Project Summary

Batch Details	PGP-DSE October'21
Team Members	Nikhil Mishra,Piyush Dalal,Rijul Choudhary,Anupriya Bhandari,Prakhar Raj Gupta,Ayush Dagar,Deepansh Saxena,Shivam Sharma
Domain of Project	Finance and Risk Analytics
Proposed Project Title	Predict the probability of customer's loan approval
Group Number	2
POC	Nikhil Mishra
Mentor Name	Mr. Ankush Bansal

Table of Contents

Project Details

Business Problem and Objective

A bank has recently started generating leads through digital channels to cope with the marketing strategy of competitor banks.They source leads through various channels like search, display, email campaigns and via affiliate partners.The bank staff is not able to prioritize the leads that are more likely to get converted into their customers.

The marketing and sales department of the bank is wondering whether they could predict lead conversion using the data captured via digital channels on the basis of the data of their existing customers.

The Bank wants to identify leads' segments having a high conversion ratio i.e, lead to buying a financial product. The focus of the bank here is to increase the number of leads getting into the conversion funnel.

The leads are identified by assessing and identifying patterns within the loan approvals as some of the better customers can be very well evaluated based on their loan history.

Business Problem Goals

1. What would you achieve by this project?

For a bank it would be helpful to know which customers could be potential customers for other financial services. By using a Machine Learning Model and predicting customers suitable for loan approvals we would be able to get an automated and better understanding of our customers who would enjoy/purchase our other financial services.

2. How would this help the business or clients?

The Machine Learning model by correctly predicting our loan approvals, would help us identify and focus on/target our potential customers. This would save the bank from incurring losses by targeting customers who would not be potential leads.

3. What is the further scope of the project?

Automated web based products on loan processing for the finance sector.

By better identifying the potential customers, we can identify which segment of the market they are originating from and we can invest the business' resources on capitalizing on the identified segment. By automating the process through our Model, we can automatically classify our future customers.

4. What are the limitations of the project?

Our Model is trying to predict the potential successful leads through loan approvals that the customers applied for. This is one way for predicting the potential customer base. There might be other influential methods to identify and predict better funneling into the lead conversion.

Topic Survey

1. Problem Understanding

Loans can only be disbursed after ascertaining the quality of customers as risks are involved. Hence, the quality of leads needs to be thoroughly analyzed. For the profitability of the financial institution, it is also necessary for them to understand whether the existing customers can be leads for other financial products as well.

By optimizing marketing campaigns with predictive analytics, the Bank can generate new customer responses or purchases more accurately. Predictive models can help the Bank to attract their future most valued customers.

2. Current Solution to the problem

Manual intervention in lending loans to applicants leading to an excessive use of resources by the business.

3. Proposed Solution to the problem

Algorithm and ML driven loan approval and lead conversion of probable applicants.

4. References to the problem

https://www.researchgate.net/publication/267864165_Loan_Default_Prediction_on_Large_Imbalanced_Data_Using_Random_Forests

https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

https://www.fool.com/the-ascent/personal-loans/articles/7-factors-lenders-look-considering-your-loan-application/

https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook

https://towardsdatascience.com/predict-loan-eligibility-using-machine-learning-models-7a14ef904057

Critical Assessment of Topic Survey

By analyzing the already existing customer/applicant database, the bank can generate more business in terms of providing further financial services. This is a calculated approach as the customer loan history and several other factors are taken into account.

An individual's Monthly Income, Employment Status, City Category, credit history, loan amount, type of loan and risks associated with it seem to be important aspects of lead conversion.

