Name: Pranav Reddy Pedaballe

Enrolment No: 23117099

Credit Card Behaviour Score Prediction Using Classification and Risk-Based Techniques Report

Objective: -

We are provided with historical data from bank A, comprising about 25,000 customer records with a labeled target next_month_default, which indicates whether a customer defaulted in the next billing cycle or not.

The goal of this project is to build a model which would accurately identify potential defaulters in advance, allowing the bank to adjust credit exposure, trigger early warning systems, and prioritize risk-based actions and thus identify who may default in advance.

The next_month_default column has two categories:

Default (1): - Indicates that the customer is more likely to miss their credit payment in the next billing cycle.

Non-Default (0): - Indicates that the customer is less likely to miss their credit payment in the next billing cycle.

We employ data preprocessing, EDA, feature engineering and evaluation metrics to improve and assess the performance of our model.

Procedure: -

1) Data Preprocessing: -

First, we drop the Customer_ID column as each Id is unique and will not help in the training of the model.

Next, we check if the values are correctly categorized.

In the marriage column there are 53 rows with the value 0, which does not correspond to any defined category, so we replace it with the value 3 , as 0 lies in the category of others.

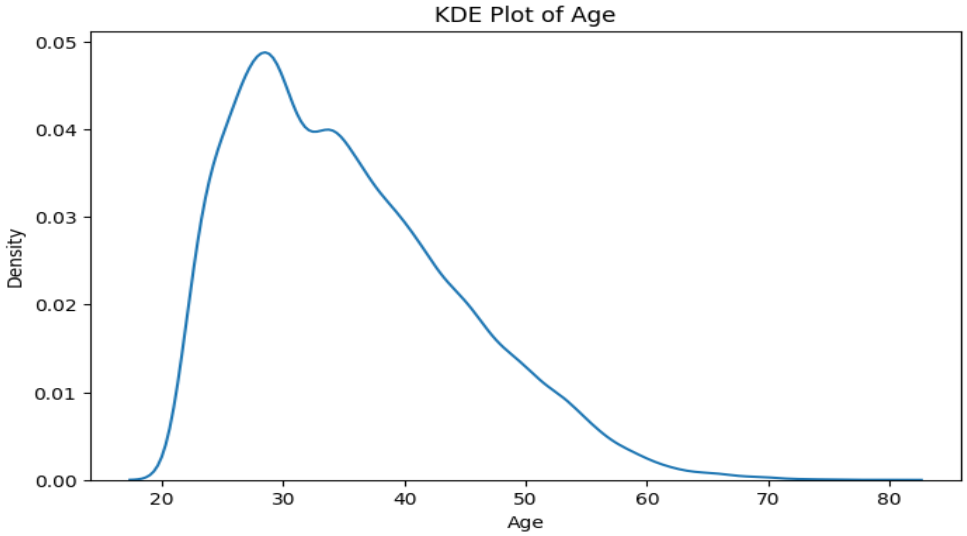
In the education column there are errors in the categorization, that is there are values other than the defined categories which are 0, 5 and 6. We replace these row values with 4 which stands for others.

Every other column has correctly categorized values.

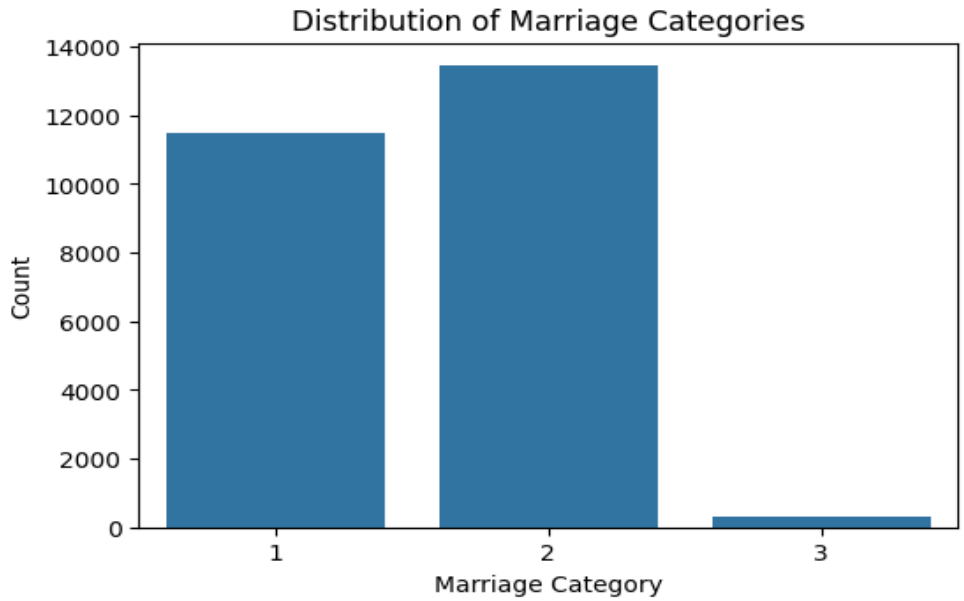
Next, we check for missing values.

We observe that the age column has 126 empty rows, so we replace them with the median of that column to preserve the data distribution.

