Data Science Project-Bank Loan Defaulter Prediction

For the purpose of this project, we have taken the roles of consultants who are working with a financial services company that provides personal loans, to develop a model that will help predict the likelihood of default for the company's customer portfolio. We have used the models logistic regression with lasso, random forest ,XG boosting and gradient boosting and evaluated the model that gave the **best accuracy and profitability**. In terms of software, we have used both R and Python.

Part 1: Business Understanding

(a) **Business problem**: Banks gain their revenue from sanctioning loans. A tedious task before approving a loan is running background checks to ensure that the customer has the willingness and ability to pay back the borrowed amount in the stipulated time. This process is tedious but is a very important step that cannot be compromised. When a customer applies for a loan, the financial service provider ideally collects a required checklist of data from the customer. This data is used to build the model to predict the likelihood of a customer paying back a loan.

(b) **Goal:** The objective of this study is to identify different patterns which would help indicate to the company if a client will have difficulty paying their loan installments, and to strategize what actions to take depending on whether they have a higher risk of defaulting on their agreement. **Our main objective is to maximize profit for the end user** (In most cases, it will be a financial services company/loan provider).

(c) **Addressing the business problem/ deployment**: The model provided by this study will primarily aid the end user with a more *clear and efficient method of accepting or denying a loan application*. In addition to this, this model could be further branched out towards reducing the amount of a loan, assessing the benefit of refinancing/restructuring the loan, determining the optimal interest rate, making impairment provisions for loans and balancing the risk of the loan portfolio although this study is not extended to cover it.

Part 2: Data Preparation

(a) **Data cleaning:** Upon our initial analysis of the dataset, we have identified the columns that have a large number of null values. We have brought the total 122 variables in the data set down to 60 variables in order to conduct further analysis. We have eliminated the columns that have more than 40% of their values as null (such as YEARS_BUILD_AVG that has 70% null values). We have dropped the column SK_ID_CURR which is the unique identifier of each record because it will not aid with the prediction and has no relationship with the target variable. Further, we have also eliminated any additional columns which we believe won't aid us in predicting (based on intuition) such as NAME_TYPE_SUIT. In addition to this, certain columns which are useful for prediction such as CODE_GENDER have a few N/A values, due to which we have dropped those specific rows from the final dataset entirely altogether. We have also made a dataset more convenient for analysis by converting certain variables to a more interpreter-friendly format. For instance, DAYS_BIRTH

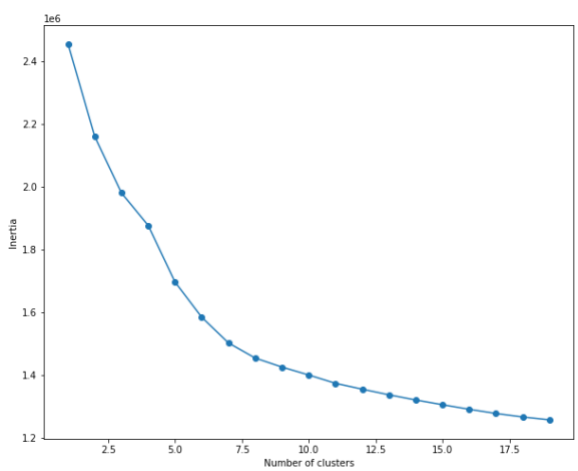
column has the total number of days a person has been alive, which we have divided by 365 to convert to the number of years (i.e age). We have also assigned dummy variables to all the categorical variables that will aid us in prediction. There are also pairs of variables such as loan amount and annuity which have a high correlation with each other. In such instances, we have decided to eliminate the column that has the weaker correlation with the target variable. We understand that in the real world there might be merit in keeping a few of these columns, but for the purpose of this project, we have dropped columns based on our judgement and analysis.

(b) After cleaning the data and filtering the columns that will be helpful in prediction, we have normalized the data in order to provide more consistency for executing the next steps.

