greatlearning

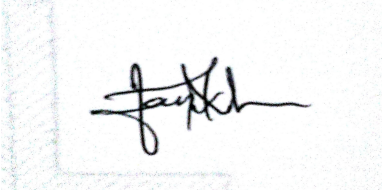
Capstone Project - Final Report

Loan Default Prediction

Batch Details	PGPDSE-FT Online June21-A
Team Members	Arjun Rajmohan, Harsh Kukreja, Jaykaran EBK, Keerthana Vivin, Ritika Kashyap
Domain of Project	Predictive Analytics
Proposed Project Title	Loan Defaulter Prediction
Group Number	5
Team Leader	Jaykaran EBK
Mentor Name	Ms. Vibha Santhanam

Date: 17-03-2022

Signature of the Mentor



Signature of the Team Leader

ACKNOWLEDGEMENT

We would like to thank our mentor **Ms Vibha Santhanam**, for providing her valuable guidance and suggestions for our Project work. We also thank her for the continuous encouragement and the interest shown towards us to complete our project.

We are extremely grateful to all our teaching and non-teaching staff members of **GREAT LEARNING**, who showed keen interest and inquired about our developments.

We would like to express our gratitude to all teaching and non-teaching staff members of **Great Lakes Institute of Management**, for providing their support and guidance for our project.

CONTENTS

1. INTRODUCTION

The main goal in the banking system is to invest their resources in safe hands wherever it is. Through this model we are able to predict the risk profile of the loan applicant and the whole method of validation of attributes is automated by machine learning technique. Loan Prediction is useful to assess the risk profile prior to loan disbursement thereby reducing the Non-Performing Assets (NPA) of the financial institution and increasing their bottomline.

1.1 NEED FOR THE PROJECT

Most individuals around the world in some way depend on banks or financial institutions to borrow loans to meet their personal and business needs thereby overcoming their financial constraints and achieving their objectives. Due to the dynamic nature of the economy and ever-increasing competition in the financial world, the activity of taking a loan has become inevitable. Also, small scale to large scale banking firms depend on the activity of lending out loans to earn profits for managing their affairs and to function smoothly at times of financial constraints. A loan is the major source of income for the banking sector as well as the biggest source of financial risk for banks. Large portions of a bank's assets directly come from the interests earned on loans disbursed.

With the improving banking sector in recent times and the increasing need of taking loans, a large population applies for bank loans. But one of the major problems banking sectors face in these unprecedented times is the increasing rate of loan defaults, and the banking authorities are finding it more difficult to correctly assess loan requests and tackle the risks of people defaulting on loans. Though lending loans is quite beneficial for both the parties, the activity does carry great risks. These risks represent the inability of a borrower to pay back the loan by the designated time which was decided mutually by both the lender and the borrower and it is referred to as 'Credit Risk'. For that, it is highly necessary to assess the clients credit suitability before sanctioning a loan.

1.2 OBJECTIVE OF THE PROJECT

The project objective is to build a model that predicts whether a borrower defaults on their loan or not. This is based on given characteristics and lifestyle of debtors and finding their relationship with defaulting. The end goal is to help lenders minimize their risk and losses.

An organization wants to predict who possible defaulters are for the consumer loans product. They have data about historic customer behavior based on what they have observed. Hence when they acquire new customers they want to predict who is riskier and who is not.

1.3 SCOPE

The scope of this project is to implement and investigate how different supervised binary classification methods impact default prediction. The model evaluation techniques used in this project are accuracy, precision, sensitivity, F1-score. The classifiers that will be implemented and studied are:

- Decision Trees
- KNN
- Logistic Regression
- Naive-Bayes
- Random Forest
- XGBoost
- AdaBoost
- LGBoost

The project will be performed using LendingClub data from 2013-2018. With regards to this fact, the results presented in this thesis will be biased towards the profile of LendingClub's clients, specifically their location and other behavioral factors. LendingClub's clients are from the United States, and thus it will be impacted mainly by the behavior of clients from here.

2. Literature Survey

Evans Brako Ntiamoah, Emmanuel Oteng, Beatrice Opoku, Anthony Siaw have studied loan default rate and its impact on profitability in financial institutions. They have used both qualitative and quantitative methods to perform the analysis. The statistical findings of the study show significantly that proper management of loans given to clients yield more profits for the firms. Also there was a significant relationship between the problem of recovery and overdue of loans and profitability.[Loan Default Rate and its Impact on Profitability in Financial Institutions, 2014]

Golak Bihari Rath, Debasish Das, and BiswaRanjan Acharya have tried to create and compare machine learning algorithms such as logistic regression, decision tree, and SVM to predict loan eligibility of individuals based on past loan repayment. They found the performance of logistic regression better than decision trees and SVM as a predictive model for future payment behaviors of the loan applicants. [Modern Approach for Loan Sanctioning in Banks Using Machine Learning, 2021]

Neel M Shah has tried to analyze the prediction of educational loan defaults as an AI and ML use case. His study concluded that ensemble models outperform simple AI models and statistical models and that the performance of both the models could be improved by stacking the models. He also observed that collateral-free loans had a considerably higher rate of default as compared to the loans with collateral. [Predicting Educational Loan Defaults:Application of Artificial Intelligence Models, 2019]

Loans disbursed to MSMEs during the pandemic could pose a higher risk of fraud as banks do not continuously monitor them, according to Deloitte India's latest banking survey.

An article by The Daily Star (Bangladesh) highlights the inconsistencies in lending practices wherein regulators had been granting defaulters (consisting of willful defaulters) a lot of leniency leading to many large borrowers to reschedule their loans practically infinitely using their political connections, violating all banking rules and norms. The article concludes that the only solution to this problem is for regulators to return to applying banking rules and regulations uniformly for everyone.

Another article by Money Control observed that credit card spending along with unsecured personal loans had seen the fastest growth in the past three years. The public sector banks have higher loan approval rates in consumer credit than private-sector lenders. This has coincided with high delinquency rates as well. Public sector lenders need to tighten their underwriting rules.

3. Dataset and Domain

The full dataset provided consists of roughly 2 million records with 152 columns from 2007 - 2018, describing characteristics of the loan applicants. These columns include annual income, funded amount, house ownership, experience, state and so on. The loan status column shows whether the customer has fully paid or charged off.

Due to software limitations, initially the data from 2016 - 2018 was extracted for model building. After removal of nulls and assessing the relevance of the columns, 500,000 rows and 33 columns were selected. The dropped columns either had a high percentage of null values or were too similar to each other. E.g. Number of derogatory public records and Number of public record bankruptcies. Nulls for the relevant columns were imputed using median and mode accordingly.

The IQR method to remove outliers removed half the data, so it has not been used. F1 scores were also reduced after removing these outliers. Instead, only the extreme outliers have been removed.

The target variable of loan_status was unbalanced. This resulted in poor performance of the base models. To rectify this additional data was added. We added 3 additional years worth of data. The dataset includes information from 2013-2018. The non-defaulters (who formed the majority class) were undersampled. This variation had a significant impact on model performance.

Source: Kaggle

3.1 Data Dictionary

addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
chargeoff_within_12_mths	Number of charge-offs within 12 months
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
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.

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.
emp_title	The job title supplied by the Borrower when applying for the loan.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
issue_d	The month which the loan was funded
issue_year	The year which the loan was funded
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.
loan_status	Current status of the loan
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower

tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently in the borrower's credit file
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

ADDITIONAL FEATURES USED IN THE FINAL MODEL

earliest_cr_line	The month the borrower's earliest reported credit line was opened
fico_avg	(Last_fico_range_high + Last_fico_range_low)/2. FICO score is one kind of credit score with a range of 300 to 850 with a higher score indicating better credit.