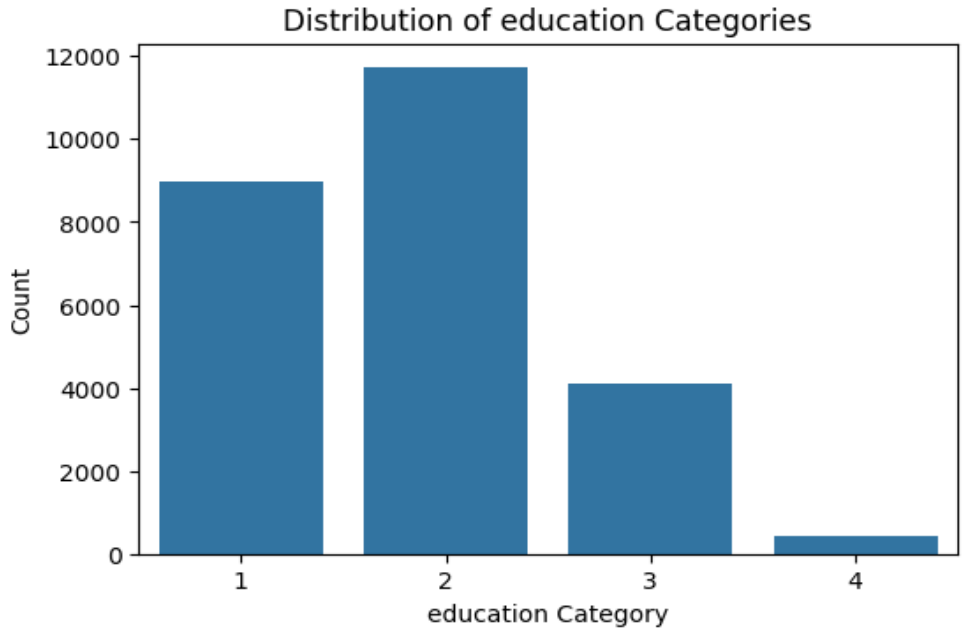
2) EDA of the given columns: -



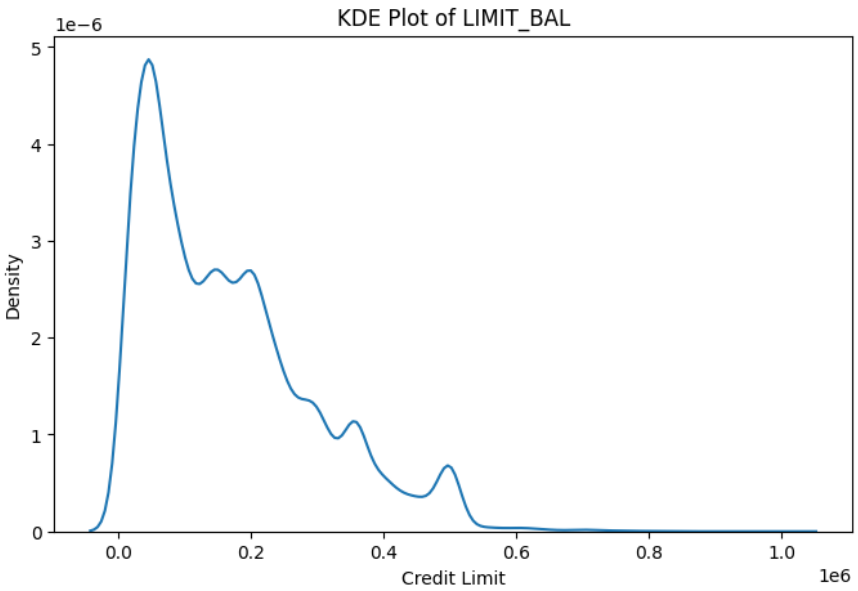
We observe that most customers who borrow from the bank are between the ages of 20 and 60 .



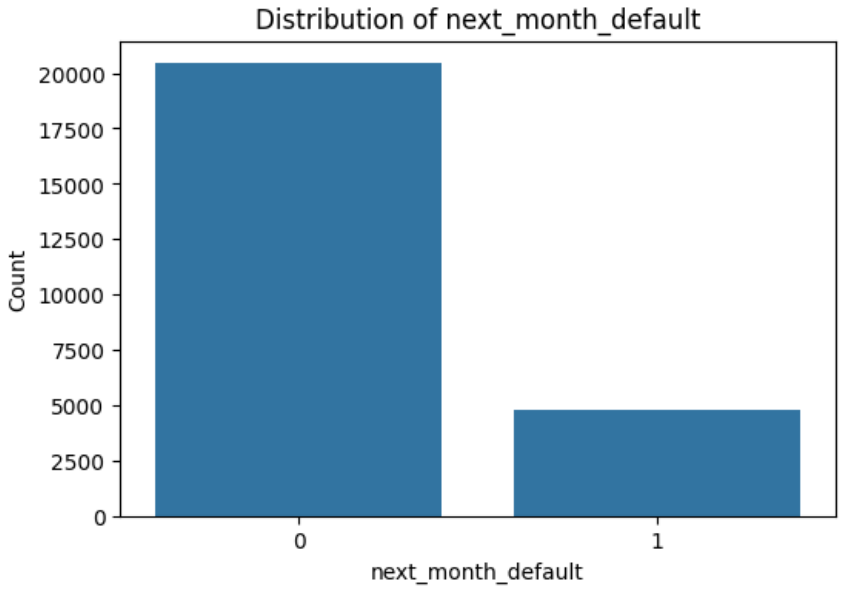
Customers who are married borrow more than single customers and very few in the others section borrow from the bank.



Customers who have an education level upto Graduate School and University borrow more from the bank than others.

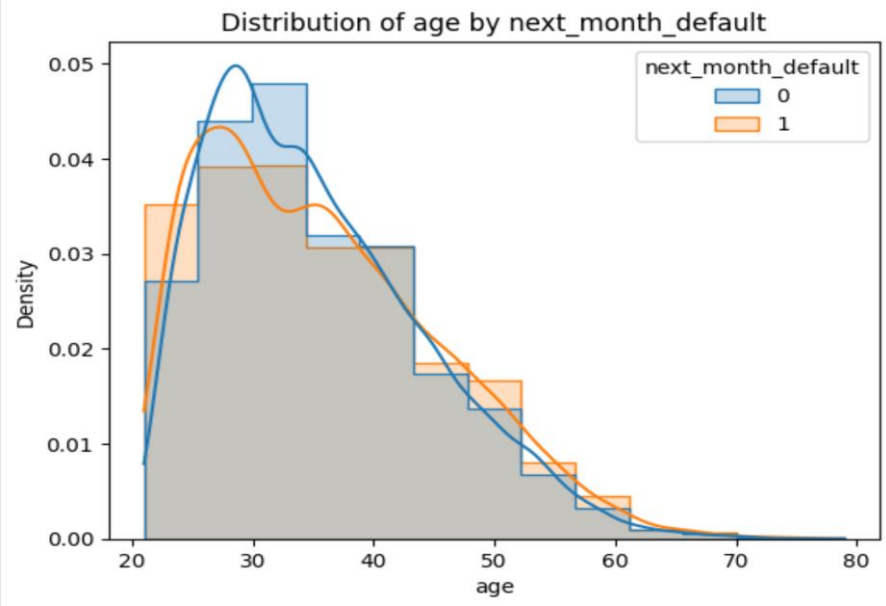


The bank generally gives a low credit limit to most customers and only few customers are given high credit limit.