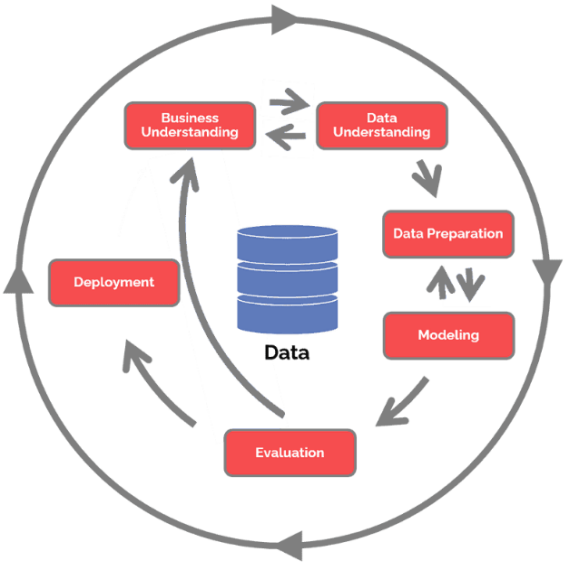
It is important to have a complete database of all the crucial information needed to approve a customer's loan. Having incomplete data and erroneous data would interfere in predicting the customer segment important to the bank.

Methodology

The entire project has been adapted to the data mining life cycle as under the methodology of Cross-Industry Standard Process for Data Mining (CRISP-DM) process model.It's a flexible and easily customizable model used for both modeling as well as data exploration and visualization in identifying particular patterns.

The six phases of CRISP-DM data life cycle are as follows -

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



Since the problem is based on a typical classification model,we have used classification techniques like logistic regression and Decision Tree while dealing with the model. Our final model is a stacking ensemble model using base learners Decision Tree and Gaussian Naive Bayes and the final estimator is Logistic Regression.

The main catch of iterating and selecting the final model was based on the ability of the model in maximizing its performance in terms of its Sensitivity (Recall) of the approved class without compromising on the model accuracy. We managed to get a good trade-off between the above parameters. By doing so we managed to reduce the false negatives in our final model to its minimalistic range which seem to be the most dangerous predictions as they would result in losing the lead for the bank.

All of the above mentioned parameters were obtained where the two classes were predicted quite distinctly by our model, with an ROC-AUC score above 0.8, Recall above 80% for the Approved class and an Overall Accuracy of around 70%.

Data Description

Our data consisted of 69713 rows and 22 columns initially. The target variable is ‘Approved’- whether a loan has been approved (1) or not (0).

The other variables are :
ID, Gender, DOB, LeadCreationDate, City_Code, City_Category, Employer_Code, Employer_Category1, Employer_Category2, Monthly_Income, CustomerExistingPrimaryBankCode, PrimaryBankType, Contacted, Source, Source_Catgeory, Existing_EMI, Loan_Amount, Loan_Period, Interest_Rate, EMI and Var1.

Data Preparation and Preprocessing

- Imbalanced Target Variable (Approved: 0, 1)

One of the limitations we face with our dataset is that the target variable is highly imbalanced (98.5% : 1.5%) . To overcome this, we explored oversampling techniques like SMOTE (Synthetic Minority Oversampling) and using the class weight ‘Balanced’ parameter in the classification models.
SMOTE balances the class distribution by randomly increasing the minority class sample (Approved: 1) by replicating the data.
The class weight ‘balanced’ automatically adjusts weights inversely proportional to class frequencies in the input data.
We used the class weight ‘balanced’ parameter in our models as this was a more suitable approach as we were focusing on reducing our false negatives and getting a better recall as per our business problem.

- Missing Values

Our data had significant no. of missing values in it, mainly for the columns of interest rate, loan period, loan amount and emi which had 40% of the data as null or missing. We could not think of dropping the null values because of the percentage of total data it beared, so we thought about mean, median imputation but it would create so many similar data points i.e around 40% of the total data would have the same value of certain columns.

So, we decided to give this responsibility to our machine itself, we used iterative imputer to deal with such an enormous amount of missing values, and we used k-nearest neighbor as the estimator for iterative imputation.
It is a strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion.