3.2 Variable Categorization

No. of Categorical Variables	16
No. of Numeric Variables	19

3.3 Data Preparation and Preliminary Analysis

1. **Missing/Null Values**

   Most records with null values were removed. However, in columns referring to months since the last incident, a null value signified the borrower had no incidents. Eg: A null value in mths_since_last_delinq meant they have not had a previous delinquency.

   For these columns, bins were formed instead and the nulls were replaced with ‘None’.

2. **Duplicate Rows**

   There are no duplicate rows in the dataset.

3. **Redundant Columns**

   - Columns like funded_amnt_inv, funded_amnt and loan_amnt have nearly identical values.
   - Issue_d is not relevant to the purpose of this study.
   - Id column is unique for each row.

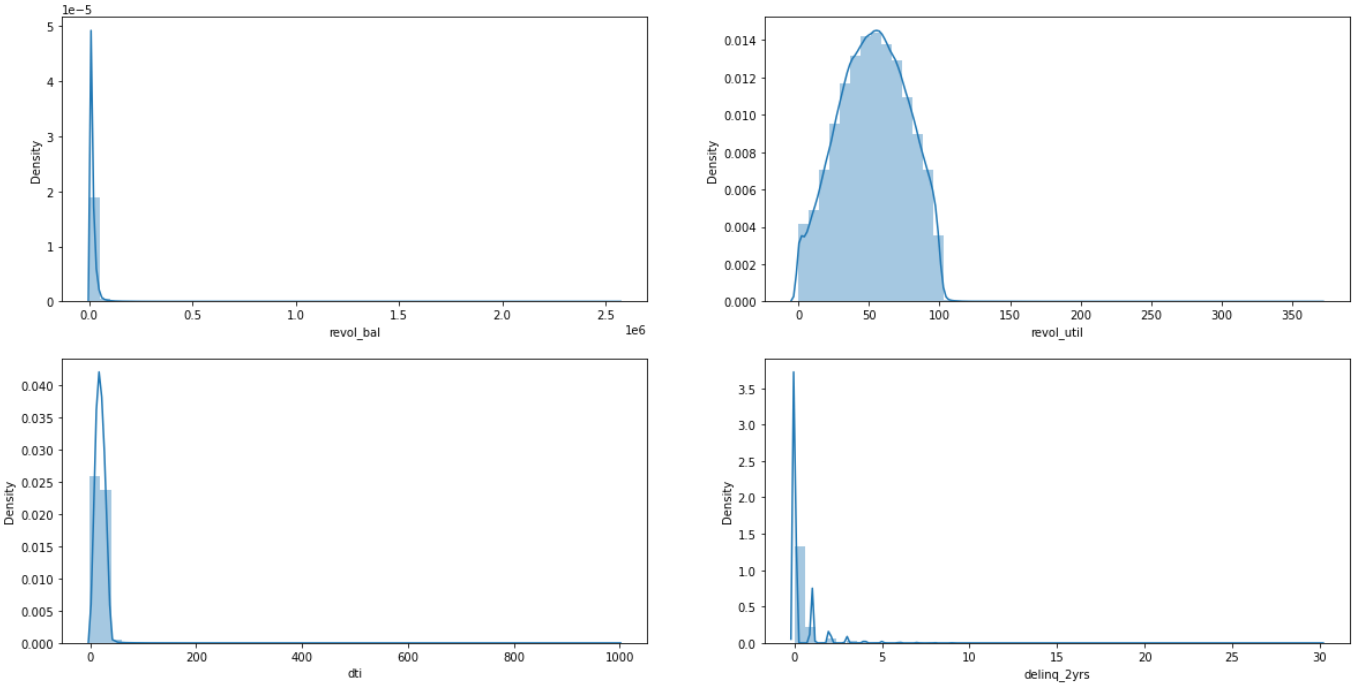
   The above columns were hence dropped.

4. Data Preparation

- The home_ownership column had two categories ‘ANY’ & ‘NONE’. These values are replaced with ‘OTHERS’.
- The FICO_Average column was created by taking the average of FICO_range_high & FICO_range_low.
    - last_fico_range_high: The upper boundary range the borrower’s last FICO pulled belongs to.
    - last_fico_range_low: The first 3 numbers of the zip code provided by the borrower in the loan application.
- All the states were now bucketed into different regions like ‘west’, ‘south_west’, ‘south_east’, ‘mid_west’ & ‘north_east’.
- There were many categories in the purpose column , these were now bucketed into major categories like ‘debt’, ‘Major_purchase’, ‘life_event’ & ‘Others ’.
- Lastly, year was extracted from the column ‘earliest_cr_line’.

5. Univariate Analysis of Numeric columns

- Histograms were plotted to understand the distributions of the columns.
  The following features appear strongly right-skewed:
  funded_amount, funded_amount_inv, int_rate, installment, annual_inc, loan_amnt, pub_rec.
- Apart from the columns mentioned above revol_bal, revol_util, dti, delinq_2yrs showed extremely right skewed distributions with extreme outliers.

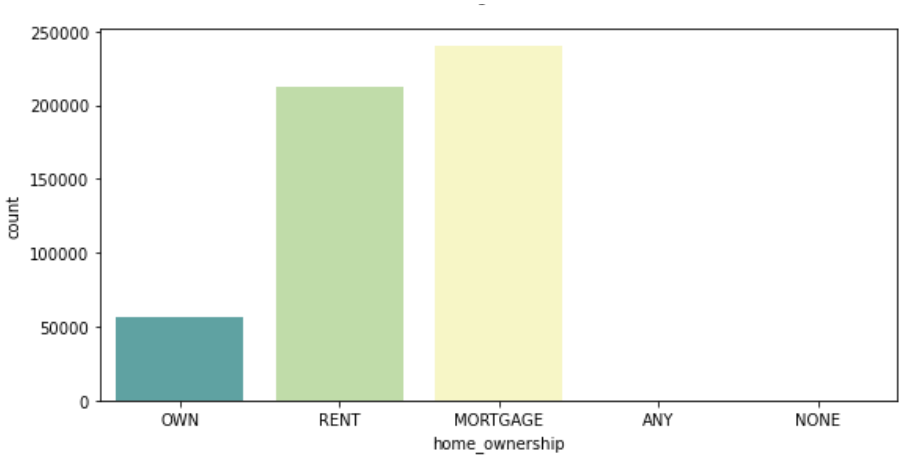


6. Univariate Analysis of Categorical columns

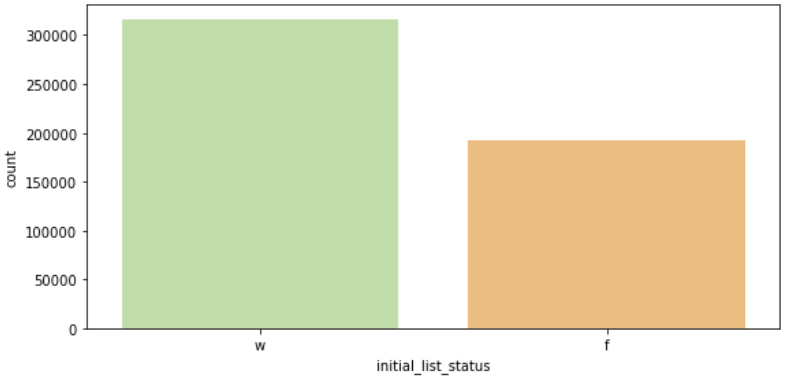
Count Plots were created for each categorical column to understand how the subcategories are distributed for this dataset.