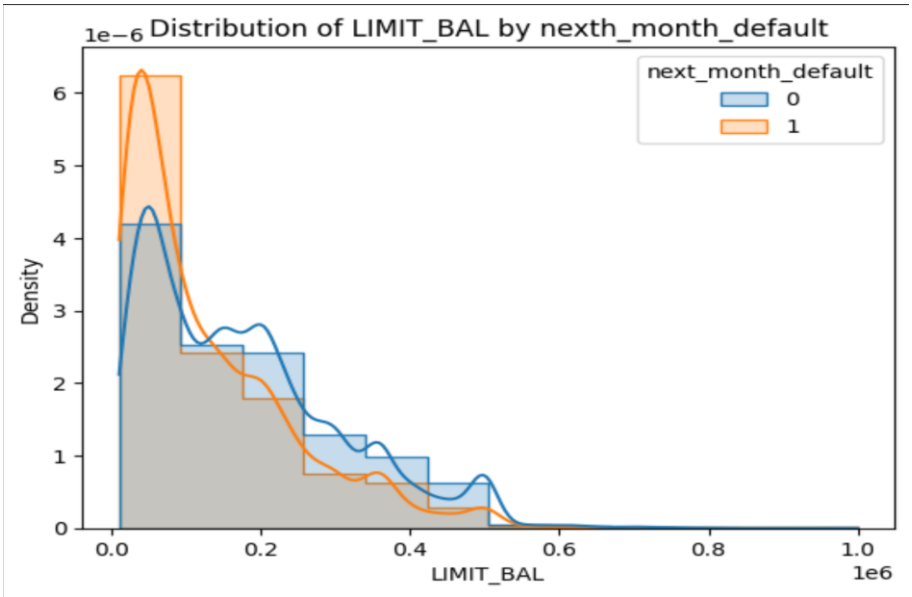
Approximately 5000 customers have defaulted from 25000 customers.

3) **Trends: -**

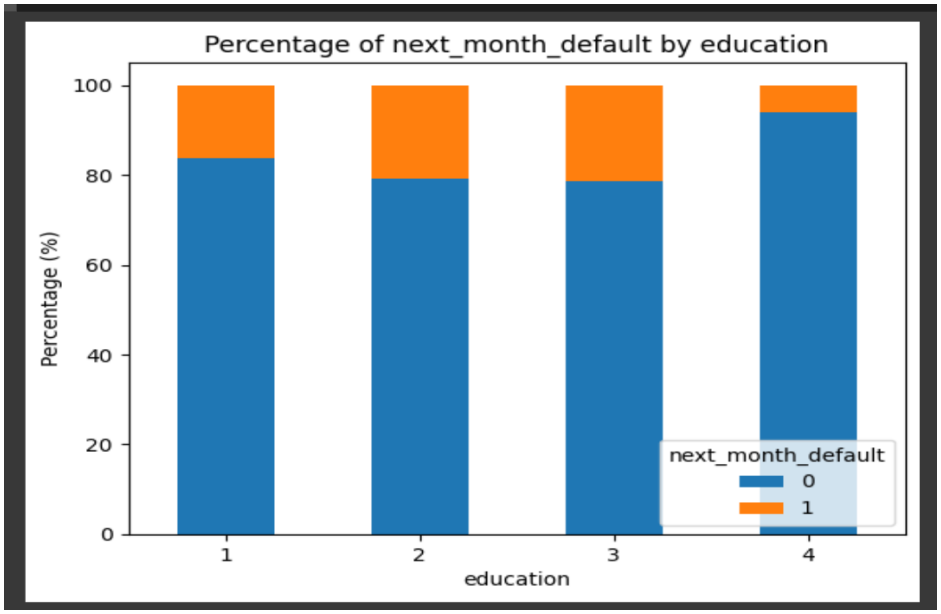


This is the plot of age vs next_month_default.

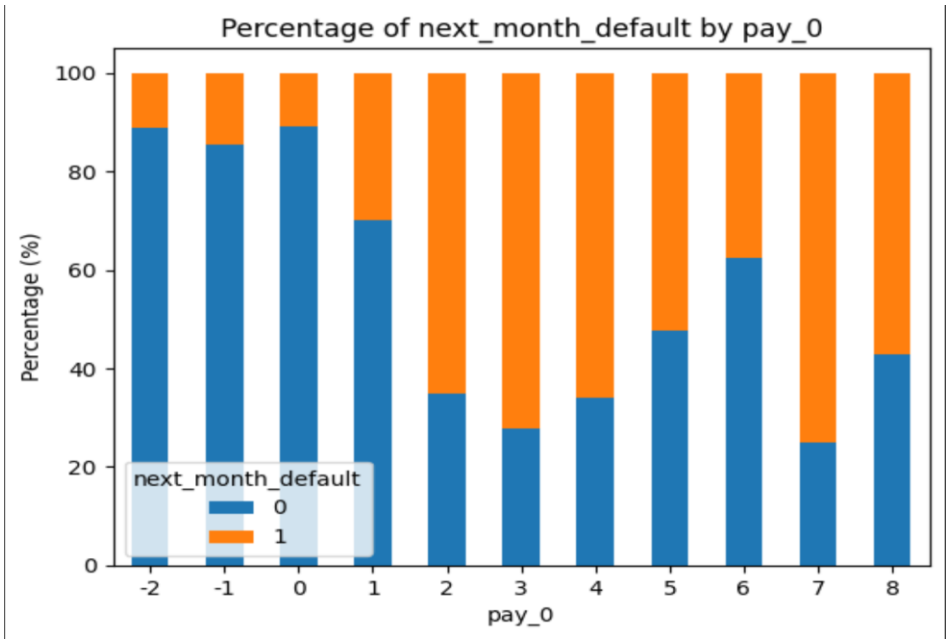
From this plot we can observe that people of the age below 25 and above 45 are more likely to default than the people between the ages 25 to 45.