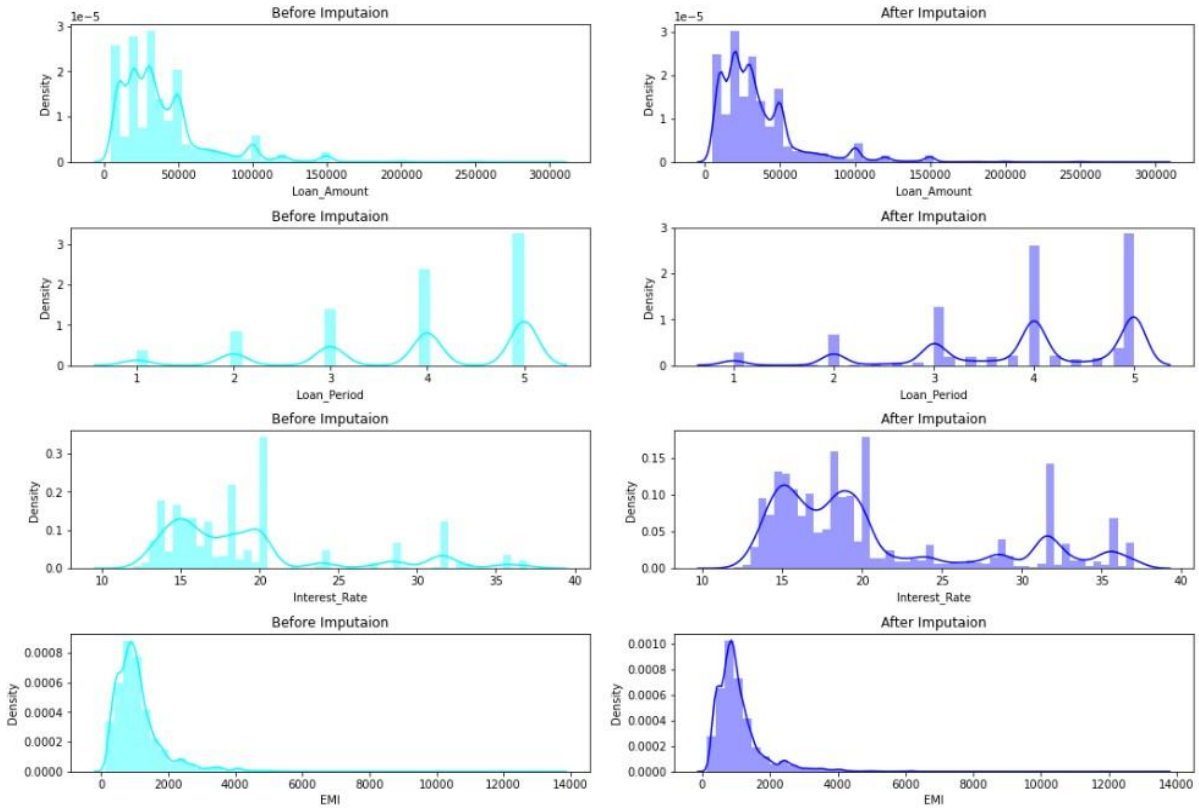
We imputed 3 columns using iterative imputer namely, loan amount, loan tenure,interest rate and calculated emi based on the formula stated below:-

$$E = P \times r \times \frac{(1+r)^n}{(1+r)^n - 1}$$

Where,

E is the EMI
P is the principal amount
r is the monthly rate of interest
n is the number of months