From the plot we can draw the following insights:

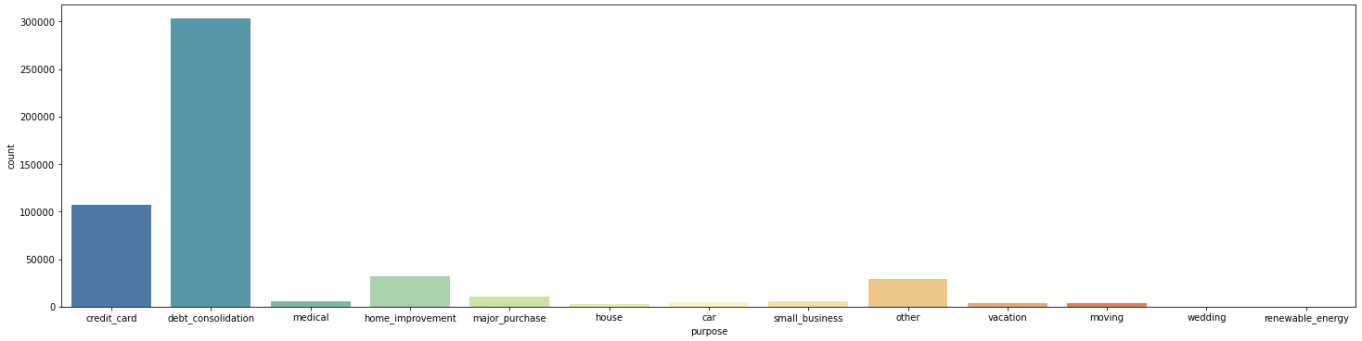
- Majority of the applicants have mortgaged or rented homes.



- Most loan applicants are given 'Wait' as the initial list status.



- The most common purpose for loan is debt consolidation, followed by credit card



7. Outliers

- IQR method removed around 50% of records.
- The F1 scores decreased after removal with IQR.
- Instead, only extreme outliers were removed.

3.4 Project Justification

Project Statement

Loan defaulting is when a debtor is unable to repay the debt. The project objective is to build a model that predicts whether one defaults on their loan or not. This is based on given characteristics and lifestyle of debtors and finding their relationship with defaulting. The end goal is to help lenders minimize their risk and losses.

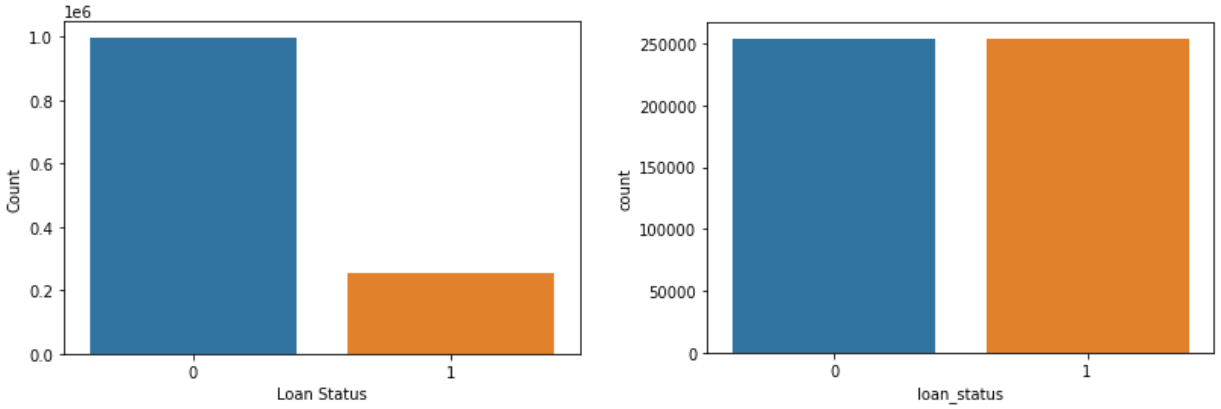
Value

The key problem regarding loans is for financial institutions to correctly identify who to lend to. Using predictive modeling, this can be accomplished. The project's aim is to create a model that can aid in this decision making.

For banks the model could play a significant role in profit maximization. It would also reduce the chances of Non-Performing Assets. For customers it could reduce bias in the loan assessment process giving everyone an equal opportunity to obtain a loan.

4. Data Exploration (EDA)

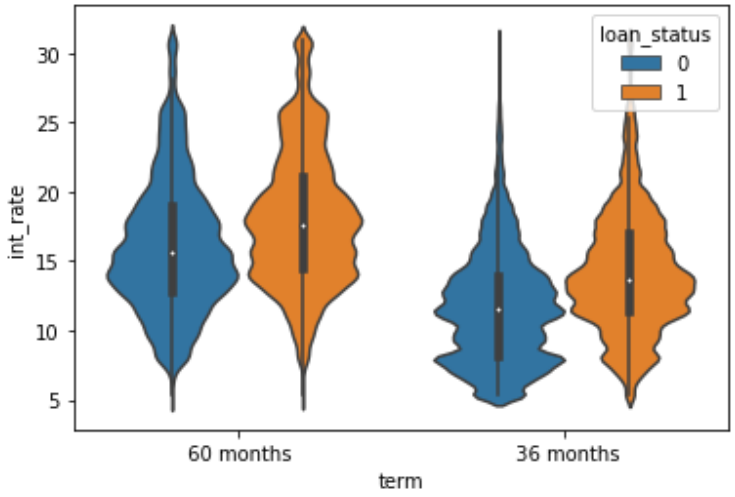
4.1 Distribution of the target



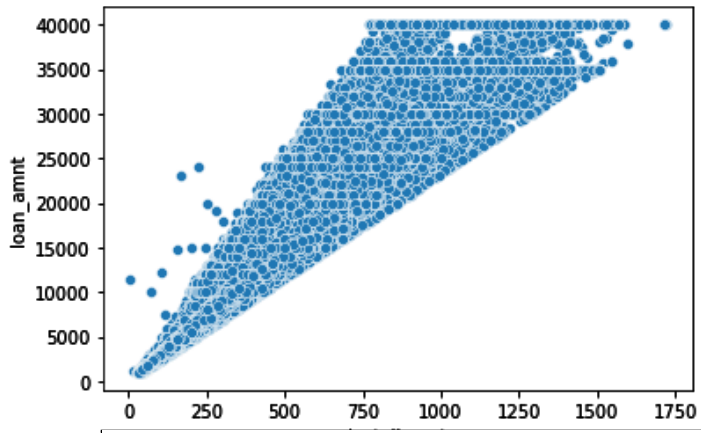
It can be seen that the target variable is very imbalanced. We have used SMOTE to undersample the target variable to balance the data for further modeling.

After undersampling the majority class, the two classes are balanced.

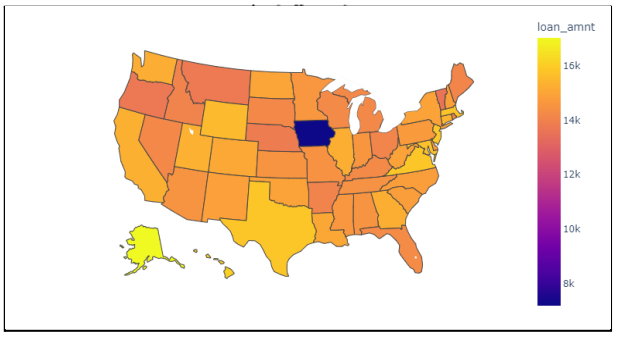
4.2 Relationship between independent variables



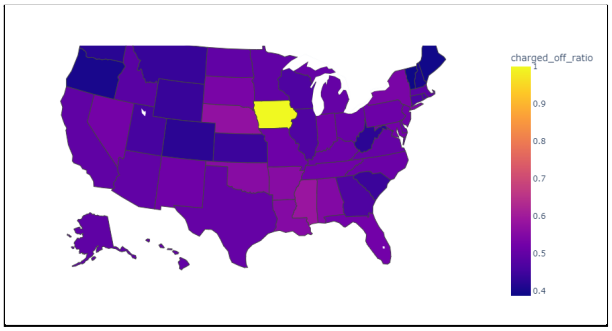
More people have higher interest rates for the 60 month term for loan when compared to 36 month term.