We observe that the people who have credit limit of less than 150,000 are more likely to default than people with a higher credit limit.



This shows that people with education levels up to university and high school have more chance of defaulting and others have a very less chance of defaulting.



We observe that people with pay_0 = 1 have a higher chance of defaulting and those with pay_0 >1 have a very high chance of defaulting.

4) Feature Engineering: -

i. **bill_over_m{i}:** - For each month we check if the Bill_amt for that month is greater than the credit limit.

   Significance: - Exceeding the credit limit may imply risky financial behavior which may indicate the likelihood of defaulting.

ii. **months_over: -** The total number of months in which the customer has bill amount exceeded their credit limit.

   Significance: - A higher frequency of over-limit usage may signify a higher chance of defaulting.

iii. **months_billdue: -** the number of months bill is due, that is full amount is not paid yet (number of months Bill_amt is greater than 0).
   Significance: A higher count of months with unpaid bills may reflect prolonged credit usage or delayed repayments, which may imply the likelihood of dafulting.

iv. **payment_delay:** - the number of months for which minimum amount has not been paid (number of months for which pay_m > 0).

v. **max_payment_delay:** - the maximum number of months for which a customer has delayed the payment.

vi. **utilization_m{i}:** - Bill_amt{i} / Limit_Bal.

   Significance: - this shows how much the bill amount at the end of the month compared with the credit limit, if it's high, there may be a high chance of defaulting.

vii. **avg_utilization and max_utilization:-** the average and maximum of all the utilizations for each month.

viii. **paybill_ratio_{i}: -** the amount paid in the present month divided by the bill generated in the previous month.

   Significance: - if the paybill_ratio_{i} is low this means the customer is not repaying well which may lead to defaulting.

ix. **avg_paybill_ratio:-** average of all the columns.

x. **tot_paybill_ratio:** Total amount paid back / Total bill generated in the 6 months.

xi. **std_paybill_ratio: -** Standard deviation of the paybill_ratios.

Significance: - High standard deviation implies irregular repayment, which may imply defaulting.

xii. **Delinquency streak: -** Counts the number of maximum streaks in which a customer has not paid back.

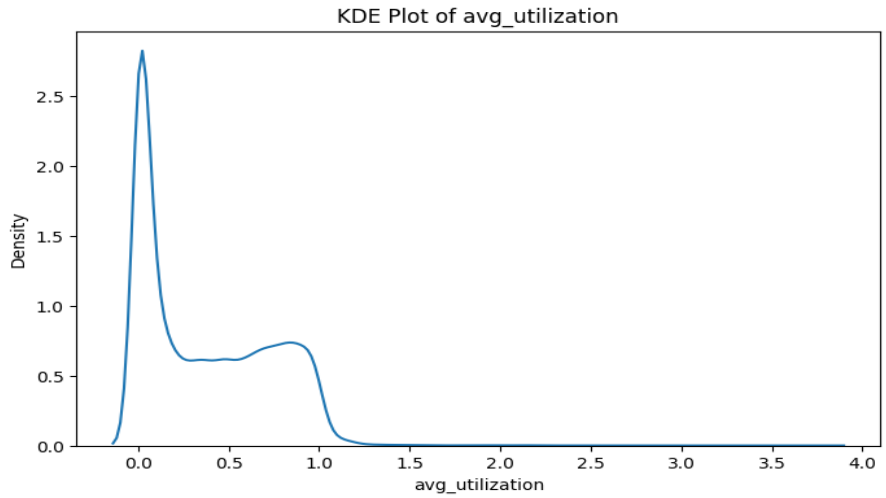
Significance: - If a customer has not paid back continuously for more than a few months there is a high chance that the customer may default.

xiii. **diff: -** The difference between the average of Bill_amts and the given AVG_Bill_amt .

Significance: - Ideally the values should be 0 or close to 0, but we are getting very high values for some rows, this shows that either the recorded Bill_amts were wrong or the average bill_amt recorded was wrong.

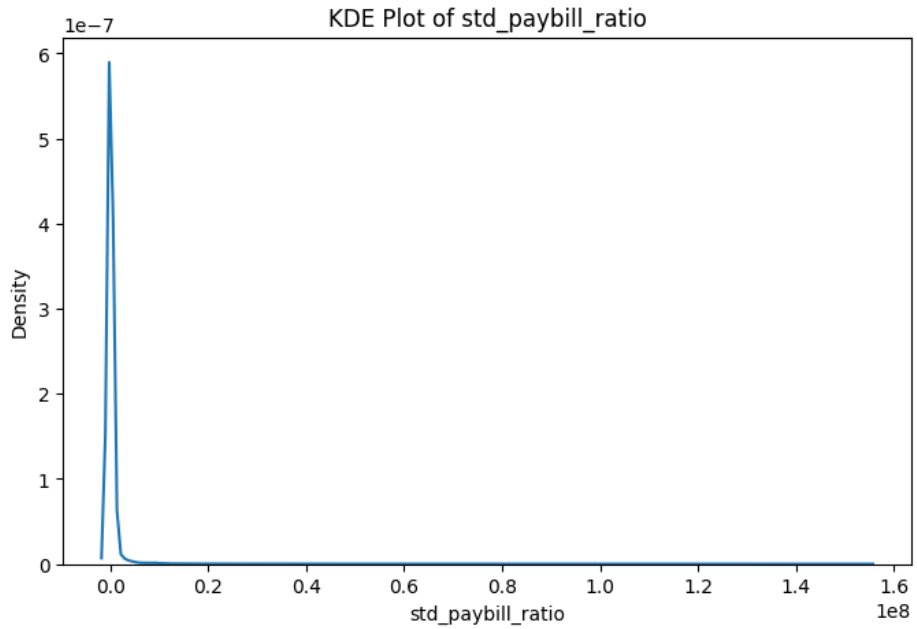
Wrong details may indicate a scam which may be the cause of defaulting.