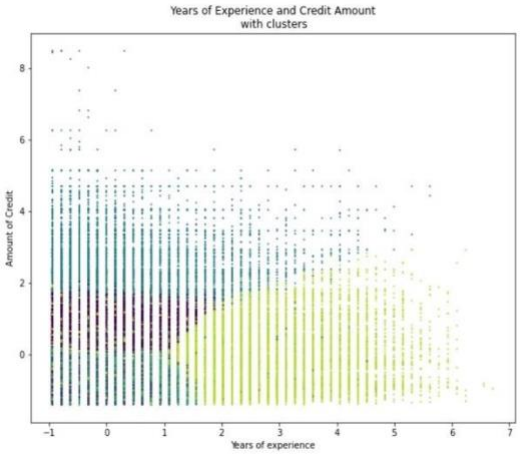
Part 3: Data Understanding

In order to understand the data better, we have performed a k-means cluster analysis. We have chosen this method over PCA , as the dataset has many categorical variables for which the PCA method would not be suitable for. We identified a separate method using the package "FAMD" in R in order to do a principal component analysis for datasets with a large number of categorical variables, however upon doing this no visible conclusions were drawn as the dimensions of the dataset were not reduced too significantly. The principal components explained a very small portion of the dataset (E.g. PC1 Represented only 1.5% of the dataset) thereby we have chosen k-means as a method to understand the dataset.

(a) **Representative – performing k-means cluster analysis**: In order to understand the data better, we decide to deploy a more sophisticated version of k-means method to do classification on both numerical and categorical variables. The method is called *k-prototype*, a clustering algorithm for mixed data types. As in this project, we have numerical variables such as AMT_INCOME_TOTAL, standing for income of the client, and categorical variables like NAME_CONTRACT_TYPE. First, we check the data type of each individual column using df.info(), finding some categorical variables are of the wrong types denoted as 'int64', from which it should be changed into the 'object' data type. Once changed, the data is all set for k-prototype implementation now that categorical and numerical variables all have the right data types. Second, we find the optimal number of clusters using the Elbow Method, with the scree plot showing the inertia Vs the number of clusters. Based on this, we have drawn the conclusion that the optimal k value is 10 which is used to calculate the centroids



derived from the k-prototype method, ensuring that there is no over fitting or underfitting of the prediction model.



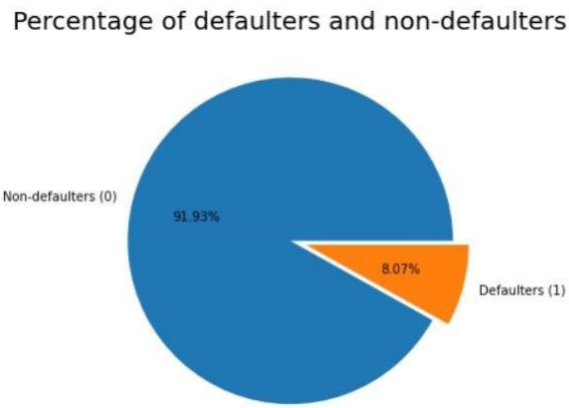
Running a k-means cluster analysis has enabled us to infer certain subsets in the dataset. For instance, the graph on the left shows the clear distinction between consumers years of experience Vs their amount of credit. Based on this graph, the below clusters can be derived:

a. Consumers with a high amount of experience and low credit amount

b. Consumers with a low amount of experience and low credit amount

c. Consumers with a low amount of experience and a high credit amount

(b) **Limitations of the model**: The original training dataset that was used to build this model is imbalanced. 92% of the model consists of non-defaulters while only 8% consists of defaulters. Having this imbalance in the dataset will make training a model more challenging due to it being biased towards the majority class (in this case, biased towards non-defaulters). However, we have decided to not balance the dataset by oversampling due to it being an arbitrary assignment of whether a borrower defaults. Our main priority is maximizing profit for the firm. We believe the company should make investments in acquiring this data from its' clients. The loan default predictor model is only as good as the data that's used to train it, and thereby **it's crucial that it has a sufficient amount of balanced and unbiased data to ensure a good accuracy figure and to maximize profit.**



Part 4: Exploratory Data Analysis

We used Exploratory Data Analysis to develop statistics and visualizations to analyze and identify the trends in data sets. Since there were numerous rows and the data is extensive, we used EDA to help us gain better understanding of the data. The first step was to determine the range and trends within the categorical variables. We created a list of the categorical variables and used a histogram for each variable. Diving into the details of each of these variables enables us to understand the target demographics and what the market is for this usecase. For example, Figure 1 below would tell us the type of occupations that the sample of loan applicants have. It is clear that majority of the applicants in this dataset are laborers. Figure 2 draws insights into the

customers previous loan history, and it is evident that majority of this sample had previous loan applications that were approved. Figure 3 shows the split between types of loans in the dataset. We can draw the conclusion that majority of the loans belong to the cash loan category.