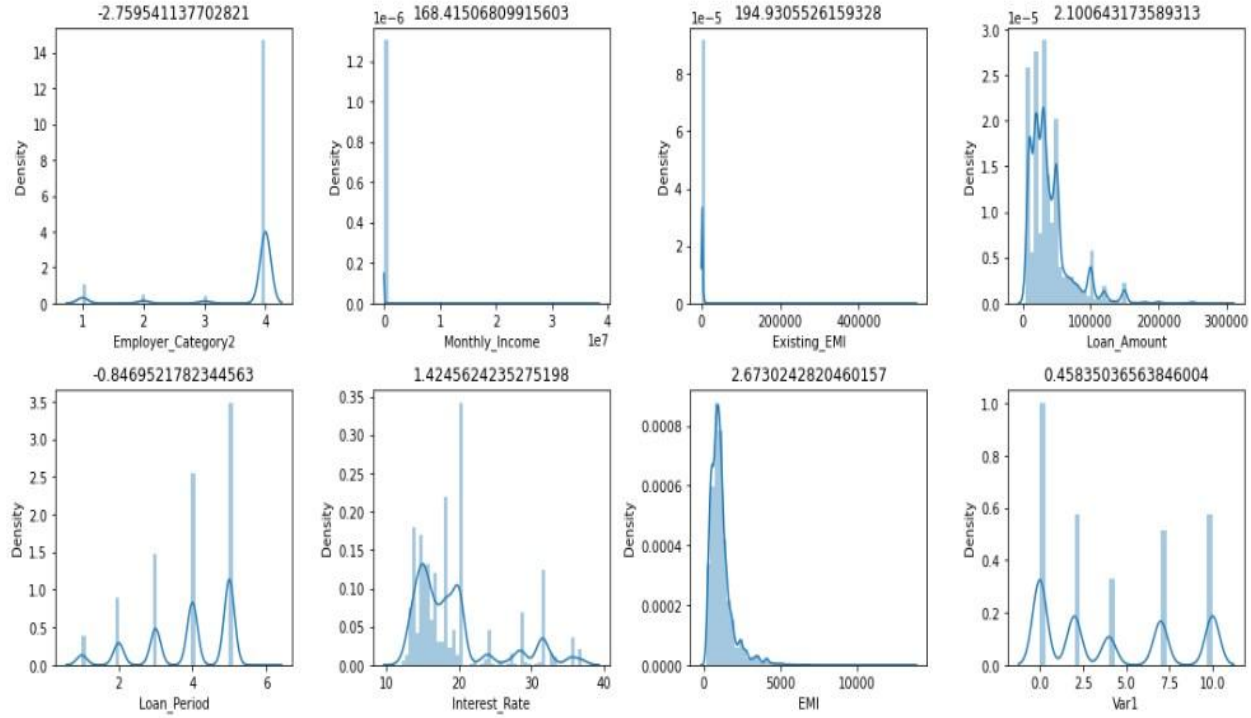
Distributions after imputation:



As we can see, iterative imputer imputes the missing values between the available values and does not have any significant effect on the original distributions of the features.

Exploratory Data Analysis

- **Numerical Variables**



Monthly Income

- It is highly right skewed.
- It might be because of few very high income leads.

Employer_Category2

- Most of the leads seem to be in category:4 , followed by category:1.

Existing EMI

- It is highly right skewed.
- Most of the leads don't have any existing Emis.

Interest Rate

- It is right skewed.
- Most of the leads are to be offered an interest rate between approx 10%-25% per annum.
- It is having a range from 10%-40% per annum (*approx).

EMI

- It is right skewed.
- Most of the leads are to be offered EMI below 5000 Dollars.

Loan Period

- Most of the leads have requested for a repayment tenure of 5 years followed by 4 years.
- Other Repayment tenures are for 1 , 2 , 3 years respectively according to the increasing density of leads.