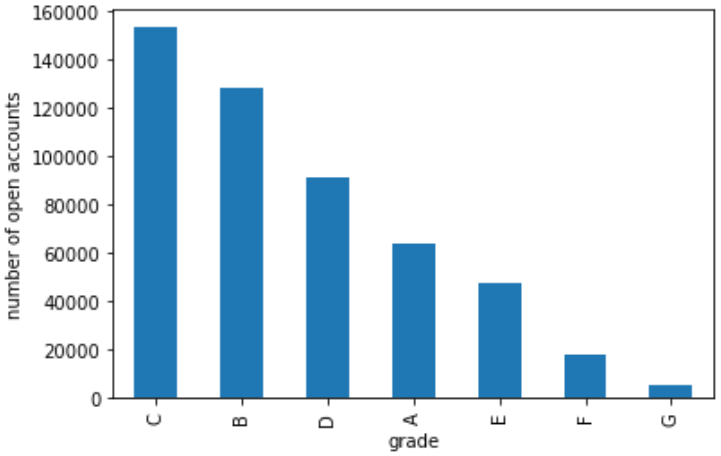
There is a clear positive linear relationship between installment and loan_amnt.



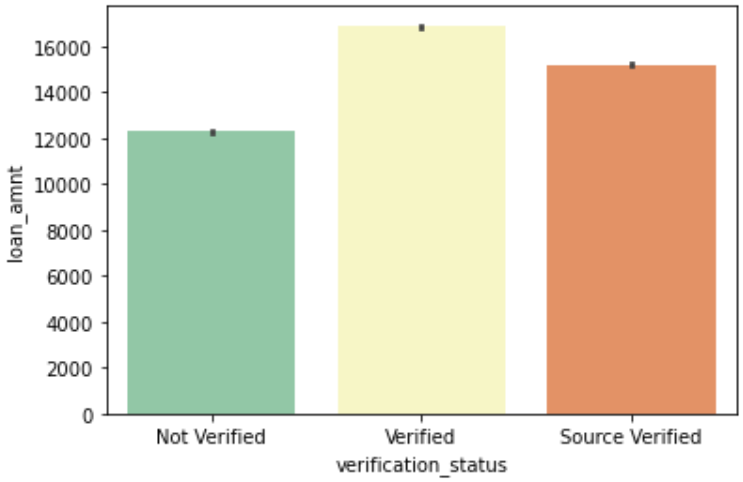
Alaska has the highest average loan amount.



Mississippi has the highest charged off ratio.

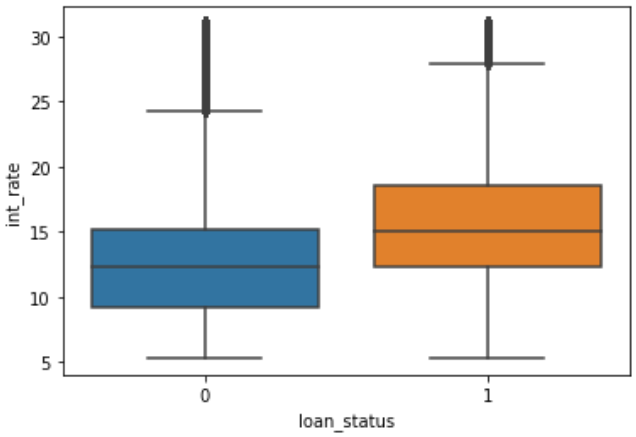


Grade C and B have the highest number of open credit lines in the borrower's credit file.



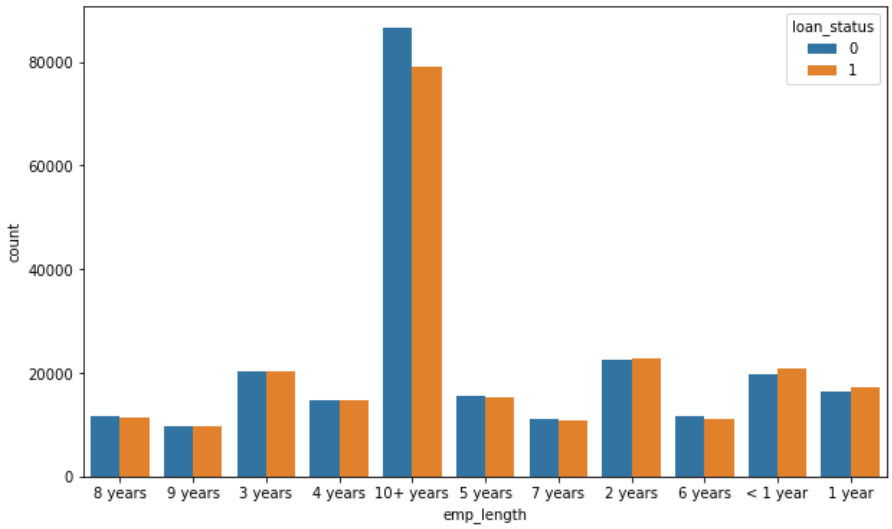
The average loan amount is slightly higher for those with verified income.

4.3 Relationship between numeric variables and target



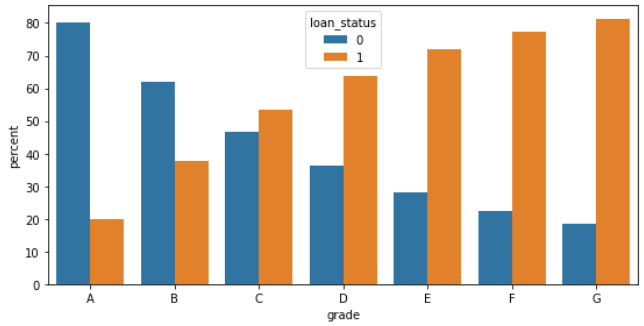
Charged off has higher interest rates as compared to full paid loan status.

4.4 Relationship between Categorical Variables and Target

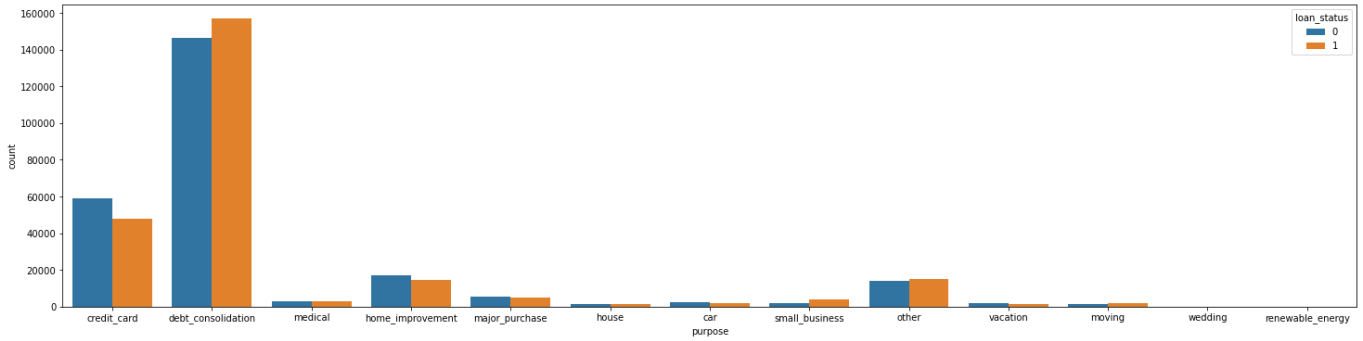


Emp length of 10+years has most number of Fully paid loans

Most of the fully paid loans are for terms of 36 months.



Grade C has the highest charged off count.