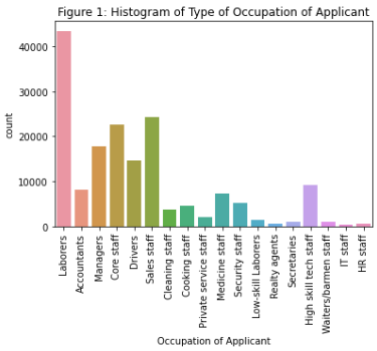


Figure 1: Histogram of Type of Occupation of Applicant

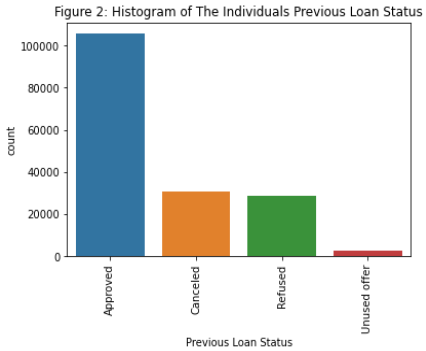


Figure 2: Histogram of The Individuals Previous Loan Status

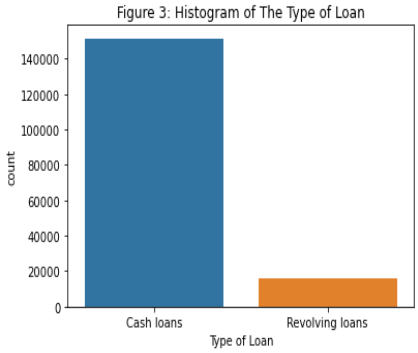


Figure 3: Histogram of The Type of Loan

Next, we tried to find relationships between different variables and their effects on the target variable. Figure 4 shows how many customers fall in to different education levels. It's evident that most customers have had at least a secondary/secondary special education. Figure 5 shows the occupation type for the default status, although this data is heavily skewed towards the target variable being 0 (as the dataset is biased). Finally, figure 6 shows the split between defaulters and non-defaulters based on the previous loan status (whether or not it was approved).

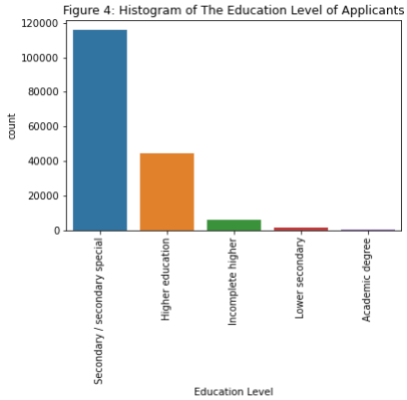


Figure 4: Histogram of The Education Level of Applicants

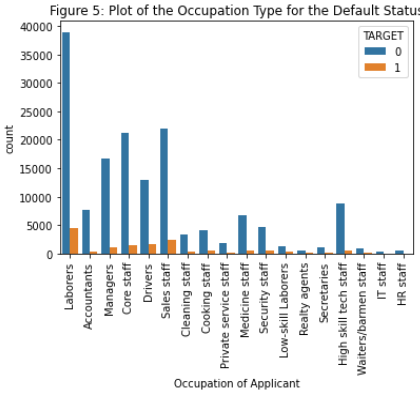


Figure 5: Plot of the Occupation Type for the Default Status

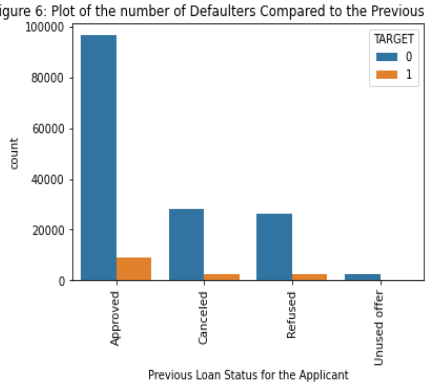


Figure 6: Plot of the number of Defaulters Compared to the Previous Loan Status

The below plots (figure 7) show the type of organizations the defaulters and non-defaulters work in.