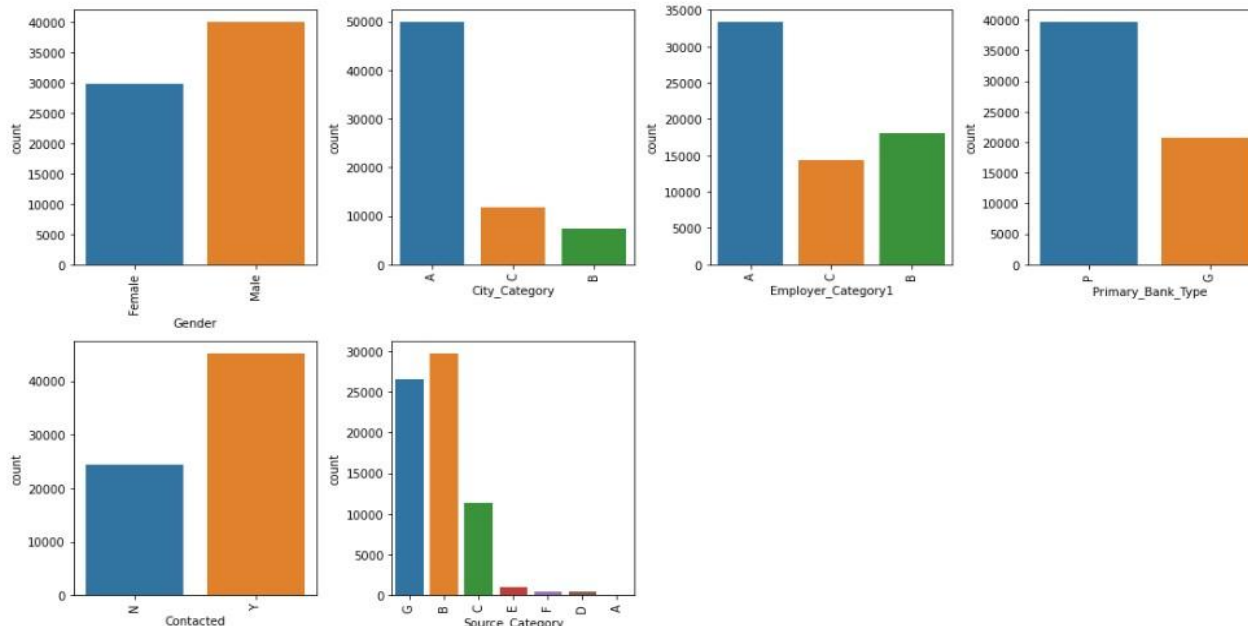
Var1

- This seems to be an anonymous rating given to the leads by the bank.
- Most of the leads are given a rating of 0.

Loan Amount

- It is highly positively/right skewed.
- The loan amount requested by most of the leads is less than 100,000 dollars.

- **Categorical Variables**



Gender

- More than 50% of the leads are Males.

City Category

- Most of the leads fall in city category:A followed by C.

Employer Category 1

- Most of the leads have employer category as A followed by B.