Most of the loans were fully paid with the purpose which they were granted being Debt_consolidation

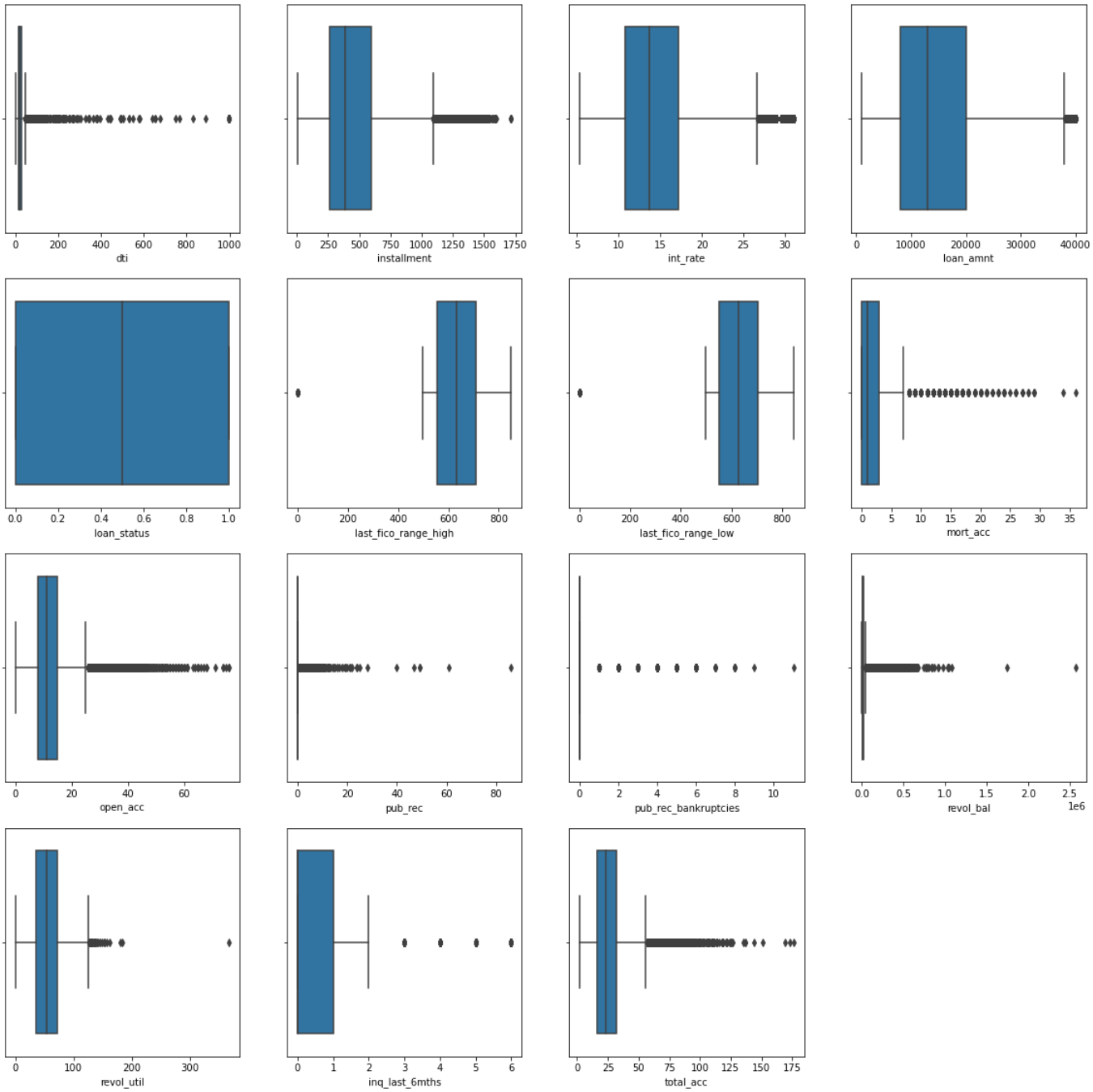
4.5 Multicollinearity

It can be inferred from the heat map that,

- total_high_cred_lim is strongly correlated with total_cur_bal
- loan_amnt and installment are highly correlated

Most of the features share a low correlation. Highest correlation can be found is 0.95 (Between Loan Amount and Installment). Hence there is a possibility of few variables exhibiting multicollinearity.

4.6 Outliers



Only a small number of extreme outliers were removed.

4.7 Statistical significance of variables

4.7.1 Statistical test for numerical vs categorical variables:

Shapiro test:

Null hypothesis: The distribution is a normal distribution

Alternate hypothesis: The distribution is not a normal distribution

Attribute	**P-value**
Annual Income	0.0
Interest Rate	0.0

If p-values lower than 5%, null hypothesis is rejected and if P-value >0.05, we fail to reject the null hypothesis.
The above attributes have been found to have a significant influence by their P-values using Shapiro-test. (p-values are lesser than 5%)

Mann-Whitney U Test:

Null hypothesis: Both distributions come from the same population

Alternate hypothesis: Both distributions do not come from the same population

Attribute	**P-value**
Annual Income	0.0
Interest Rate	0.0
Pub_rec_bankruptcies	0.0
Revol_bal	0.0

Since p-value is lower than 5% for both the attributes, we conclude that two distributions do not come from the same population

Kruskal Wallis Test:

Null hypothesis: The medians of all groups are equal

Alternate hypothesis: At least one median is significantly different from the other

Attribute	**P-value**
Home Ownership	0.0

Since p-value is lower than 5% for both the attributes, we conclude that at least one of the group's median is different from other.

4.7.2 Statistical test for categorical vs categorical variables:

Chi-Square test for independence:

Null hypothesis- The two variables are independent
Alternate hypothesis- The two variables are dependent

Variable	P-Value
Grade	0.0
Home Ownership	0.0

The above attributes have been found to have a significant influence by their P-values using chi square test. (p-values are lesser than 5%)

5. Feature Engineering

5.1 Encoding of categorical variables:

We have used different encoding techniques for different features. We have used Ordinal encoding for 'Grade' and 'Emp_length' columns. We have split the 'purpose' column into three levels: 'debt_consolidation', 'credit card' and 'other' as the first two were high in count. We have used Target encoding for 'addr_state' as it had too many levels.

For the remaining features we have used dummy encoding.
Our final dataframe has 54 features.

5.2 Scaling
We have used Standard Scaler to scale our numeric variables.

5.3 Feature Creation
- The creation of the feature, fico_avg, improved results from the base model significantly. This was created taking the average of 'last_fico_range_high' + 'last_fico_range_low'.
- Regions were added as West, Southwest, Southeast, Midwest, Northeast based on states.
- Purpose was streamlined into debt, major purchase, life event or other.

5.4 Dimensionality Reduction
- VIF removed too many important features which was reducing model performance. So features were manually dropped instead using heatmap.

6. Assumptions

6.1 Logistic Regression
1. Independence of error, whereby all sample group outcomes are separate from each other (No duplicate reports) Duplicates have been dropped
2. Linearity in the logit for any continuous independent variables.
3. Absence of multicollinearity. Multicollinearity has been checked.
4. Lack of strongly influential outliers. Outliers have been removed.

6.2 Naive-Bayes

1. The biggest and only assumption is the assumption of conditional independence.

6.3 Decision Tree

1. At the beginning, we consider the whole training set as the root.
2. Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
3. On the basis of attribute values records are distributed recursively.
4. We use statistical methods for ordering attributes as root or the internal node.

6.4 Random Forest

1. Assumption of no formal distributions. Being a non-parametric model it can handle skewed and multi-modal data.

6.5 KNN

1. The KNN algorithm assumes that similar things exist in close proximity.
2. Since KNN is a non-parametric algorithm it does not make assumptions about the distribution of the data

7. Base Model Results

	Train set				Test set			
Model Name	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.78	0.54	0.11	0.18	0.78	0.54	0.11	0.18
Naive Bayes	0.27	0.23	0.97	0.37	0.27	0.23	0.98	0.37
Decision Tree	1.00	1.00	1.00	1.00	0.68	0.30	0.31	0.30
Random Forest	0.98	1.00	0.91	0.95	0.77	0.43	0.11	0.18
KNN	-	-	-	-	0.73	0.34	0.22	0.27

Interpretations:

- Naive Bayes, being a rather simple model, did not perform very well. It had overall lowest scores across all metrics.
- Logistic regression performed slightly better than Naive Bayes but was still lacking as the scores in isolation were not good.
- Decision Tree without any pruning resulted in an overfitted model which did not perform well on the unseen test set.
- Similar to Decision Trees, Random Forest had the same problem of overfit.
- KNN required a lot of time to predict on the test set and precision and recall scores are also low.

