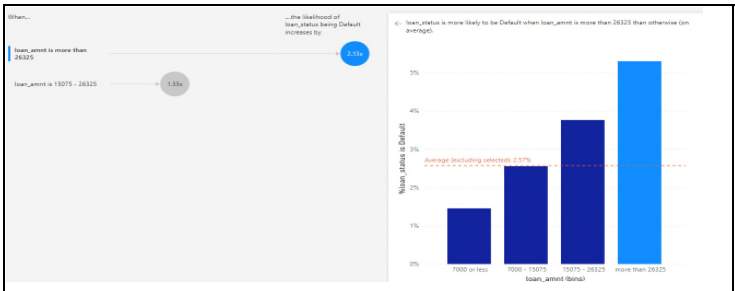
Maximum loan defaults happen for interest rates between 10 – 20 percent.



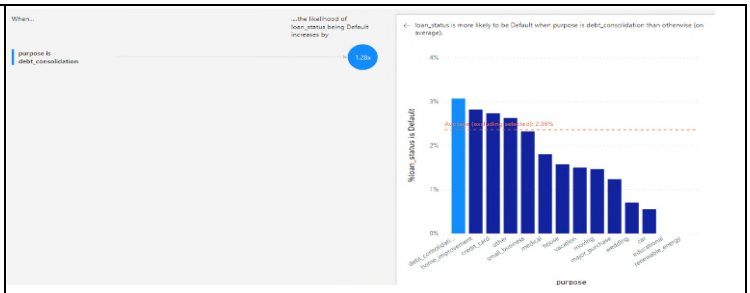
Maximum loan defaults happen for dti ratios between 15–20.

Multivariate Analysis

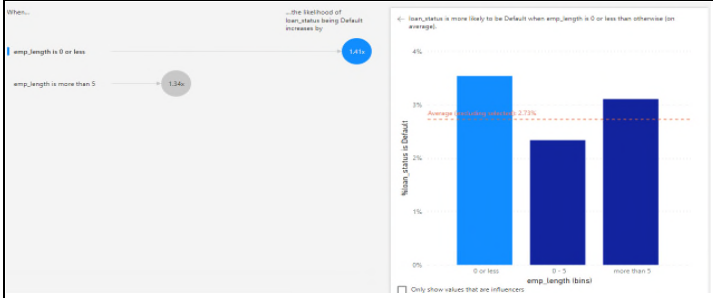
Multivariate analysis helps finding empirical relationships between more than two variables. Specifically, the One Dependent vs Two Independent Variables.



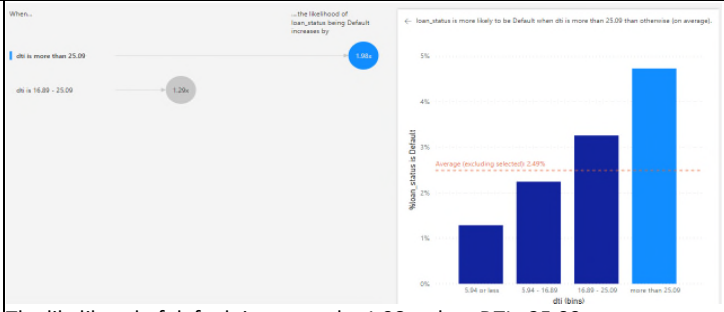
When Loan amount is >26325, the default rate increases by 2.13x



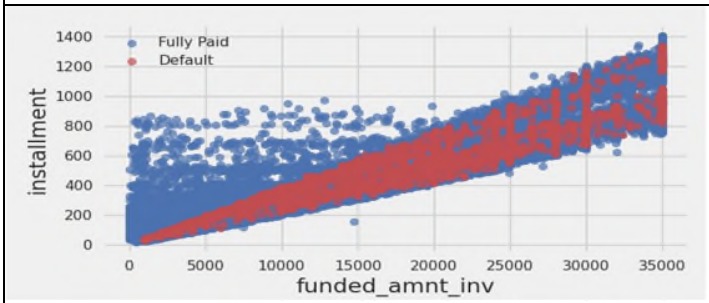
When purpose of loan is debt-consolidation, default rate increases by 1.28x



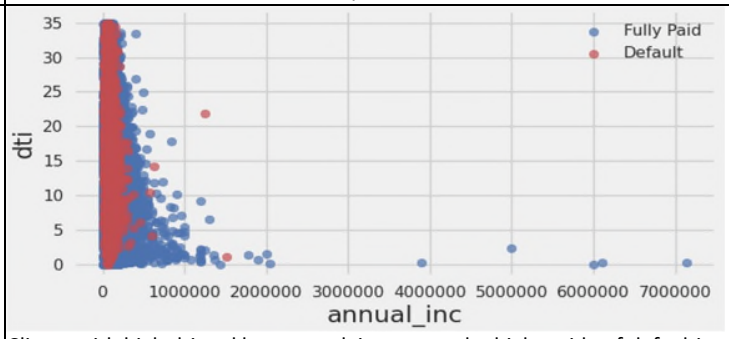
When employment length is less than 0 the default increases by 1.41x.



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

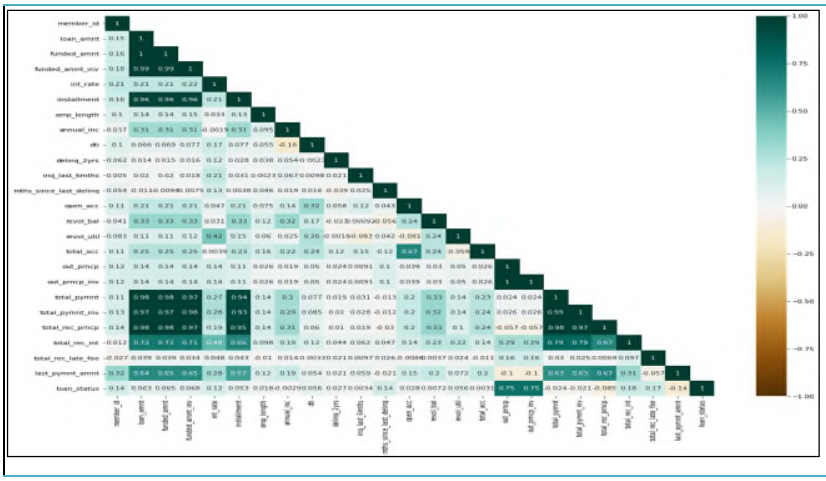


Clients with high installments and high funded_amnt_inv are on the higher side of defaulting.



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

Visualizing Correlation/Associations via Heat map



By plotting a heat map — **SPEARMAN METHOD** — it is easy to understand that the variables

a) 'out_prncp' and 'out_prncp_inv' are the two most correlated features with (75%) loan_status.

b) Also the variables set [total_payment, total_payment_inv, total_rec_prncp total_rec_int], and [funded_amount, funded_amnt_inv, installment] show high levels of correlation.

Key Inferences from the EDA

Dependant variable i.e., 'Loan status' having highest default points is tabled below

Category	Pick points	Default
Customer	• Grade C customers • Customers with Debt consolidation loans • Customers with mortgages • Customers with High debt to income ratio (dti) • Customers with Higher annual income • Customers with high Number of credit lines • Customers with high funded amount • Customer with member id 4419405	High
Nature of Loan	• Loan with High instalment • Loan Highest Tenure • Loan amount	High
Macro economical influence	• Year 2013	High
Demographic influence	• California	High
Loan Disbursement status	• High loan disbursement to people with highest experience of 10 years, followed by 1 year and 2 years	High

6. Modelling Approach

Before starting the modelling, **it is important to make the dataset proper fit and tuned for prediction modelling**. Considering the given dataset, we have applied following measures:

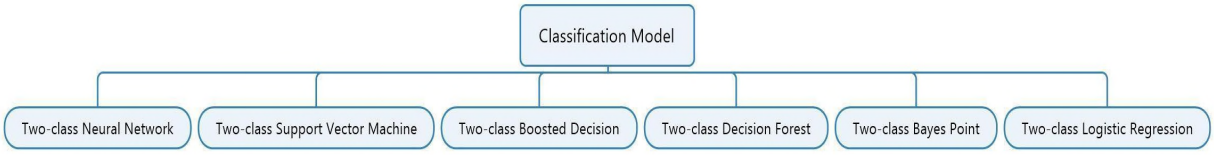
Model Tuning Techniques

#	Description	Outcome/ Results
1	**Model runs on batches** Before the revelation of Pre-Facto and Post facto the model was run on different batches and overfitting was observed in the results. It was then concluded that the model needs only select columns that need to be run for the accurate prediction. **Divide data into Pre-Facto and Post Facto** As per the Banking Loan data provided, it is important to separate the features which are available during Loan application from those which are tracked post	These columns are used for the modelling during column select module**: loan_amnt, funded_amnt_inv, term, int_rate, installment, grade, emp_length, home_ownership, annual_inc, loan_status, purpose, dti, open_acc, revol_bal, revol_util,total_acc, addr_state**

	loan issuance. This will ensure our model is predicting based on information present during the loan application. **Pre-facto- variables** that are available during loan application **Post facto- variables** that come into factor after the disbursement of loan and during the tenure of loan repayment or default.	
2	**Multicollinearity** is the occurrence of high intercorrelations among two or more independent variables, having the same value.	**Funded amount** was removed owing to collinearity with **Loan amount.**
3	**SMOTE -** Visualisation through Tableau and python programming showcased highly unbalanced dependent variable - Loan status the Fully Paid vs Default loan as **97% vs 3%.**	With a highly imbalanced classification dataset, '**SMOTE transformation' feature was used to build models.**
4	**Skewness or degree of distortion to normal data** Visualization through tableau and Python helped understand the skewed data points.	In order to remove the skewness, **Apply Math function was used to make the data near normal. Transformation of numerical values using LogPlus1 Transformation.[Yeo-Johnson]**
5	**Splitting the Dataset**	Splitting the dataset into test and training data in the ratio of **'70:30'**, with **stratified set to 'True'**, and **random seed to '12345'**
6	**Deciding factor of selecting Model- Confusion Matrix**	The accuracy and precision on various classification models were similar and close. Hence, it was concluded that - **Confusion matrix takes precedence over accuracy and precision for the choosing the right model.** Details on the confusion matrix are provided below.

Modelling Approach

We have built various classification models on Azure ML using the cleaned and tuned dataset. We evaluated the performance of all the models using the **CONFUSION MATRIX**. The model with better performance would be used for deployment and 'default'predictions. The following classification models have been used for the prediction model.



The following are the results of the confusion matrix of models with the accuracy, precision, recall and F1-score.

Confusion matrix

Classification Models	True Positive	False Positive	True negative	False Negative
Two-class neural network	1247	4	981	33512
Two-class support vector machine	34759	985	0	0
Two-class Boosted decision	**34610**	**961**	**24**	**149**
Two-class decision forest	34641	979	6	118
Two-class logistic regression	34759	0	0	985
Two-class Bayes point	34740	984	1	19

Classification Models	Accuracy	Precision	Recall	F1 Score
Two-class neural network	0.062	0.997	0.036	0.069
Two-class support vector machine	0.972	0.972	1.000	0.986
Two-class Boosted decision	**0.969**	**0.973**	**0.996**	**0.984**
Two-class decision forest	0.968	0.973	0.997	0.984
Two-class logistic regression	0.972	0.972	1.000	0.986
Two-class Bayes point	0.972	0.972	0.999	0.986

Observations

From the above results of confusion matrix and score 'Two-Class Boosted Decision' stands out as a better model. The reason for selecting this model is not only **accuracy** but also the **performance of the model.**

Classification Models	True Positive	False Positive	True negative	False Negative
Two-class Boosted decision	34610	961	24	149

True Negative

The negative side of the confusion matrix is the default status of a customer. True negative of the model has count of **24**, which means the model is able to predict the defaulters from test data who have defaulted.

False Negative

This indicates borrowers who were classified as defaulters but instead would repay loan amounts in full. The figure **149** in this model is comparatively lower from other models. This cannot be a high number because banks would then end up rejecting loans and lose business with customers having good credit profiles who would not actually default.

True Positive

This indicates the fully paid status of loans, which are identified appropriately by the model.

False Positive

The positive label indicates the fully paid loan status. The model selected as a top performing has false positive count of **961**.This means the model is predicting that these would be fully paid loans and would not default, whereas these are likely to default. Ideally this should be on lower side since banks may end up giving loans to customers who are likely to default without adequate risk and benefit evaluation**.**

7. Actionable insights and recommendations to the stakeholders

There are higher defaulters in Debt-consolidation category, state of California, dti ranging between 15 to 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 instalments are high risk customers who are shown to default more based on the given dataset.

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 also largely benefit from the application of Machine Learning. **If the bank decides to lend the borrower, who has been predicted to default, it can ask the borrower for collateral or guarantee or both, and charge a higher rate of interest. 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.**
Bank can aim to provide good quality data with minimum missing values which would help to build a more robust prediction model. To summarise, a better understanding of the default behaviour and of the regional differences in these **credit markets** could help policy makers to undertake more effective **risk-mitigating actions.**

Feature Engineering – Way Forward

A significant number of anomalies are detected in the data, nothing but the suspicious cases. The anomalies [D] could be analysed using a model. Along with anomaly study, additional details provided by **banks pertaining to customer age at loan, age of the account, customers with or without credit cards, number of family members operating the account would help build a robust model. Data points related to Heuristics, household size, marital status, and gender could also substantiate model tuning.**

8. References and Bibliography

[1] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
[2] https://www.geeksforgeeks.org/machine-learning-for-anomaly-detection/
[3] https://analyticsindiamag.com/7-types-classification-algorithms/

9. Appendix

Azure ML Link:

[A] https://gallery.cortanaintelligence.com/Experiment/Bank-Loan-Default-Two-class-Boosted-decision-and-Decision-forest
[B] https://gallery.cortanaintelligence.com/Experiment/Text-Mining-Bank-Loan-Default

Python Link:

[C] https://colab.research.google.com/drive/19ClQXQjfLPMzqJyNZcya_N2JgAxbbFmn?usp=sharing
[D] https://colab.research.google.com/drive/1PbAJA9loBthDs6pJfzHS-ytH315SL2O7?usp=sharing
[E] https://colab.research.google.com/drive/13TCAwMB3H2ETAvH9FCwHwOFLNt-2ai-l?usp=sharing
[F] https://colab.research.google.com/drive/1rm7jU7j1AufhIjXtthuwJ5Bf_3ObrZ42?usp=sharing

Tableau Link:

[G] https://public.tableau.com/app/profile/narendra.kr.gupta/viz/NKG_BL0512/Dashboard1?publish=yes