Primary Bank Type

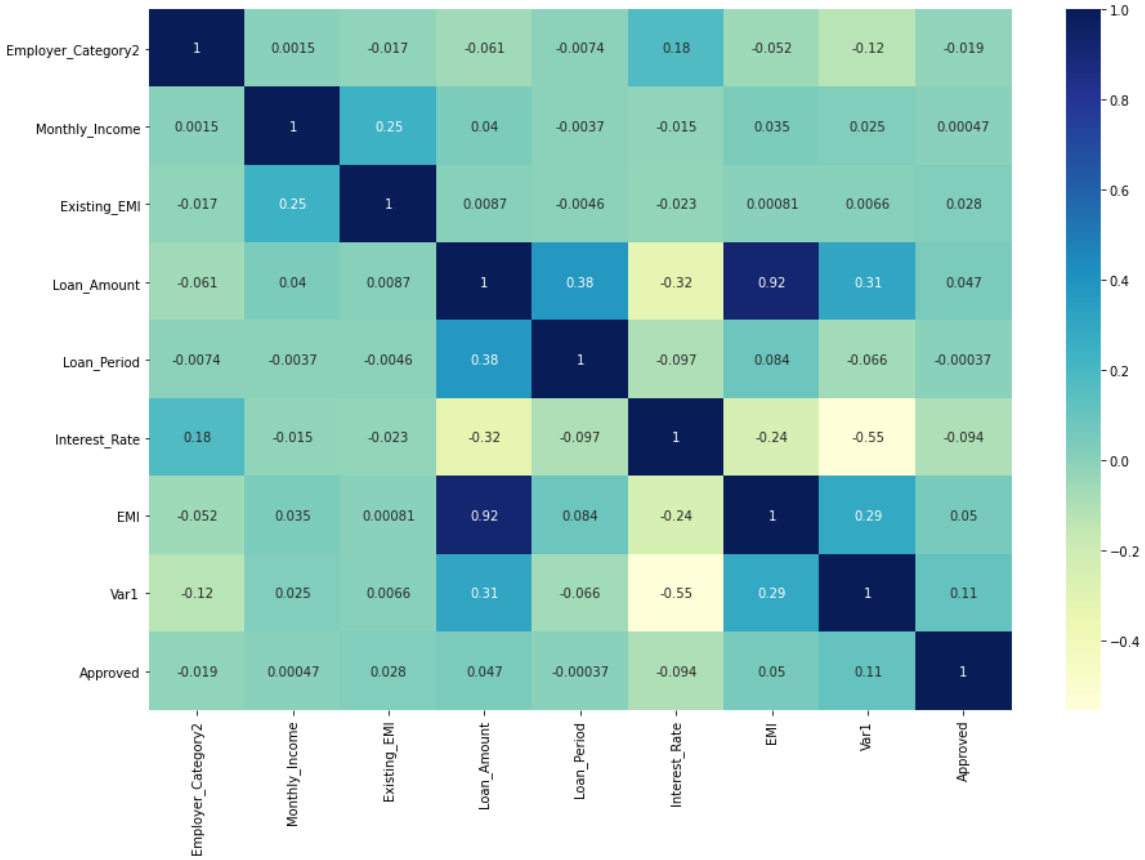
- As per our understanding the two categories (P and G) can be interpreted as PRIVATE and GOVERNMENT
- Approx more than 60% of the leads are having their primary bank type as PRIVATE.

Contacted

- Approx more than 55% of the leads have been contacted by the bank.

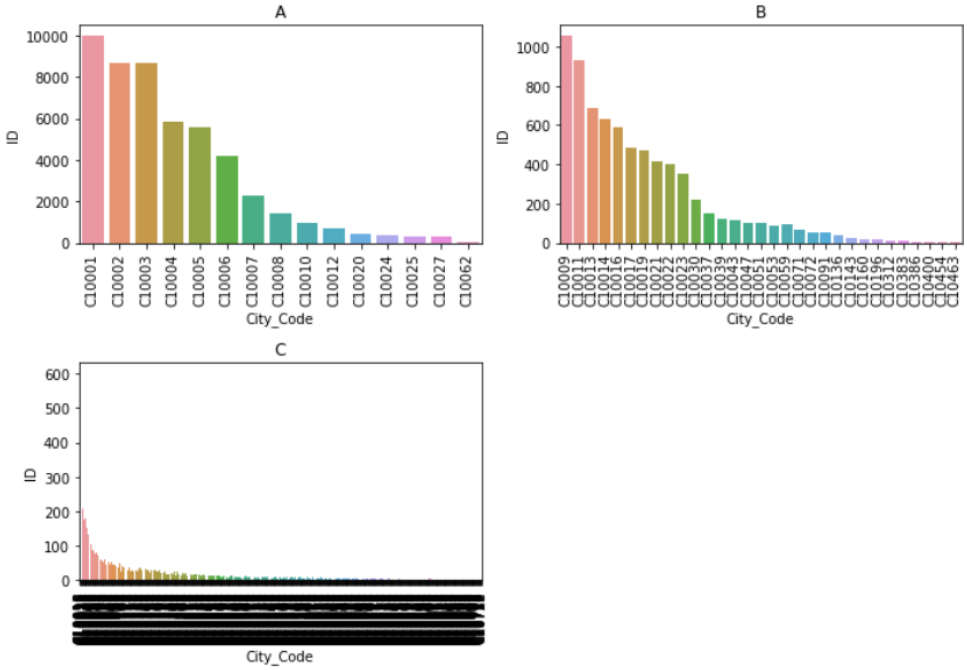
Source Category

- Most of the leads are in source category B, G, C respectively as per increasing density of leads.

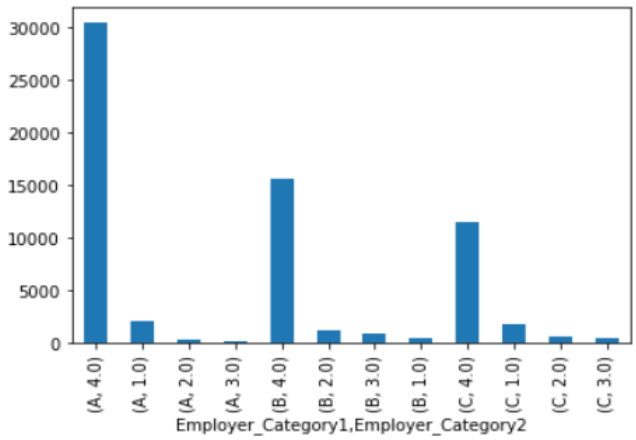


- Emi and loan amount have strong positive correlation
- var1 and interest rate have negative correlation

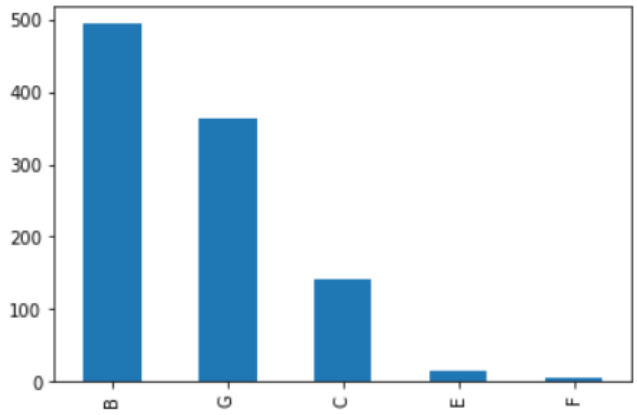
Key Insights from EDA



- City Category C has the highest number of city codes followed by City Category B and the least city codes exist in City Category A.
- However the density of population in City Category A is the highest followed by B and C.



- Across employer category 2, the 4th category has the highest number of leads



- Most leads are converted from Source Category B, G and C

- Var 1：Better Credit Rating has more chance of lead conversion.

It seems to be the credit score of leads as understood by statistical tests. It has a significant relation with Monthly Income, Interest Rate, Loan Period and Existing EMI which are all factors taken into consideration while assessing a lead's credit history.

Feature Engineering and Extraction

- We manipulated the DOB and Lead Conversion Date to obtain the approximate age of the lead and further tagging it into an Age_Cateory column.
- We re-calculated the EMI basis the emi calculation formula and used to impute the missing EMI values.

$$E = P \times r \times \frac{(1+r)^n}{(1+r)^n - 1}$$

Where,

E is the EMI
P is the principal amount
r is the monthly rate of interest
n is the number of months

- We created 4 Business Quarters as per the Lead Creation Month, to tag leads and their respective conversions segregated based on the quarterly business of the bank.
- We indulged in creating a scale of business columns namely Employer_Cat3, which depicts the volume of applicants coming from a certain Employer Code into 5 categories.
- We further created two features namely Monthly Income to EMI ratio and Loan to Income Ratio as is needed for the acceptance of the loan by the bank. Higher the value of the former and lower the value of the latter seem to have a better chance of loan approval.

Hyperparameter Tuning

Hyperparameters in Machine Learning Algorithms help us tailor a model to a specific dataset. They are given explicitly as inputs to the model.
It is challenging to understand which hyperparameters would suit a particular dataset and thus, Grid Search technique is used to find the best parameters.

For our dataset, in the Decision Tree Algorithm we gave hyperparameters such as criterion, max_depth, max_features and min_samples_split to the grid search algorithm.

The best parameters were criterion: entropy, max_depth : 8, max_features : 15 and min_samples_split : 1250.

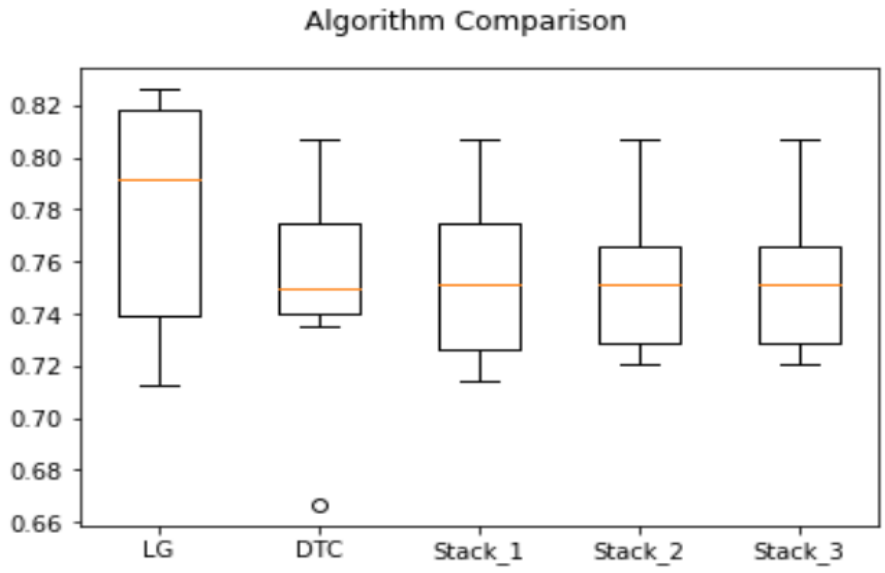
Using the best parameters we proceeded to tune our Decision Tree and got an accuracy score of 0.68 and recall 0.77 for class 1 on our test data.

Modeling, Model Comparison and Final Model Selection

After conducting wide research on loan prediction models used in the financial sector, we decided to use the Logistic Regression Algorithm as our base model. It is a classification algorithm that is widely used in such business cases.

Our Intermediate Model was made using the Decision Tree Algorithm which we tuned as per our dataset using the Grid Search Strategy.

Our final model was the Stacking Classifier Algorithm using base learners - Decision Tree,KNN Classifier, Gaussian Naive Bayes and final estimator- Logistic Regression. This model helped us reduce the false negatives and also gave us a good overall accuracy.



The models we finally selected were stack 2 and stack 3 which were made using the Stacking Classifier. These models gave us a high recall and the standard deviation of scores was the least among them.
In the Logistic Regression model, although our recall was high, the standard deviation was also high which meant the scores were not consistent. In the Decision Tree model, we could see an outlier having a very low recall and thus, we chose the Stacking classifiers 2 and 3 as our final models.

Conclusion and Recommendations

Monthly income, employment status, city category, credit history, loan amount, type of loan and risks associated with it seemed to be important aspects of lead conversion as these were the important features given to us by our models.

These features are also relevant in the real world while assessing whether a customer's loan should be approved or not and finally, whether the customer is indeed a potential lead.
False negatives in this business case are dangerous. A false negative would be an actual lead for the business but the model unfortunately does not predict the customer as a potential lead. This would be a cost for the business and thus, reducing the false negatives here is of utmost importance.

In the real world, working with a database with so many missing values and erroneous data would be a huge risk to the business and thus, collection of data and entering the data accurately into the database would also need to be monitored for accurate predictions.

References

https://www.kaggle.com/arashnic/banking-loan-prediction