8. Final Model Results

Second Round of Model Building:

- Added new independent variables to the data which we believed were relevant to the classification problem
- Eliminated Naive Bayes based on its poor performance compared to other models and tried to optimize the rest of the models.
    - Removal of Multicollinearity
    - Hyperparameter tuning
    - Feature Selection using Recursive Feature Elimination (RFE)
- Attempted three new models:
    - XGBoost
    - AdaBoost
    - LGBoost

  and performed the same optimization techniques to improve the models.

Results:

	Model_Name	Dataset	Accuracy	F1 Score	Precision	Recall
0	Logistic Regression	Train	0.895331	0.895198	0.896782	0.893620
1	Logistic Regression	Test	0.893725	0.895478	0.879945	0.911569
2	Logistic Regression: Youden's Index	Test	0.896491	0.897547	0.887459	0.907868
3	Random Forest	Train	0.992221	0.992199	0.995405	0.989015
4	Random Forest	Test	0.890362	0.890502	0.888321	0.892694
5	Random Forest: RFE + Hyperparamter Tuning	Train	0.893909	0.896328	0.876764	0.916786
6	Random Forest: RFE + Hyperparamter Tuning	Test	0.893679	0.895977	0.876036	0.916846
7	Decision Tree: RFE + Hyperparamter Tuning	Train	0.897741	0.899269	0.886458	0.912456
8	Decision Tree: RFE + Hyperparamter Tuning	Test	0.896012	0.897511	0.883756	0.911701
9	KNN	Test	0.837040	0.830065	0.866079	0.796926
10	AdaBoost	Train	0.894887	0.897307	0.877532	0.917993
11	AdaBoost	Test	0.895167	0.897490	0.877048	0.918907
12	AdaBoost: RFE + Hyperparameter Tuning	Train	0.896598	0.898783	0.880603	0.917730
13	AdaBoost: RFE + Hyperparameter Tuning	Test	0.896261	0.898325	0.879812	0.917634
14	XGBoost	Train	0.939243	0.940269	0.925097	0.955947
15	XGBoost	Test	0.896937	0.898555	0.883662	0.913958
16	XGBoost: RFE + Hyperparameter Tuning	Train	0.903225	0.905079	0.888498	0.922290
17	XGBoost: RFE + Hyperparameter Tuning	Test	0.900227	0.902001	0.885243	0.919406
18	LGBoost	Train	0.902239	0.904042	0.888101	0.920566
19	LGBoost	Test	0.900260	0.902001	0.885524	0.919104

8.1 Key Observations

- The model performances have improved significantly
- There is little to no difference in most model metrics.
- Based on the metrics XGBoost after RFE and HyperParameter Tuning performs the best on train as well as test set

8.2 XGBoost

- XGBoost is parallelizable which solves ML tasks by training with high performance. It basically helps to run core algorithms on clusters of GPUs.
- The XGBoost library implements the gradient boosting decision tree algorithm.
- This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines.

- Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made.
- Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

Results:

- Train accuracy and F1-Score: 0.94 and 0.94
- Test accuracy and F1-Score: 0.90 and 0.90
- This shows that there is slight overfitting. RFE and Hyperparameter tuning can be performed.

RFE was used to find the significant features, which returned 22 significant features.
Hyperparameter tuning was done using GridSearchCV which gave more balanced results between test and train.

	Model_Name	Dataset	Accuracy	F1 Score	Precision	Recall
14	XGBoost	Train	0.939243	0.940269	0.925097	0.955947
15	XGBoost	Test	0.896937	0.898555	0.883662	0.913958
16	XGBoost: RFE + Hyperparameter Tuning	Train	0.903225	0.905079	0.888498	0.922290
17	XGBoost: RFE + Hyperparameter Tuning	Test	0.900227	0.902001	0.885243	0.919406

Accuracy: 90%. Out of 100 loans, the model can predict 90% of results correctly. However, due to the nature of the prediction, it is not the most important metric.

Precision: ~89%. Quality of positive prediction made by the model. Refers to the number of true positives divided by total positive predictions. A high precision value indicates the model out of all the applicants the model predicted to be a defaulter, 89% of them were correct.

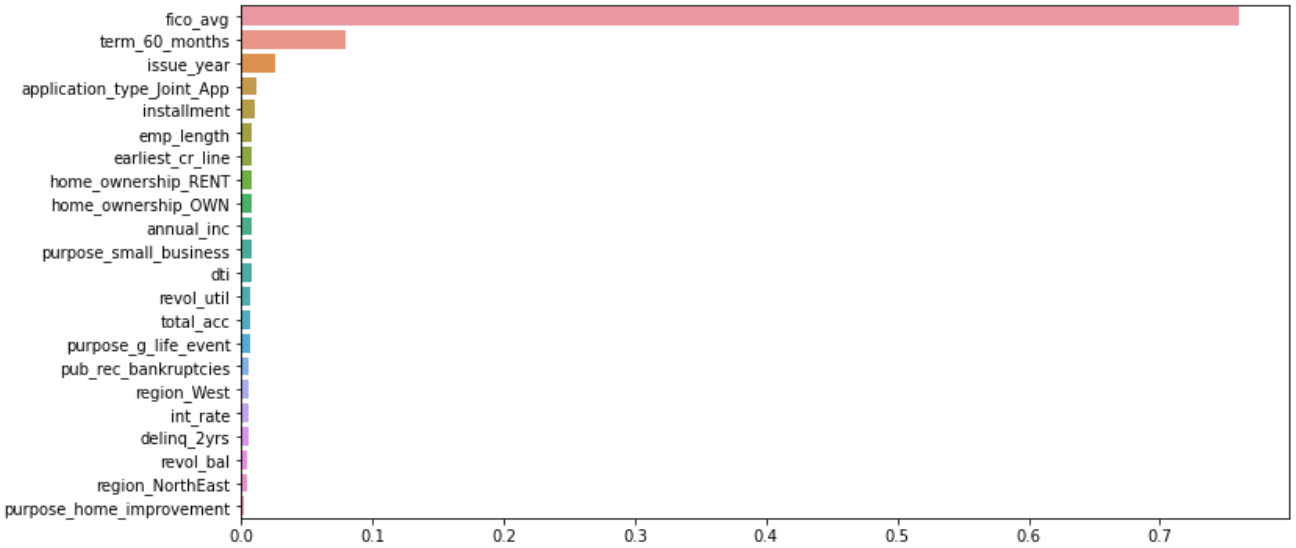
Recall: 92%. Target for optimization. Gives an indicator on the number of false negatives. Number of true positives by true positive + false negatives. False negatives could be costly, as they will cost the lender if there is default.

F1 Score: A combination of Precision and Recall. The high F1 score of 0.89 indicates that the model is good overall , since F1 score will only be high if both precision and recall are sufficiently high.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

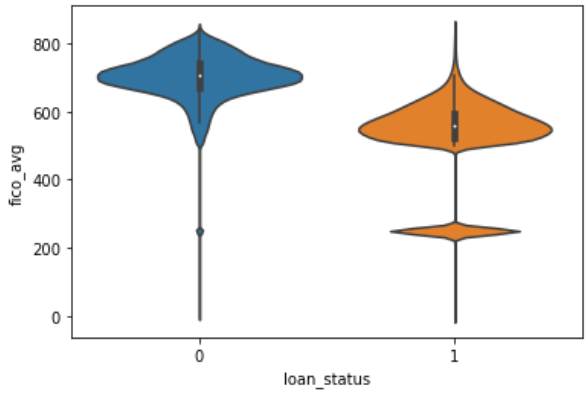
Since recall is our optimisation objective XGboost is the most suitable model since it gives us the highest recall scores. The recall score of 92% indicates that in the test set the model was able to successfully detect and classify 92 % of the defaulters. This is a high score and indicates that the model has a high potential to assist lending institutes in identifying potential defaulters.