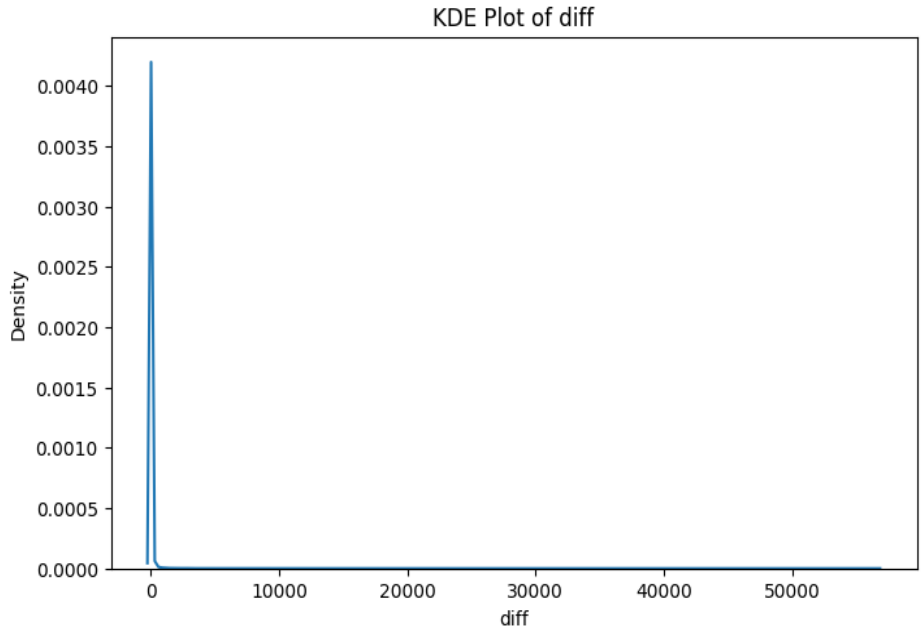
5) Trends in the new added columns: -

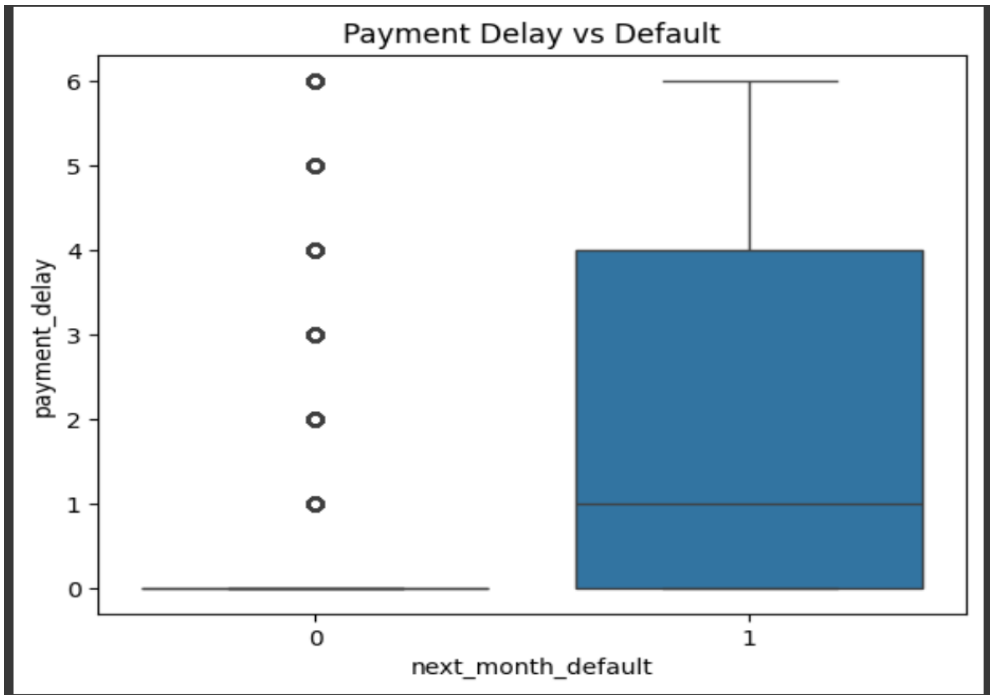


Most customers have an avg_utilisation <1 which means that few customers borrow more than the credit limit on an average.

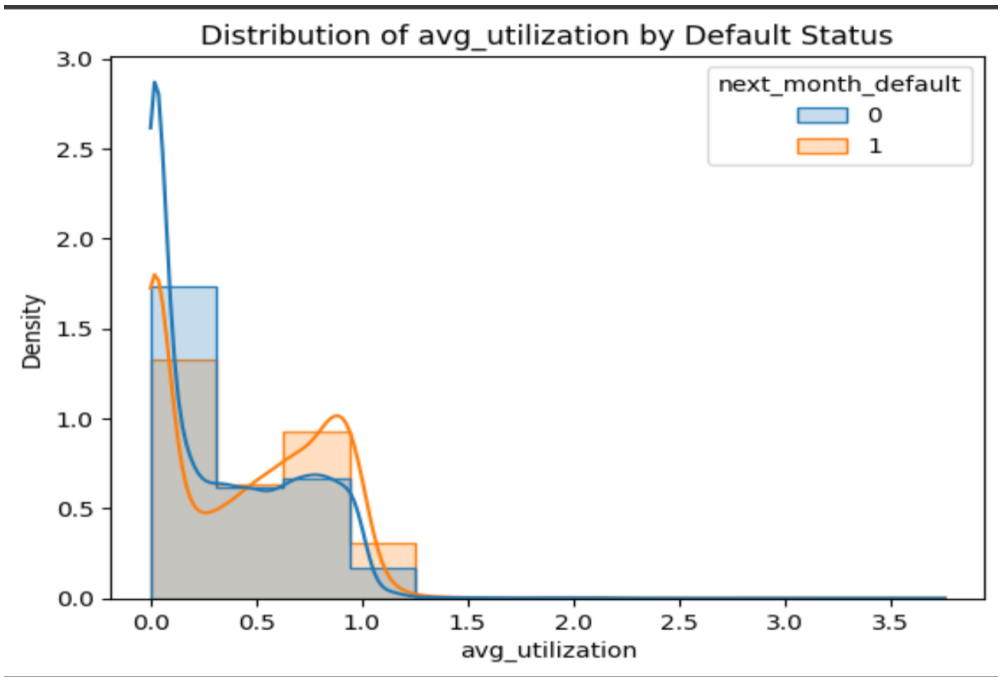
The kde plots for diff and std_paybill_ratio are similar, peak at 0 and very few customers for other values.



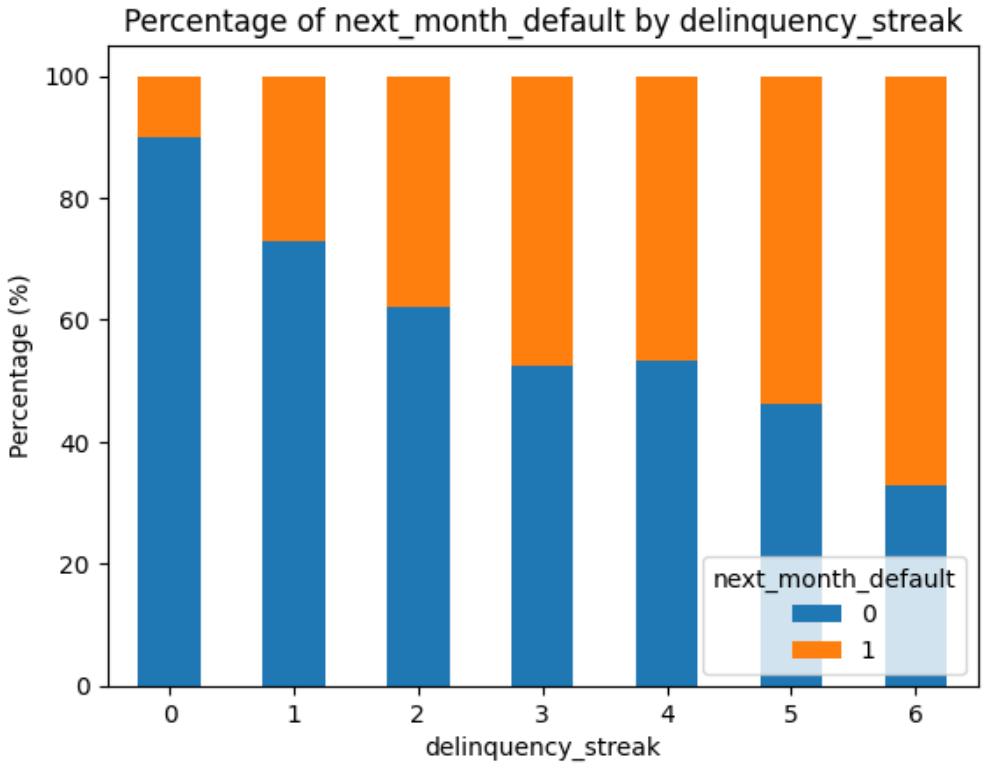




We observe that in most cases non-defaulters have a payment delay of 0 but defaulters have a general history of payment delays.



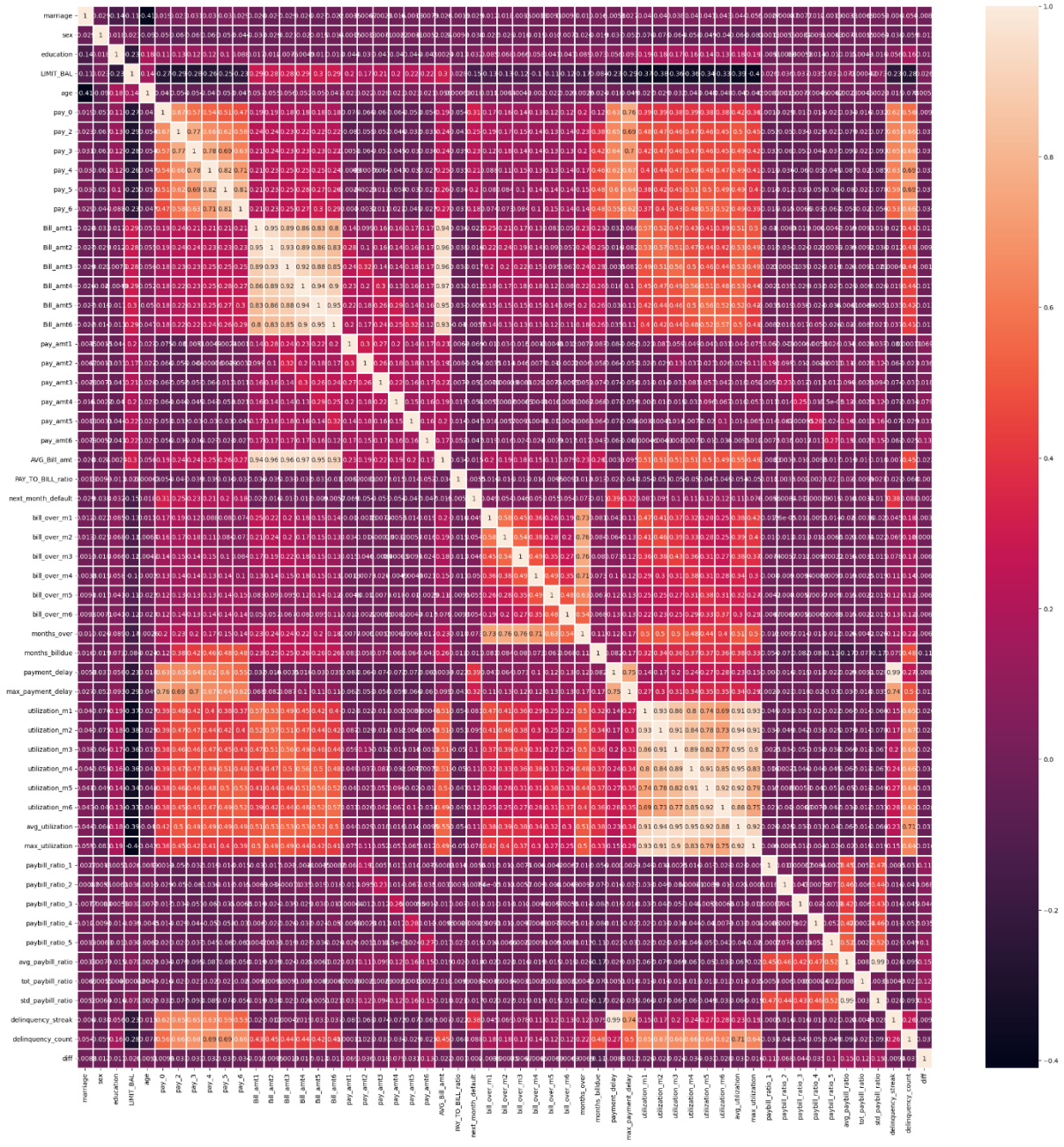
People with average utilization are greater than 0.5 are more likely to default. Average utilization is the average of each Bill_amt_m/LIMIT_BAL.



Higher the delinquency streak higher is the likelihood of the customer of defaulting.

6) Feature Importance: -

We create a correlation map, for every feature with each other.



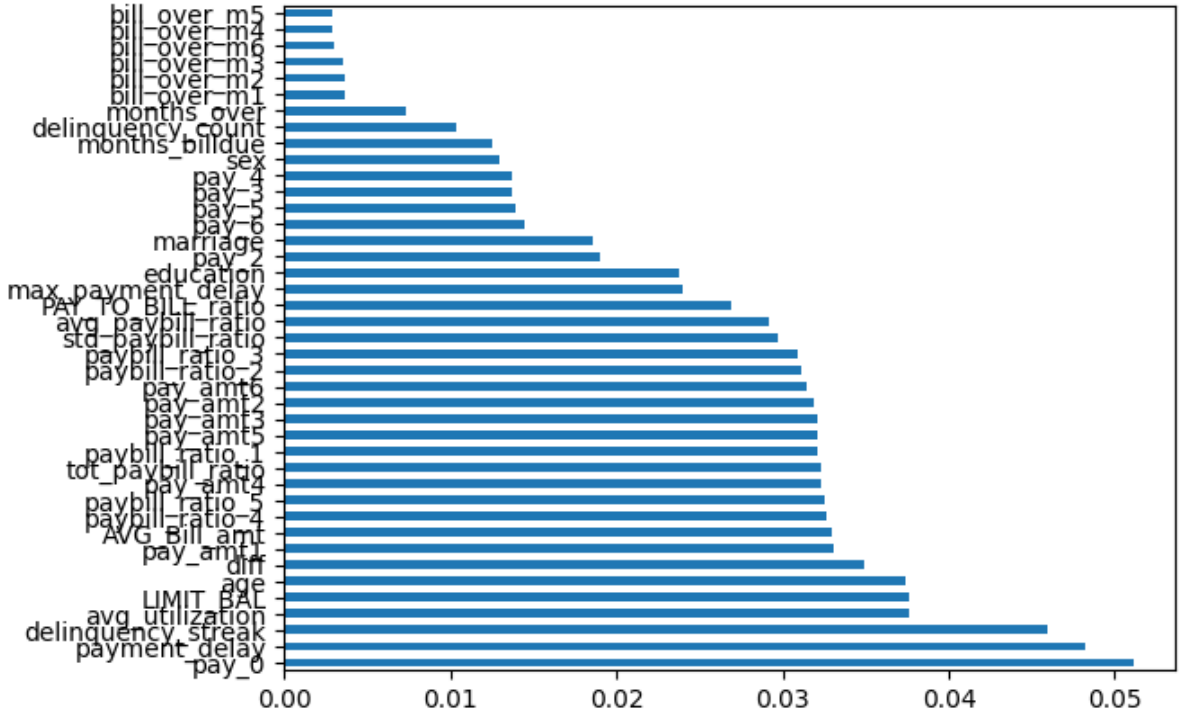
We notice there are very high correlation values between some columns.

To address multicollinearity which may negatively impact the model performance we drop features with a correlation coefficient greater than 0.9.

There is a high correlation between all Bill_amt{i} columns and the AVG_BILL_amt column, so we drop all Bill_amt{i} columns.

There is a high correlation between all utilization_m{i} columns and the avg_utilization column, so we drop all utilization_m{i} columns.

For checking the feature importances we use an extra trees classifier model and run the model and check the feature dependencies with the target variable.



The feature importance plot reveals a noticeable drop in feature importances in few columns.

We therefore drop all bill_over_m{i} columns and months_over , delinquency_count and months_billdue.

7) Class Imbalance and Training: -

We trained the data on 7 models, Logistic Regression,XGB Classifier, Random Forest Classifier, Light GBM, Decision Tree Classifier, Gaussian NB and K Neighbors Classification.

Then once we apply SMOTE + Class Weighting and another time we apply SMOTE + Downsampling.

In the case of credit risk detection banks generally want to predict who is more likely to default.

Thus, it is more acceptable to incorrectly classify a non-defaulter as a defaulter than to miss identifying a defaulter. The bank wants to catch as many defaulters as possible, that is we want false negatives to be less.

Recall = True Positive / (True Positive + False Negative)

Thus, we want the recall of class 1 to be high, but at the same time we cannot just identify all customers as deafulters. Thus, we try to maximise the F2 score metric.

F2 score prioritizes recall 4 times more than precision while still maintaining a balance between them.

Accuracy is not a suitable metric to use here as it does not reflect the model's ability to identify defaulters effectively.

Thus, we vary the threshold for each model from 0.1 to 0.9 with 0.01 difference and choose the threshold which gives the best F2-score.

After checking with 7 models with the two different ways of handling class imbalance we get the best model as Random Forest Classifier on the SMOTE + Class weighted data which gives the best F2-scrore of 0.6216 at a threshold cutoff at 0.28.

Other models give a lesser F2-score.

Metrics Result on Train Dataset: -

```
Accuracy :  0.6578217821782179

Classification report:
              precision    recall  f1-score   support

           0       0.93      0.63      0.75      4095
           1       0.33      0.80      0.47       955

    accuracy                           0.66      5050
   macro avg       0.63      0.71      0.61      5050
weighted avg       0.82      0.66      0.69      5050


Confusion matrix :
 [[2562 1533]
 [ 195  760]]

F2 Score: 0.6216
AUC-ROC Score: 0.7850
```

We save the above model as final_model.

8) Validation data: -

First, we save the customer_ID column as Customer_ID.

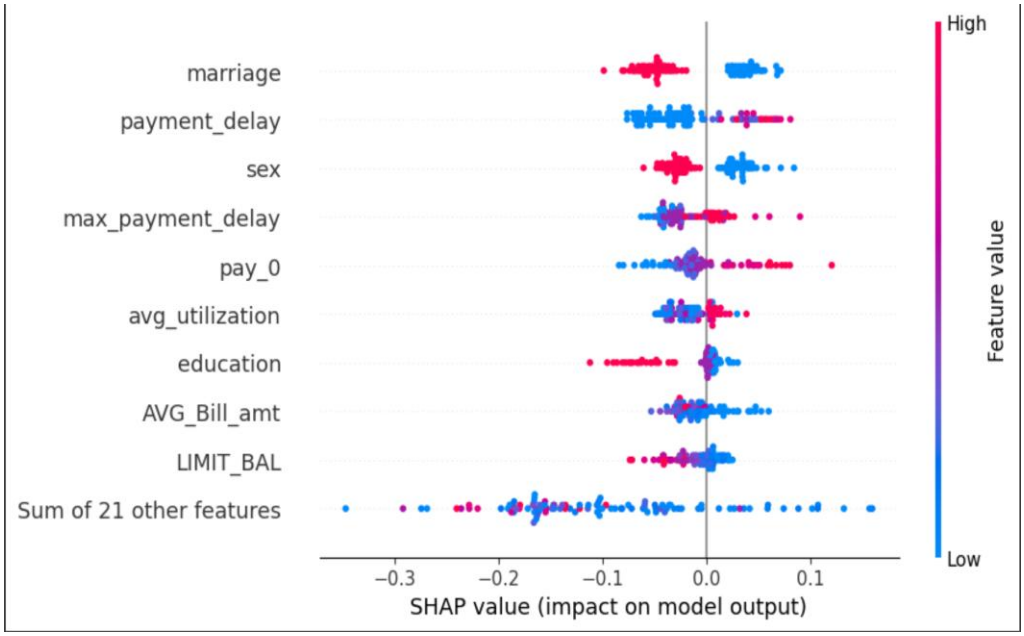
Then we remove the missing values, correct the data into correct categories and add and remove columns to make it the same as the columns on which the model was trained on.

Then we run the validation data on our final_model and save our results along with the respective Customer_ID to a new csv file submission_23117099.

We observe that we get our model predicts **3179 customers to not default** and **1837 customers are likely to default** in the validation dataset.

Inferences: -

SHAP was applied to the final model.



For marriage (1=married, 2=Single, 3=Others), we observe that married customers push the model to non-default whereas singles and others contribute positively to default prediction.

Payment_delay: - we observe that higher values influence the model prediction positively and contribute more to predicting default.

Sex: - Female customers contribute more positively to default risk predictions.

Similarly higher values of max_payment_delay , pay_0, avg_utlization, contribute to the model positively towards default.

Lower values of education, LIMIT_BAL and AVG_Bill_amt contribute to the model positively towards default.

Business Implications: -

The model flags predict customers as high-risk and low risk based on the data of the past 6 months.

False Negatives: - The customer has defaulted but was predicted not to default.

False negatives can cause high financial loss in the banks due to unpaid which the banks don’t want.

False Positives: - The customer has not defaulted but was predicted to default.

False Positives may cause reduced customer satisfaction and unnecessary checks on customers.

For a bank false negatives cause more harm than false positives, hence banks should focus on reducing the false negatives.

However, at the same time reduction of false positives should not be ignored, hence the F2-score metric was used.

Based on the predictions made by the model the bank could: -

- Lower the credit limit for high-risk customers.
- Reach out to the high-risk customers to recover the credit.
- Adjust the interest rates based on the default likelihood.

A transparent and interpretable model can help the bank justify their decisions and help satisfy the grievances of the customers and create a form of trust.

Summary and Key Findings: -

This project aimed to build a predictive model to identify potential credit defaulters using customer data from a bank. After data cleaning, EDA, and feature engineering, key indicators like age, credit limit, payment delays, and utilization were found to influence default risk. Class imbalance was addressed using SMOTE and class weighting. The final model was obtained with Random Forest achieving the best F2-score. The model was interpreted using SHAP. The model helps the bank reduce financial risk by identifying high-risk customers in advance.

Key Findings:-

1. **Age Impact**: Customers below 25 and above 40 are more likely to default.

2. **Credit Limit**: Customers with credit limits under 150,000 show a higher risk of default.
3. **Payment Behavior**: Those with payment delays (≥1 month) and high delinquency streaks are more likely to default.
4. **Utilization**: An average utilization above 0.5 significantly increases default likelihood.
5. **Marital Status & Education**: Single and less-educated customers tend to have higher default risk.
6. **Model Performance**: Random Forest with SMOTE + class weighting gave the best F2-score (0.6216), balancing high recall with acceptable precision.
7. **SHAP Analysis**: Key drivers of default include payment delay, low credit limit, high utilization, females ,high average utilization and low average bill amount.