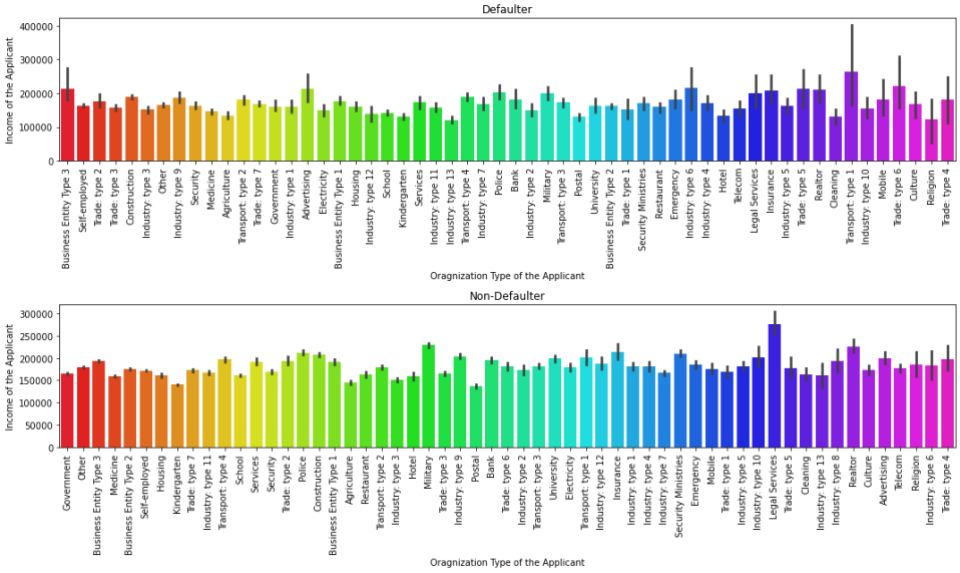


Figure 7: Comparing Organization Type and Income for Defaulters and Non-Defaulters

People working in advertising, transport and trade are more likely to default than those working in Legal Services, Military, Police, Ministries, Bank. Hence, the latter form better options to sanction loans to.