8.3 Feature Importance + Visualizations

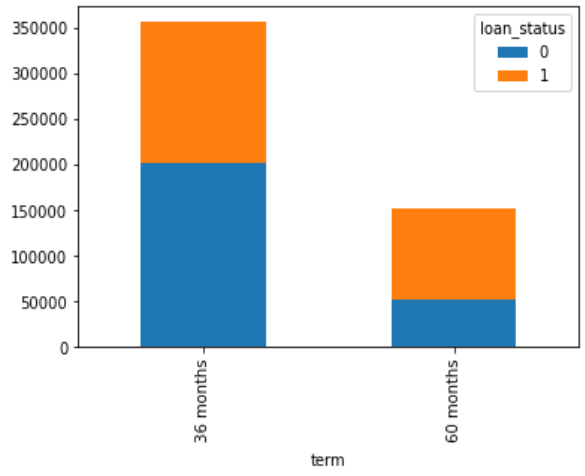


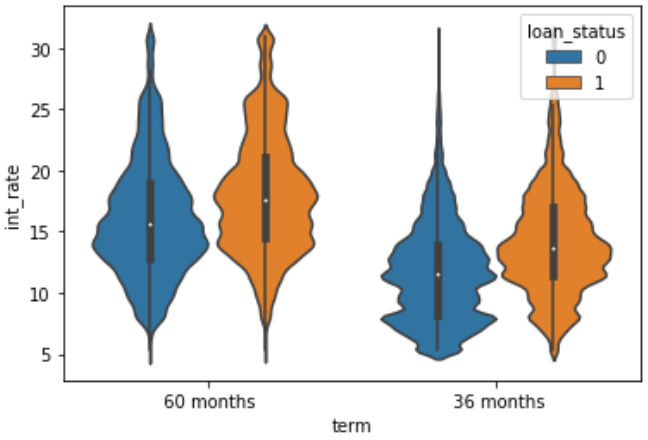
Fico_avg: The FICO score is one kind of credit score with a range of 300 to 850 with a higher score indicating better credit. The range is the reason for the large number of defaults around a score of 300. It is made up of the following five factors:

- Payment History (35% of score)
- Debt amount to credit limit ratio (30%)
- Age of credit (15%)
- Recent applications for credit (10%)
- More than one type of credit (10%)

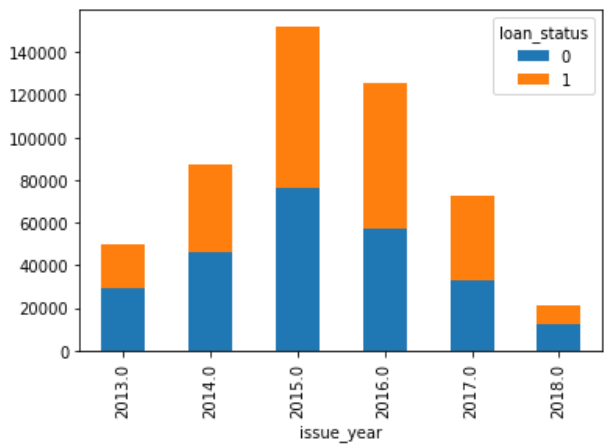


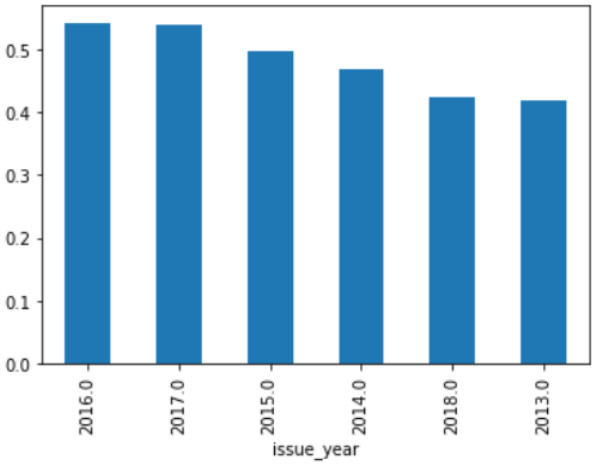
term_60_months: Terms are either 36 or 60 months. The 60 month term loans are significantly more likely to default as shown below. Longer loans have higher interest rates.



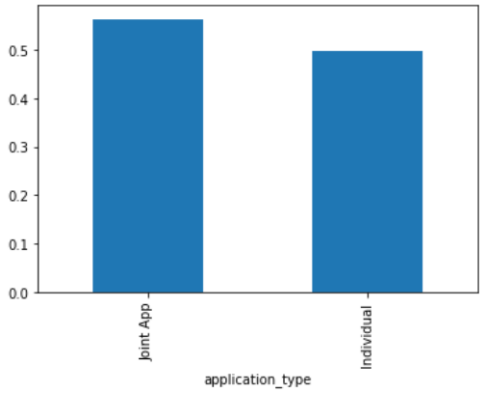


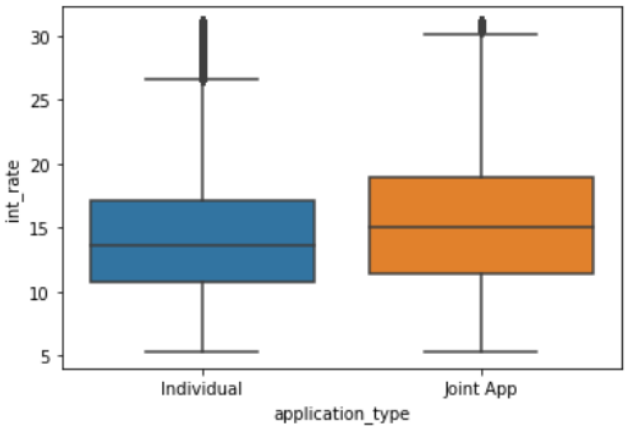
Issue_year: The graph on the right is a normalized representation of defaults in the years. We can observe 2016 and 2017 have the highest rate of defaults. As a new US president was elected in 2016, a possible reason may be the revised policies affecting loans. 2018 could be low considering the fact that not all of these loans are complete yet.



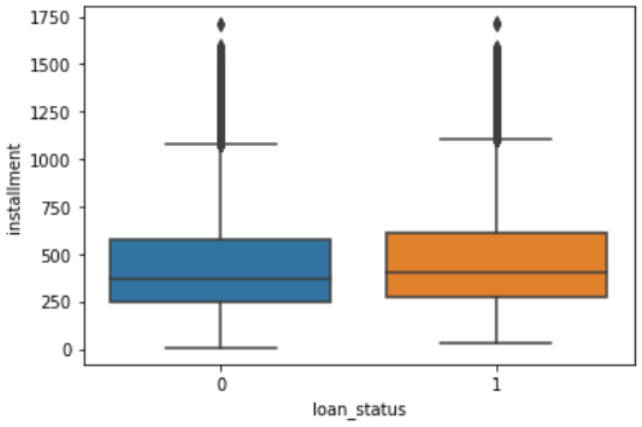


Application type: Indicates whether the loan is an individual application or a joint application with two co-borrowers. Loans with two co-borrowers are more likely to default. This may be because they were not creditworthy enough to begin with and hence needed a co-borrower to help finance their loan.