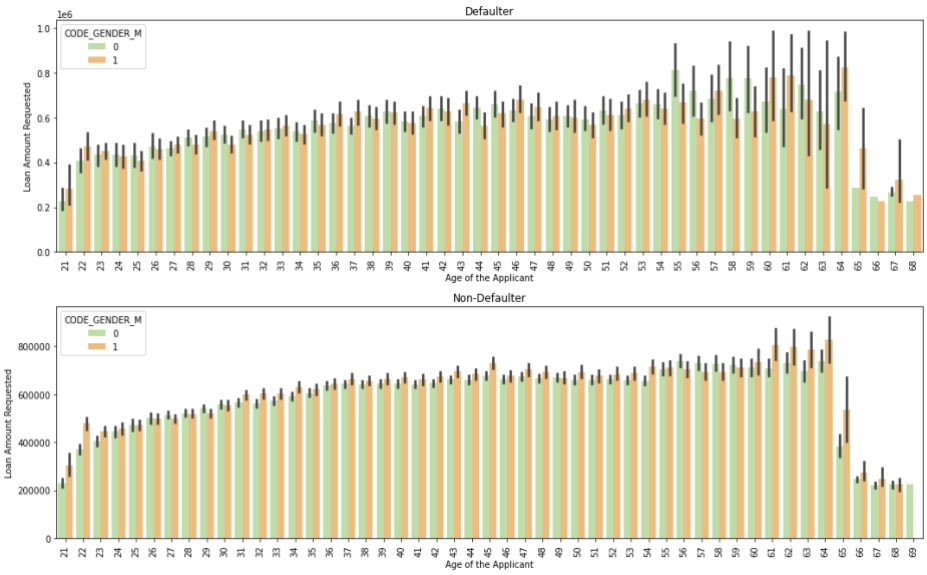
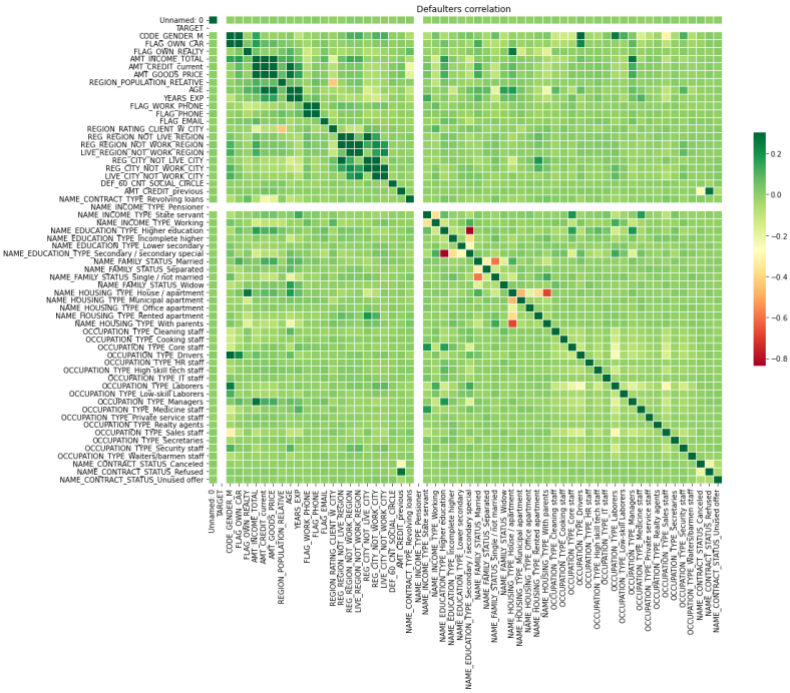


Figure 8: Comparing age group and amount of credit (amount loan applied) for defaulters and non-defaulters categorized by male and female.

From this plot we see that middleaged people taking higher credit are more likely to pay back their loan (3050) whereas adults above 50 years of age, on a higher credit, are riskier.

Inference from EDA:

1. Banks should focus more on income type 'Working' as they are the largest group and pay their loan back in a timely manner.
2. People in the age group 30-50 can be focused on because they have a good history of repaying bank loans.
3. People working in advertising, transport and trade are more likely to default than those working in Legal Services, Military, Police, Ministries, Bank. Hence, the latter form better options to sanction loans to.
4. People having high income and owning car should not be given higher credit amount.
5. Of all the occupation types, we see that laborers and drivers were the most likely to default.



From the correlation matrix we can see that our dataset does not have many variables that have high correlation but we can infer the following: Most Laborers and Drivers are Male. Managers earn more than most of the other occupations. A loan is most likely refused when they have requested for a really high loan amount previously. People with high income, own a house/apartment as compared to renting/living with parents.

Part 5: Modeling Framework

Since this is a loan default prediction, we have a classification problem on our hand. **Our Target Variable is the Defaulter Status (if a person will default or not).** Our uncertainty is whether they are going to pay us back or not. What we can control and hence our decision would be if we accept their loan application or deny it. From

Methodology	**Logistic Regression with Lasso**	**Random Forest**	**Gradient Boosting**	**XG Boosting**
Metric used to evaluate the models	Mean AUC : 0.673 Accuracy score : 0.911042	Mean AUC : 0.649 Accuracy score : 0.911020	Mean AUC : 0.646 Accuracy score : 0.911037	Mean AUC : 0.682 Accuracy score: 0.911047
Decomposing the problem	Our goal is to maximize profit and hence, we have divided the problem into 2 steps: 1. We are using the model to provide the probability that an individual will default. So, we are doing a probability estimation. 2. We categorize each individual into 3 buckets of low, medium and high risk to determine the interest rates in each category. 3. Combine the above two statistics into the given profit equation explained in part 6			
Recomposing a solution	We can use the different interest rates on the dataset to calculate the average gain and loss margins. This can be plugged into the profit equation. We can now set different thresholds and find the value for which we can maximize profit.			

our EDA and Data Understanding procedures we noted that there are no set of features that help explain the target. Hence, we would need to use most of the features to determine the target. Additionally, we have not considered models with linear objective functions (e.g SVM) as there exists no strong linear relationship between most of the variables and the target variables as seen by the correlation plot.

After testing out the models logistic regression with lasso, random forest and gradient boosting, we have picked the **XG Boosting** model as our final model based on the accuracy and the AUC metric in order to calculate the profit. All four models perform the same in terms of accuracy, therefore we have used the **AUC** metric to determine the best model and subsequently calculate profit for the firm.

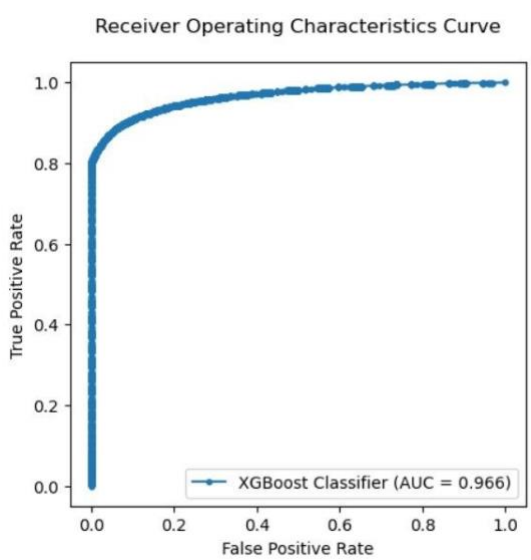
(a) Hyperparameter Tuning using K – fold cross validation:

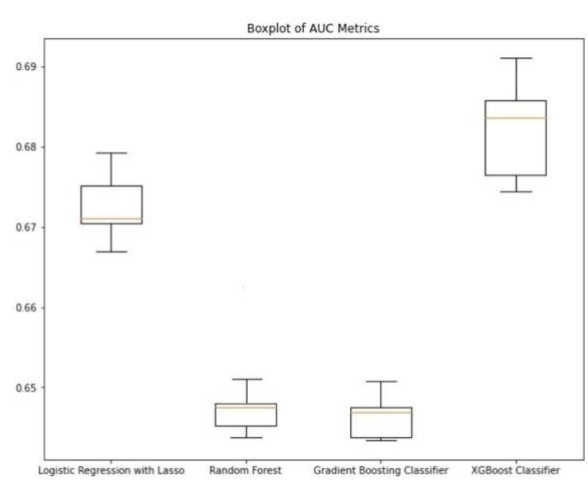
	model	best_score	best_params
0	random_forest	0.911042	{'class_weight': {0: 0.1, 1: 10}, 'n_estimator...
1	logistic_regression	0.911020	{'C': 1}
2	gradient_boosting	0.911037	{'learning_rate': 0.01, 'n_estimators': 50}
3	xg_boosting	0.911047	{'gamma': 10, 'learning_rate': 0.3, 'n_estimat...

We have performed a *grid search* on all four chosen models, which is a cross validation method in order to run all the defined combinations of the estimators. For instance, for the random forest model we have defined the number of estimators (number of trees) as 50,100,150 and 200 and defined the class weights as {0:1,1:1}, {0:0.1,1:10}, {0:1,1:100}, {0:0.0001,1:1000}, {0:0.1,1:10000}. What this means is that the grid search will perform iterations of the random forest model on the dataset using all the above combinations such as 50 trees with class weight {0:1,1:1}. This grid search is run on the random forest model, Logistic regression with lasso, gradient boosting and XG boosting in order to determine the best combination of estimators for the models. Running the grid search will then give the below output:
According to the grid search, these combinations of estimators yield the best possible results for each model.

(b) Modelling and Prediction:

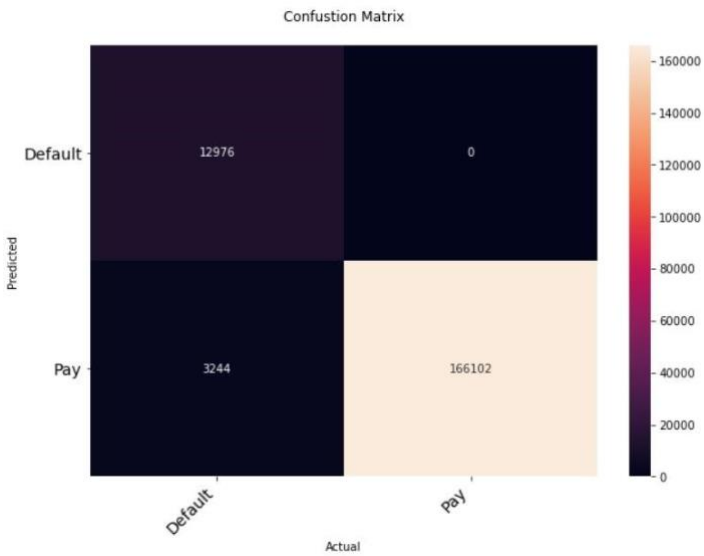
We have used the estimators from the Hyperparameter tuning to further proceed in determining the model with the best results. We then perform k-fold cross validation with 5 splits using the above parameters for each model. Python's function to run logistic regression automatically applies to the penalization parameters. We have picked the **XG Boosting** model as our final model based on accuracy and the AUC metric in order to calculate the profit.





The XG Boost Classification model has an AUC of 0.966. Hence, we can conclude that it works well after 5-Fold Cross Validation.

Part 6: Evaluation



(a) Confusion matrix

The confusion matrix is computed basis the actual data for the target variable and the model predictions. We see that 12976 are the number of predicted defaulters and they actually defaulted. Hence, our true positives, TP=12976. Our model does not misclassify any person who will default, hence our FP is 0. Our model has classified 3244 people who actually defaulted as FN= 3244. We also have a true negative I.e actually defaulted and the model also predicts default TN= 166102.

(b) True positive rate and false positive rate:

$False\ positive\ rate = \frac{False\ positive}{All\ negative}$ = 0/(0+166102) = 0%.

$True\ positive\ rate = \frac{True\ positive}{All\ positive}$ = 12976/(12976+3244) = 80%

As per the confusion matrix, we have a true positive rate of 80% and a false positive rate of 0%.

(c)Choosing interest rate to calculate profit: in order to calculate the cost and benefit to the loan provider, we refer to the interest rate of loan from Fair Isaac Corporation (FICO) website. FICO credit score is the most common credit scores used by many lenders to determine if a person qualifies for a credit card, mortgage, or other loan in the United States. Originally, FICO credit score divides people into 6 categories whose interest rate ranges from 6.481 % to 8.07 % .

In our project, combing the above interest rate and available variables in our dataset, we use age and income to measure the credit level of clients.

- Income credit level (112500 is 1st quartile and 225000 is 3rd quartile):
  low: income<112500 **medium**:112500<income < 225000 **high**: income>225000
- Age credit level (40 is the median):
  low: age<40 high: **age**>40
  Based on these two factors, the credit levels will be
- **Low credit level**: low income + low age, low income + high age
- **Medium credit leve**l: medium income+ low age, high income +low age
- **High credit level:** medium income+ high age, high income +high age

We set 6% as the annual interest rate for clients with high credit level, 7 % as the annual interest rate for clients with medium credit level and 8% as the annual interest rate for clients with low credit level.

Monthly loan formula = $\frac{Loan\ Amount \times r \times (1+r)^{\cdot n}}{((1+r)\cdot n-1)}$:

Loan revenue: $Monthly\ loan \times 120$ （We assume the duration of loan for each client is 10 years)

(d)Cost and Benefit Matrix:

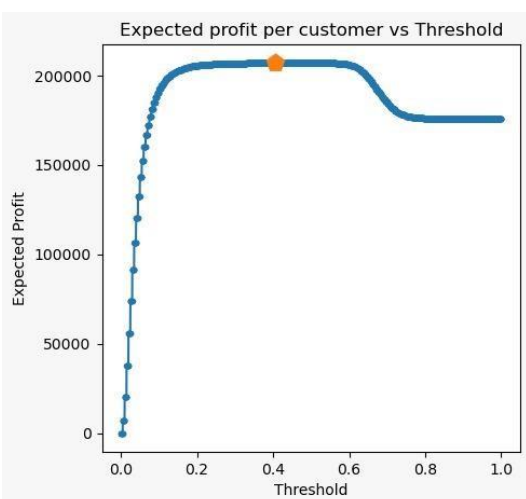
Predict \ Actual	Default	Pay
Default	0	0
Pay	-434395	235677

The average gain: $\frac{sum(Revenue - loan\ amount)}{number\ of\ client}$

The average loss: $\frac{sum(-0.7 \times Loan\ Amount)}{number\ of\ client}$

Based on our assumption on interest rate and loan duration, if we predict the client pay the loan but he/she actually defaults, the average amount that loan provider loss will be 434395. If we predict the client pays the loan and they actually pay it, the average amount that loan provider loss will be 235677. In the other two scenarios, the loan provider will gain/loss nothing.

(e)Profit curve



Profit equations:

[Expected profit per customer |No loan] = 0

[Expected profit per customer| Loan] = 235677X Probability (Repayment) – 434395 X Probability (Default)

Expected Profit (majority rule)

Based on our analysis we recommend that the user picks a threshold of **0.4** to maximize their profit. This suggestion takes into account assumptions as stated. **Hence, we will choose XGBoost Classifier model and the threshold of 0.4 in the deployment.**

(f) Baseline Comparison: We cannot estimate the baseline profit from the original dataset because there are too many arbitrary variables that we do not have information about, eg: interest rate and amount of collateral used. This is why we need to stick to our model metrics to determine the maximization.

Assumptions:

i. We have assumed a term of the loan is 120 months for the profit calculation.
ii. There is no reinvestment or refinancing risk for all the loans; the customer will pay all loans in due time and not sooner or later.
iii. For the purpose of this project we have bucketed the customers into 3 groups. However, we realize that there might be more parameters considered when picking an interest rate for a customer. We have assumed the following interest rates based on a combination of intuition and FICO rates. The break up has been mentioned in the Evaluation above.
iv. We have considered that the loss only comes from a defaulter. We are assuming all costs to be zero and the opportunity costs to be zero. Additionally, from the dataset we see that there is a chance that a person refuses/ does not accept the loan offer, in this case too, we consider that the cost is zero.
v. We have assumed that the customer who defaults pays 30% of their loan amount as they did not begin defaulting on their loan immediately. Hence, the loss is –0.7 times the loan amount per defaulting customer. vi. We have averaged out the gain and loss on the whole data set. We have used this estimate in our cost benefit matrix.

Part 7: Deployment

(a) Deploying the result of the analysis

This model computes a threshold in terms of granting loans based on given customer dataset. Banks can customize the threshold according to their own date so that they can generate most revenue while limit the risk of losing money on default by identifying trustworthy customers.

(b) Potential issues with regards to deploying the solution

This model is highly dependent on input data and can only do as good as the data provided. Therefore, it's necessary to guarantee the precision and relevance of data being used, otherwise the threshold calculated won't be of any merit for banks to maximize profits.

(c) Ethical considerations

Customer demographic data is sensitive and should be classified. For example, income and address should be handled with care and not be provided for mass use or go into the wrong hands.

(d) **Potential risks and mitigation plan :** Deploying a model as such comes with a certain degree of risk as the model is only as good as the data provided to train it. When implementing this at a financial services company, it is possible that there could be instances where the model falsely predicts that the consumer will not default, where in reality the consumer actually does default. How this can be mitigated is by tuning

the threshold to capture this requirement, taking in to account that there will be people who default when the model says they won't. While this risk won't be 100% managed by adjusting the threshold, it can most definitely improve.

Part 8: References

a. *Estimate your loan savings using credit score calculator* (2022) *myFICO*. Available at: https://www.myfico.com/credit-education/calculators/loan-savings-calculator/ (Accessed: October 16, 2022).
b. *Dataset used : https://www.kaggle.com/datasets/arkapravasen/bank-loandefault?select=application_data.csv*