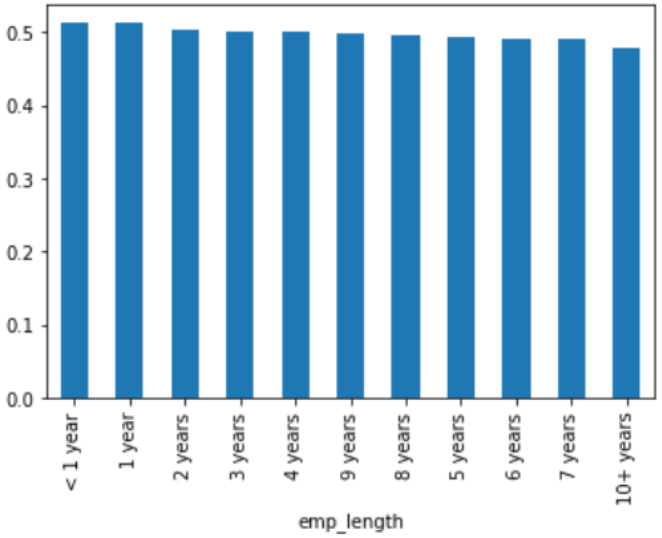




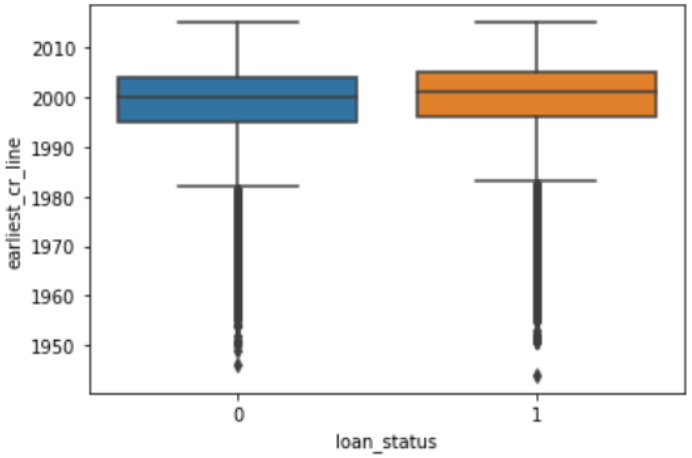
Installment: Larger installments are more likely to cause difficulties for the borrowers. Larger installments leads to more defaults.



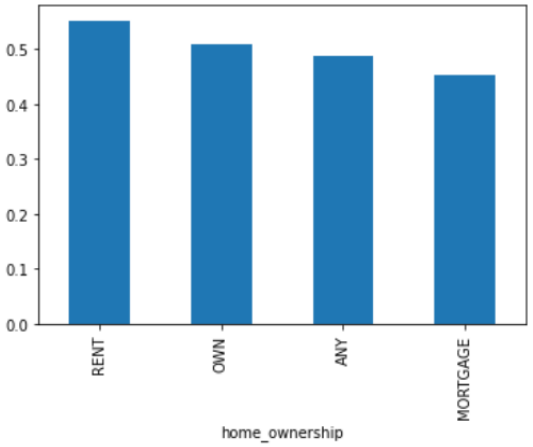
Emp_length: Employment length in years. As seen in the diagram below, those who are employed more than 10+ years have the least percentage of defaulters. For other experience levels, the number of defaulters vs non defaulters are roughly the same. The lowest five categories have the highest defaulter percentage.



Earliest_cr_line: The month the borrower's earliest reported credit line was opened. As shown in the box plot, defaulted loans have their median earliest_cr_line higher than non-defaulters.

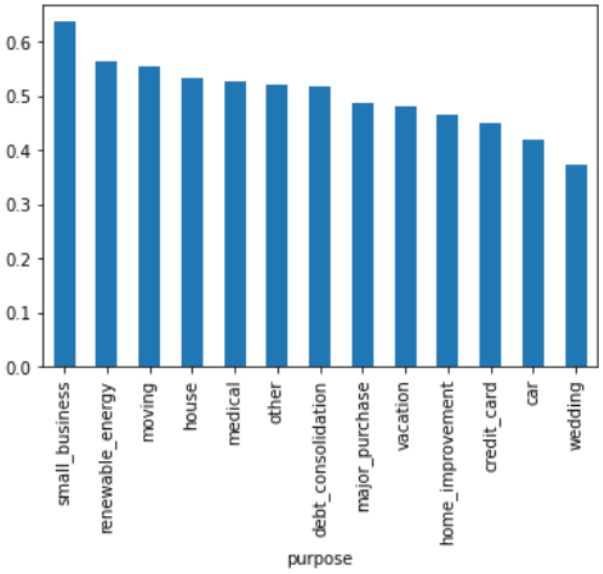


home_ownership_RENT and home_ownership_OWN: Those who live in rented homes are more likely to default on their loans.



Annual_inc: Higher annual income means they are more likely to repay their loans. This is because low income borrowers may not have as much disposable income (income after subtracting expenses). Naturally, this leads to more failure to repay installments and ultimately, loan default.

Purpose_small_business: When it is for the purpose of small business, the percentage of default is more than half. From the graph, we can infer small businesses are by far the riskiest to lend to.



Dti: The higher the debt to income ratio, the higher rate of default. This is self explanatory. If there is more debt than the income of the borrower can handle, they are likely to default.

8.4 Comparison to the benchmark

Random Forest was chosen as the benchmark against which we could evaluate our model performance.

Results from benchmark : Random Forest

	precision	recall	f1-score	support
0	0.86	0.72	0.78	70700
1	0.37	0.58	0.45	20124
accuracy			0.69	90824
macro avg	0.62	0.65	0.62	90824
weighted avg	0.75	0.69	0.71	90824

Results from final model: XG Boost

	precision	recall	f1-score	support
0	0.92	0.88	0.90	76363
1	0.89	0.92	0.90	76184
accuracy			0.90	152547
macro avg	0.90	0.90	0.90	152547
weighted avg	0.90	0.90	0.90	152547

We can see that in the final model, all test scores are higher than the benchmark model. Particularly precision and recall scores for the positive class have improved. The overall f1 score has doubled from 0.45 to 0.9 . Compared to the benchmark, the final model has significantly reduced both bias and variance.

8.5 Implications

Small scale to large scale banking firms depend on the activity of lending out loans to earn profits for managing their affairs and to function smoothly at times of financial constraints. A loan is the major source of income for the banking sector as well as the biggest source of financial risk for banks. Large portions of a bank's assets directly come from the interests earned on loans given.
The model:

- Can be used to assist in the decision making process.
- Can possibly reduce bias towards loan applicants.
- Can minimize the risks associated with lending.

Since our model was able to detect and correctly classify 92 % of the defaulters in the test set, we would recommend that it would be more advisable to reject or further scrutinize the applications of those applicants that the model predicts as defaulter.

8.6 Limitations

- XGBoost is Black Box. It is difficult to interpret the inner workings of the model. This makes it difficult to judge how individual factors affect the model's accuracy.
- Unable to capture trends that have occurred post 2018.
- Adjustments may need to be made frequently.
- Black swan events can be difficult to apply the model to. The algorithm cannot discern these situations. E.g. Wars, pandemic, etc.
- Demographic information such as age, marital status and family size are unavailable. These may have helped us improve the model.
- There is little to no information on loan applications from the state of Iowa.

9. Conclusion

In this paper, we have successfully used the Extreme Gradient Boosting (XGBoost) for bank loan default prediction. The task was to predict if a loan applicant will default on loan payment or not. The analysis was implemented in the python programming language, and performance metrics like accuracy, recall, precision, F1-score were calculated. From the analysis, we found out that the most important features used by our model for predicting if an applicant would default depends heavily on the average FICO score of the applicant. Additionally, term duration of the loan and issue year are also important features. This paper provides an effective basis for loan credit approval in order to identify risky customers from a large number of loan applicants using predictive modeling.

10. Deployment

We have deployed the model as a web application on Heroku using Streamlit. Streamlit turns data scripts into shareable web apps. Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud.

https://loan-default-prediction-app.herokuapp.com/

Here are few snapshots: