PREDICTION OF CREDIT RISK USING MACHINE LEARNING MODELS

Project of Nguyen Vo Phuong Thao

GROUP 1- MENTOR NGUYEN HUU HOP

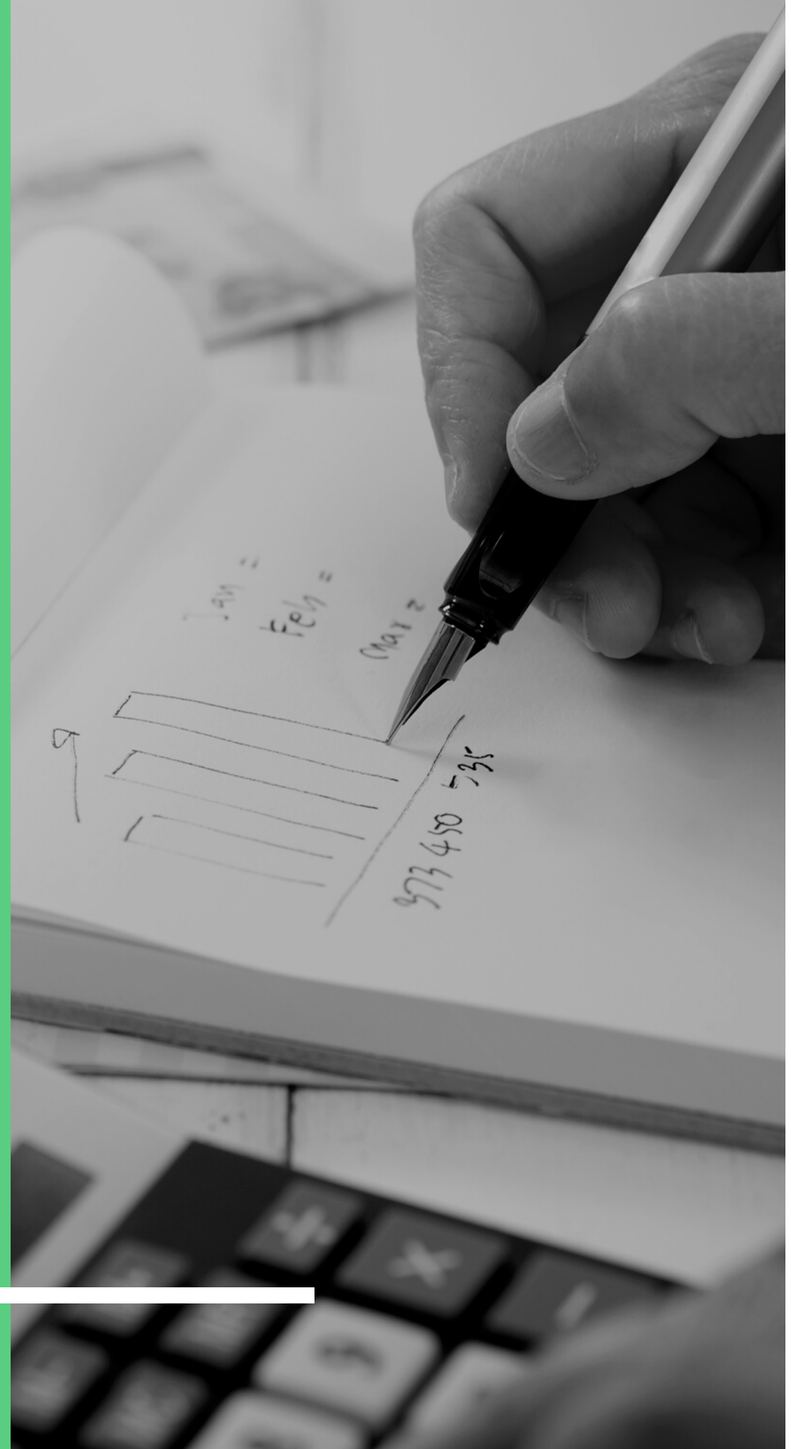
PRESENTATION OUTLINE

Introduction
Technical Background
Method and Implementation
Results

TODAY'S HIGHLIGHTS

WHAT IS CREDIT RISK?

Credit risk is the probability of a financial loss resulting from a borrower's failure to repay a loan. Essentially, credit risk results in an interruption of cash flows and increased costs for collection. Lenders can mitigate credit risk by analyzing factors about a borrower's creditworthiness, such as their current debt load, income, etc



RISK MANAGEMENT IS IMPORTANT IN BANKING

PURPOSE & GOALS OF THIS PROJECT

PURPOSE

This project aims to investigate different ML techniques and find and apply such that is best for predicting credit risk

GOALS

1. Investigating ML techniques and finding the best suited for credit risk modeling

2. Implement techniques to handle imbalanced data, that should be able to achieve a ROC-AUC score of higher than 65%

3. Find the otimal values for these hyperparameters to improve model performance

DATASET

This dataset classifies customer of Bank X described by a set of attributes as **GOOD** or **BAD** credit risks.

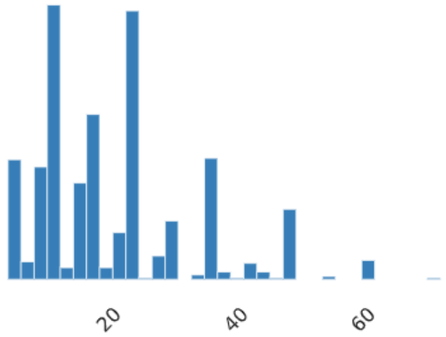
The initial data given by Bank X is given in an excel sheet of 1000 rows and 20 columns,

- Each row represents one customer application
- Each column contains represents one feature of this exact customer.

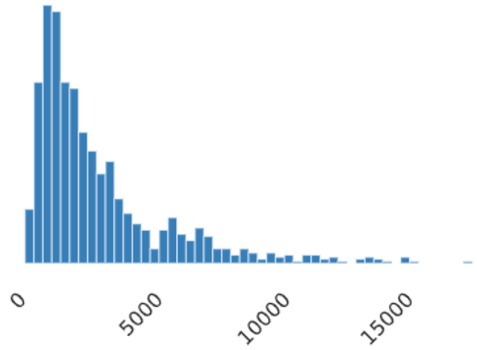
This means that the dataset contains information about 1000 customer applications.

Attributes

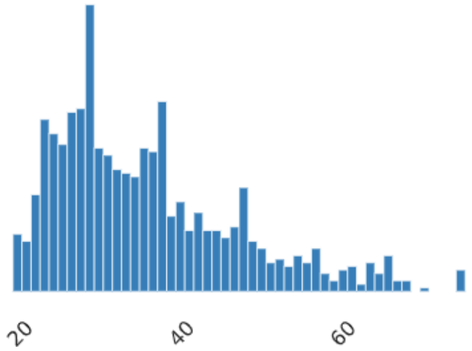
duration
Real number (ℝ)



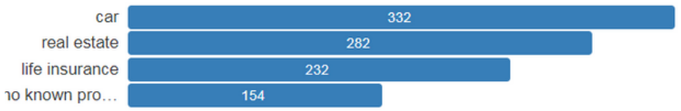
credit_amount
Real number (ℝ)



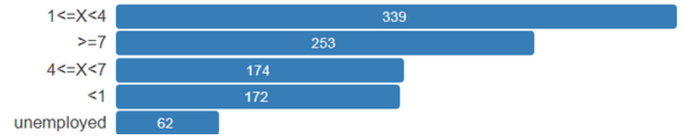
age
Real number (ℝ)



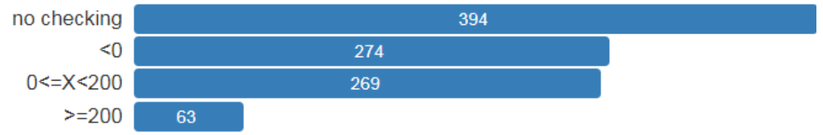
property_magnitude
Categorical



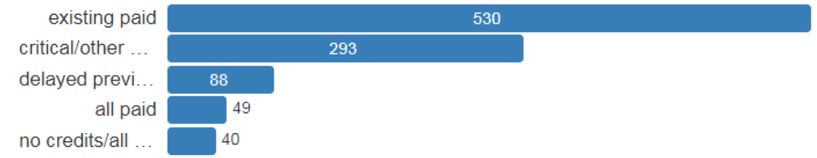
employment
Categorical



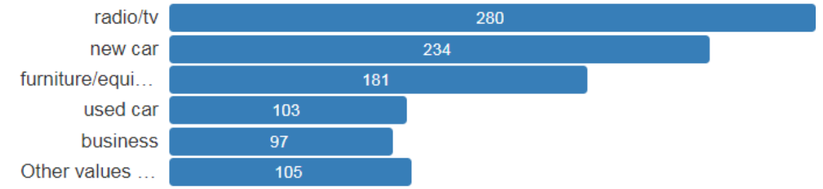
checking_status
Categorical



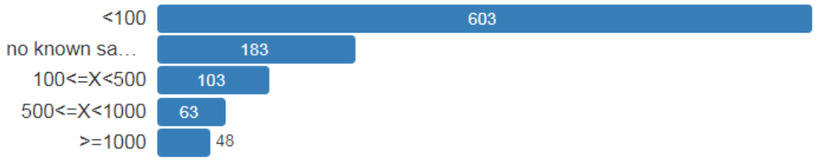
credit_history
Categorical



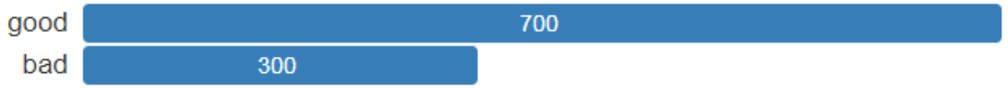
purpose
Categorical



savings_status
Categorical



class
Categorical



Data pre-processing techniques

MISSING VALUE

Checking if there are any missing values using np.isna()

ENCODE DATA

Encode categorical colums by OneHot method using pd.getdummies()

CHANGE LABELS OF TARGET

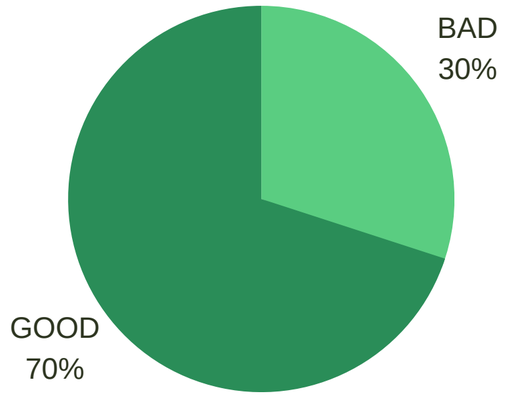
Change label of target y from Good and Bad to 1 and 0

SCALE DATA

Standardization making data points centered on the mean of all data points with a unit standard deviation

70%

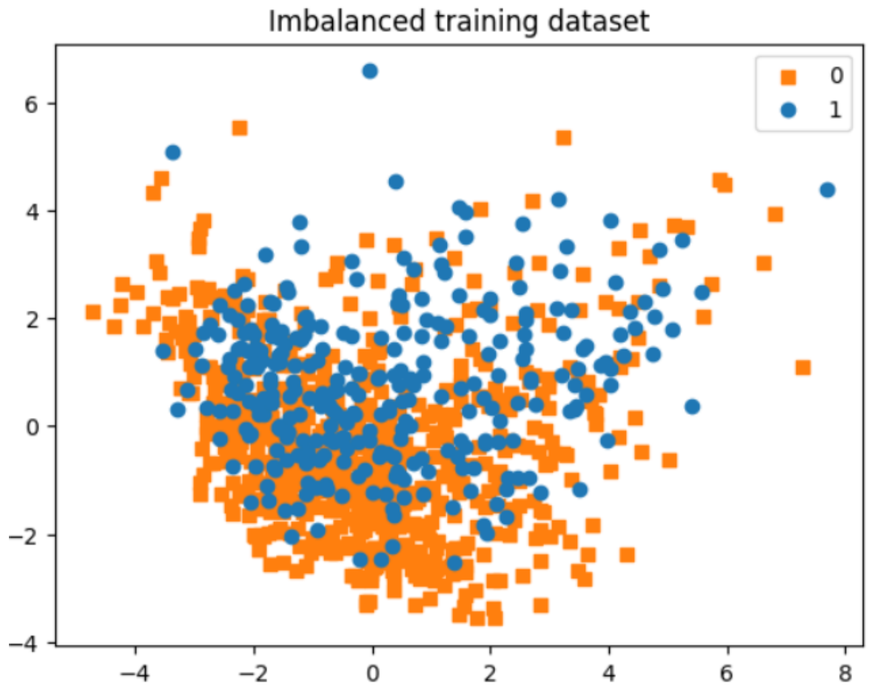
OF APPLICATIONS IS CLASSIFIED AS **GOOD** CREDIT RISK



30%

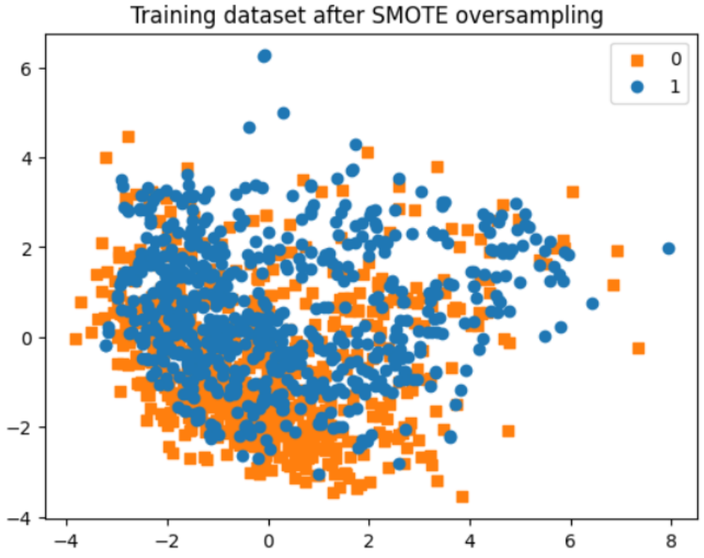
OF APPLICATIONS IS CLASSIFIED AS **BAD** CREDIT RISK

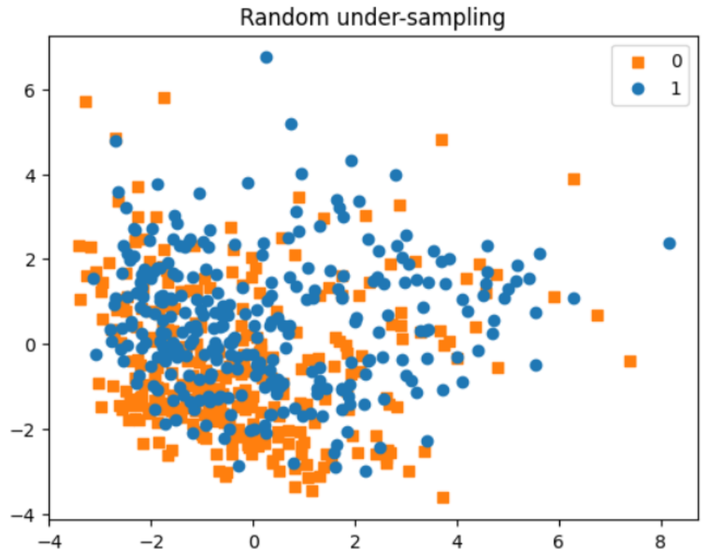
IMBALANCED HANDLING TECHNIQUES

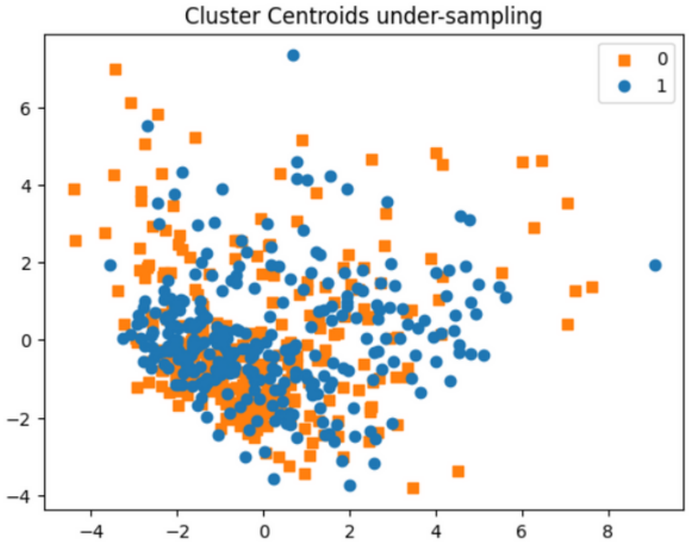


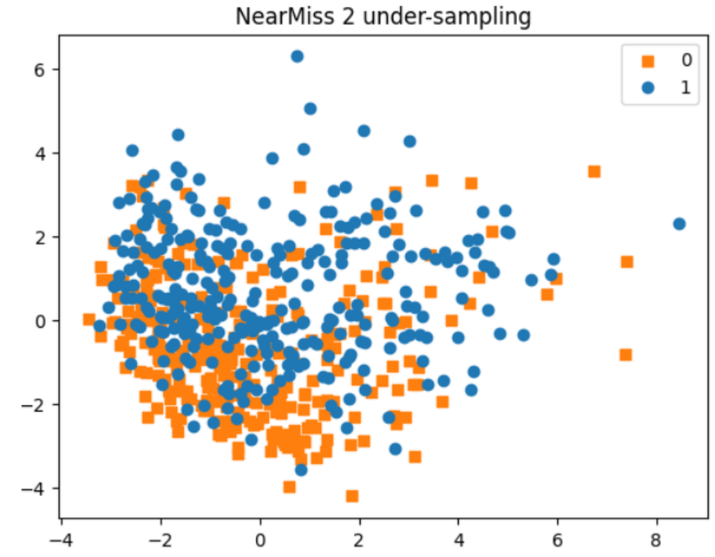
Mainly three things we can do when we have imbalanced data:

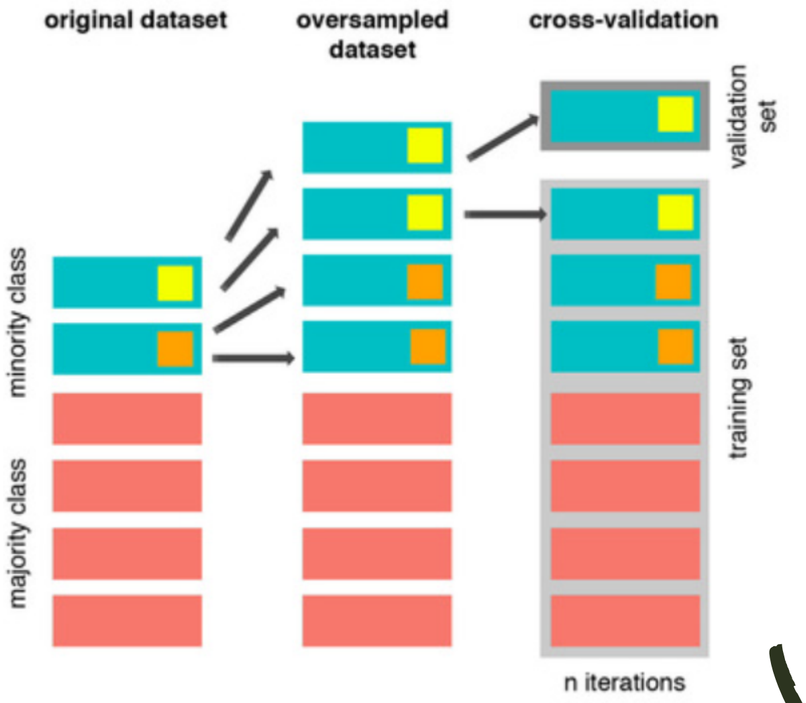
- Ignoring the problem.
- Undersampling the majority class.
- Oversampling the minority class.





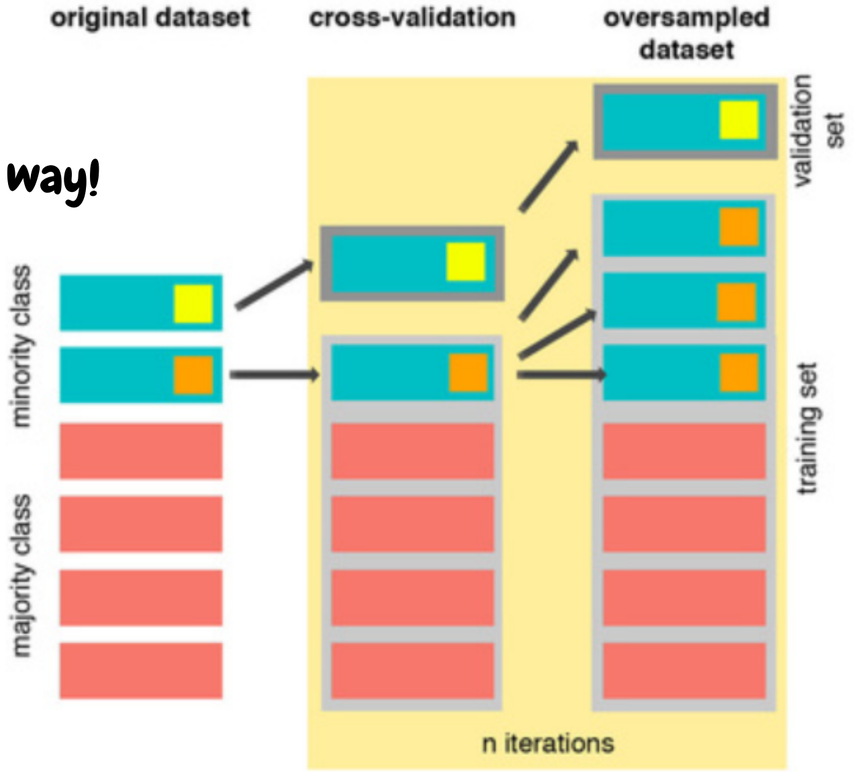






Oversampling the minority class can result in overfitting problems if we oversample before cross-validating

the right way!

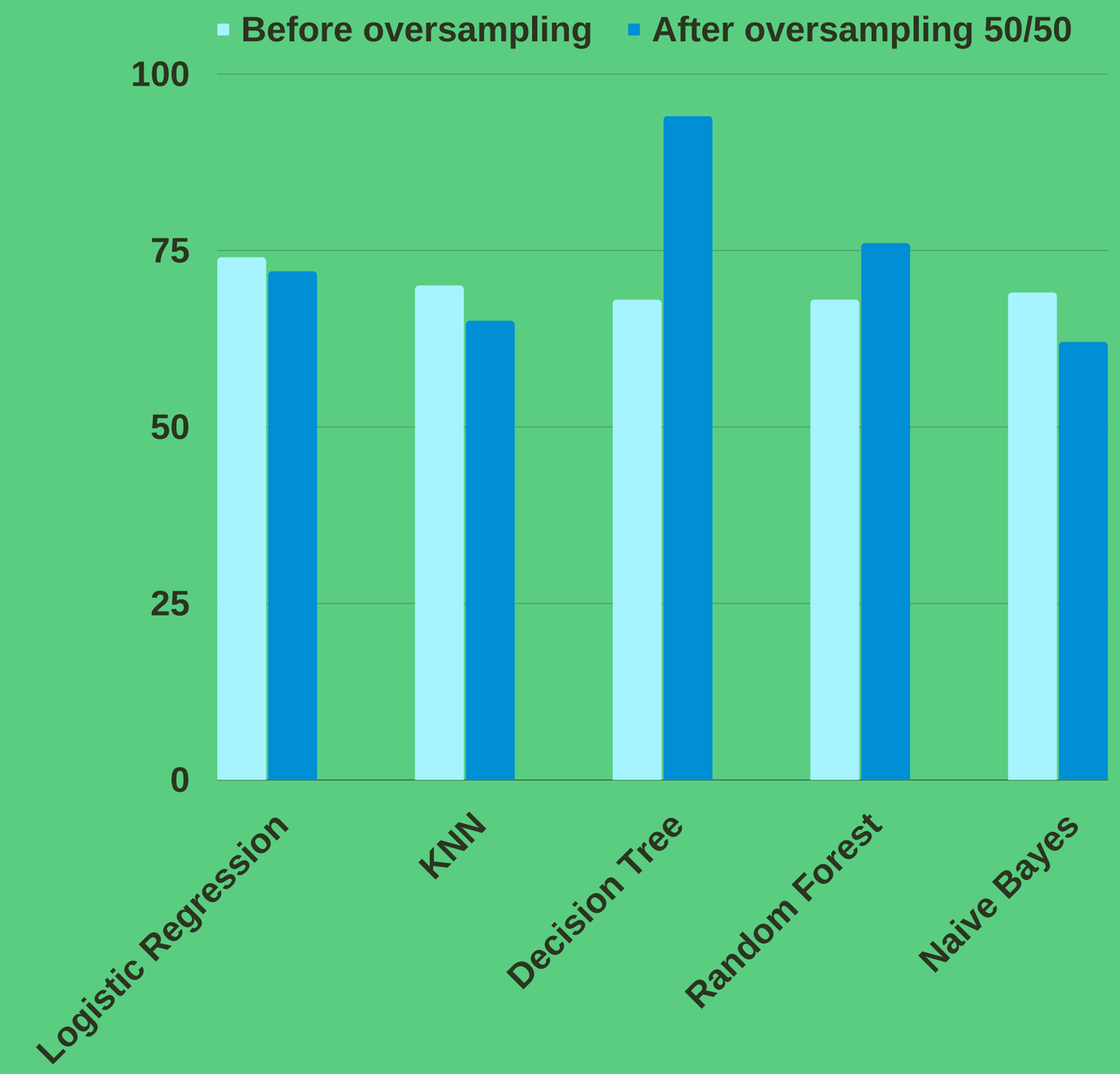


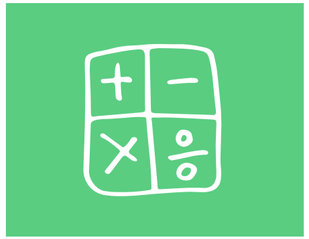
Proper Cross Validation when Oversampling

ACCURACY OF MODELS

$$\text{Accuracy} = \frac{TP + TN}{\text{total sample}}$$

Some machine learning models have higher accuracy score after optimzation





Precision

Recall

F1- score

Confusion Matrix

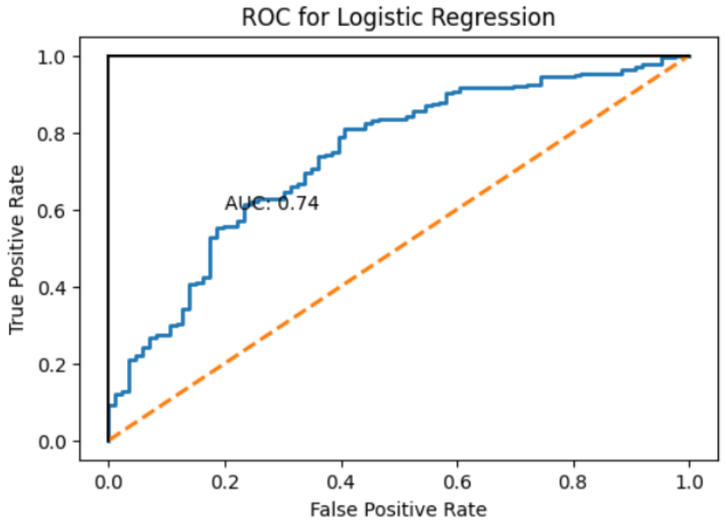
ROC curve-
AUC score

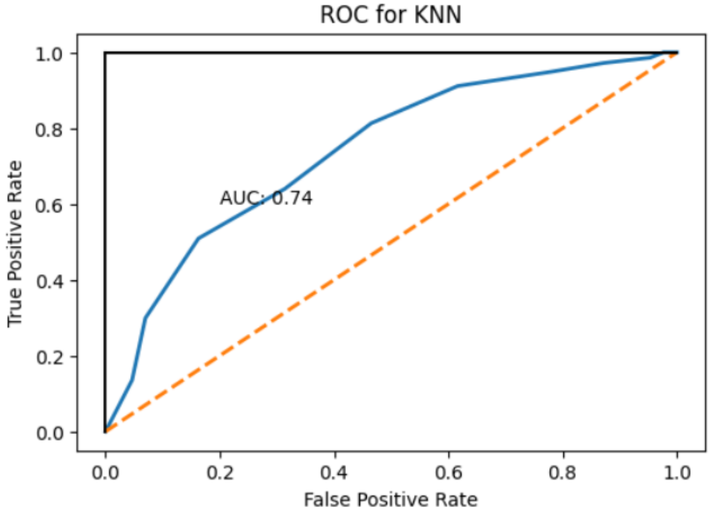
- Accuracy score limitation is that it measures on all labels without regard to the accuracy on each label.

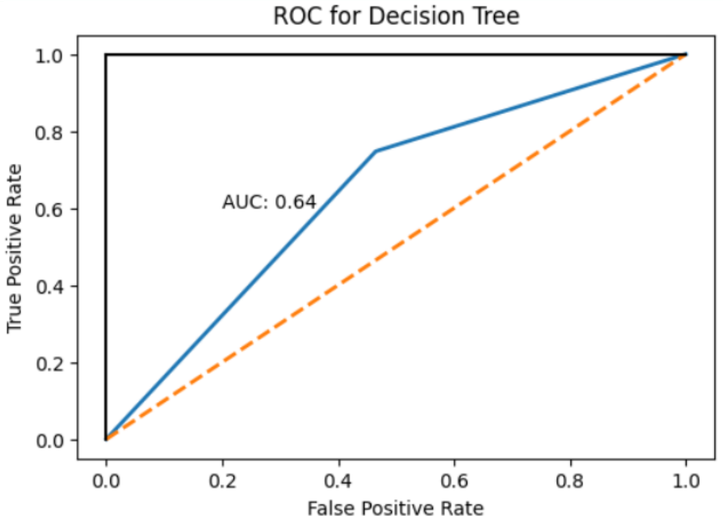
- It is therefore not suitable for evaluating tasks where the importance of predicting labels is not the same. It is more important for us to correctly detect a bad debt record than correctly detect a good profile

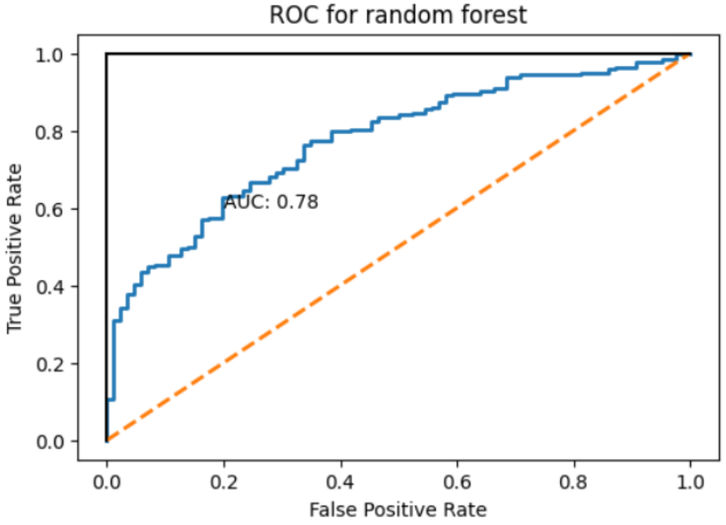
METRICS TO EVALUATE MODELS

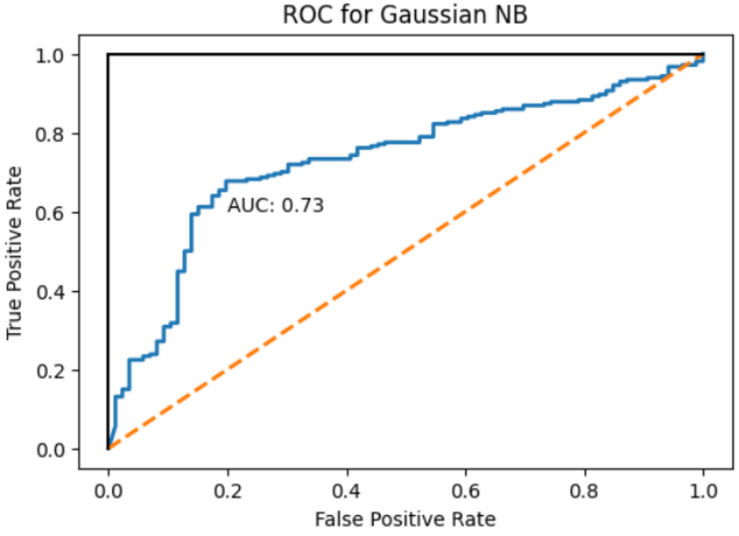
BEFORE TUNNING HYPERPARAMETER









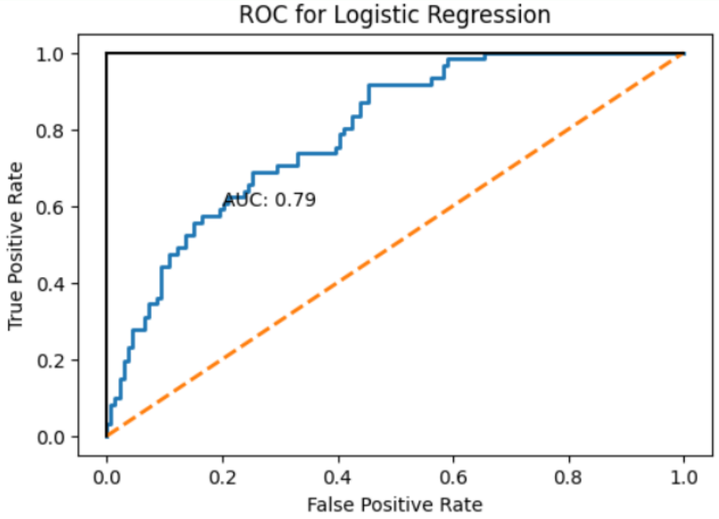


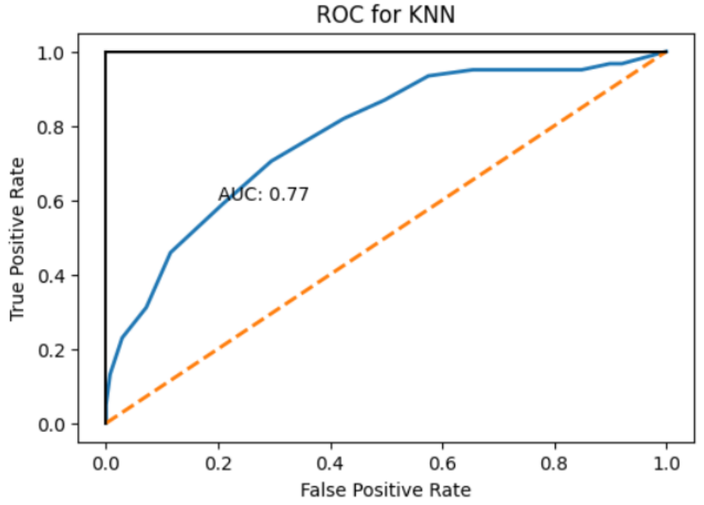
ROC is a curve representing the classifier ability of a classifier at threshold .

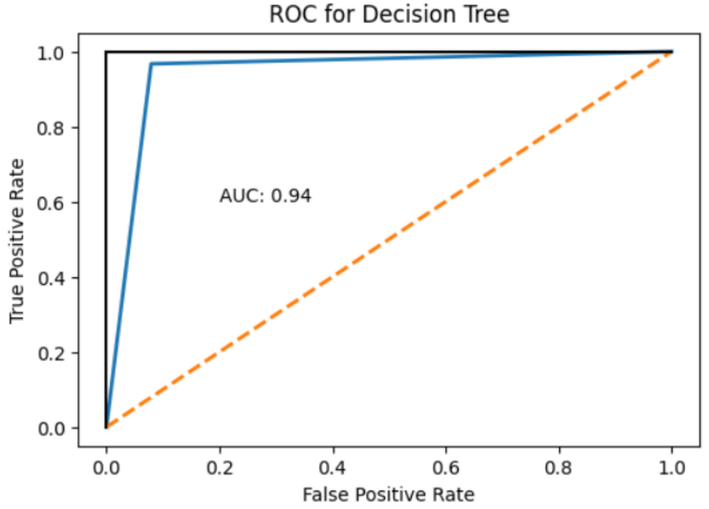
The larger AUC (area under curve), the ROC curve tends to be asymptotic to the y=1 and the better the model's classifier

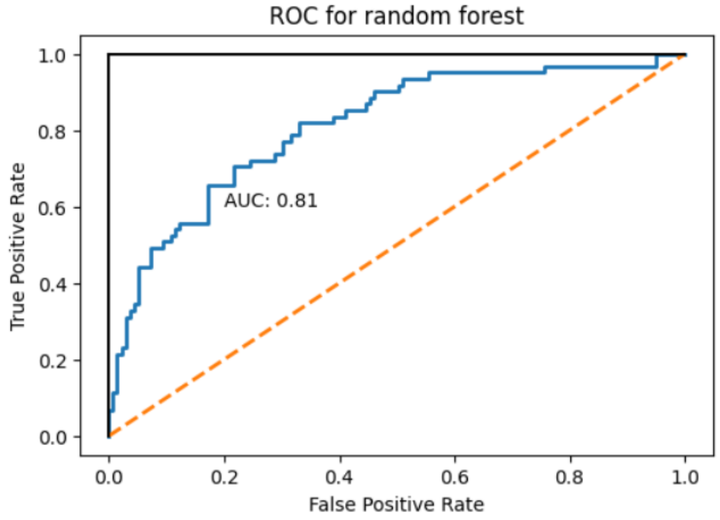
Random Forest have the highest AUC score

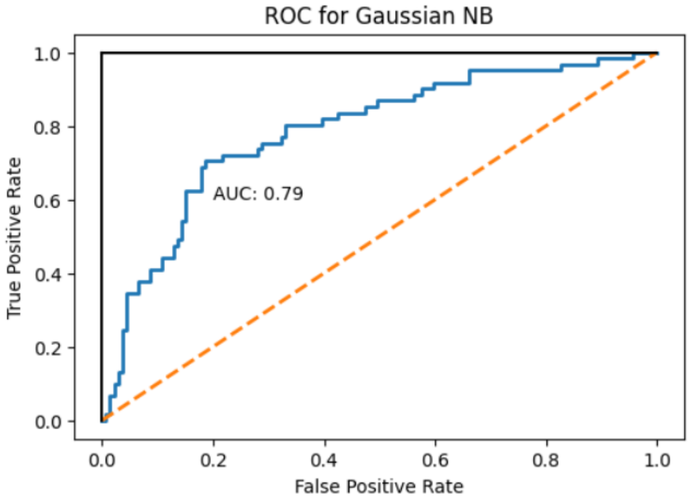
AFTER TUNNING HYPERPARAMETER BY GRIDSEARCHCV









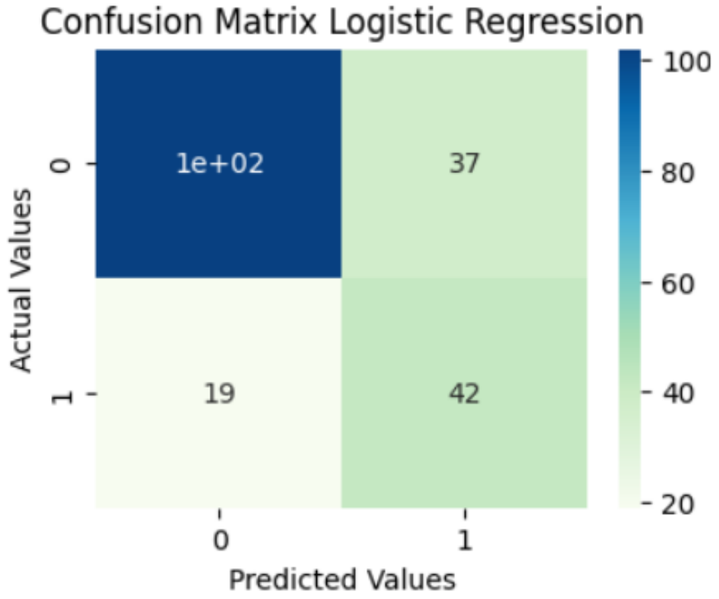


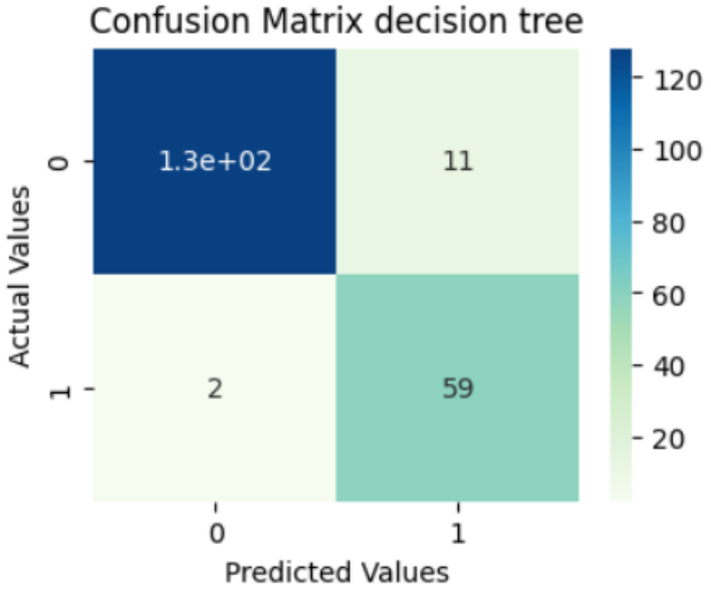
Decision Tree have the highest AUC score

- The performance of a mode can improve significantly if we can find optimal values for the hyperparameters.
- GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model

```
clf.best_params_
```

```
{'criterion': 'entropy', 'max_depth': 40}
```



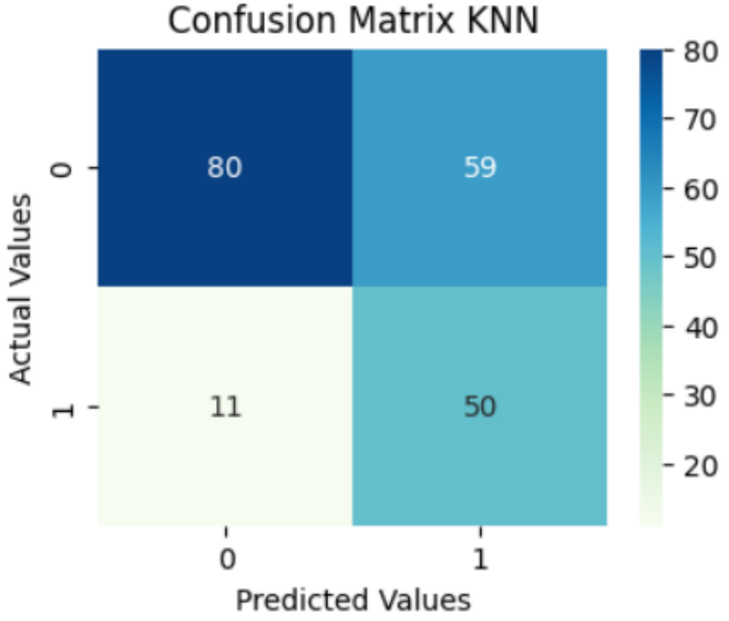


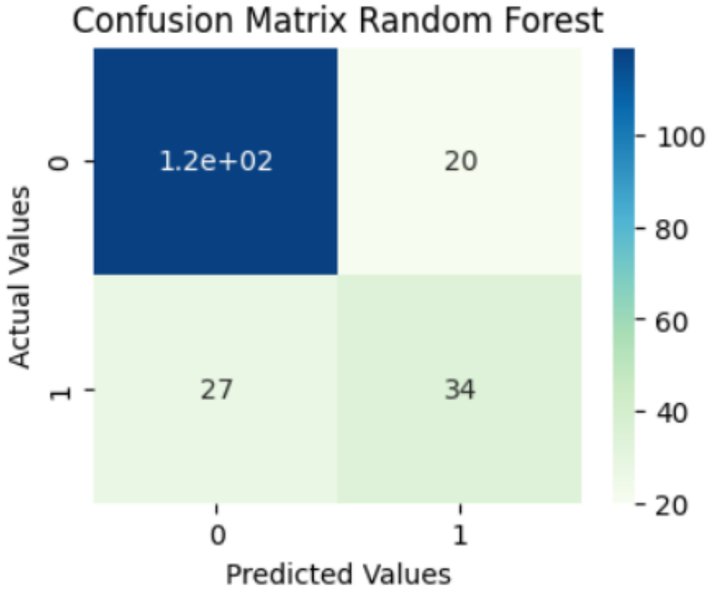
CONFUSION MATRIX

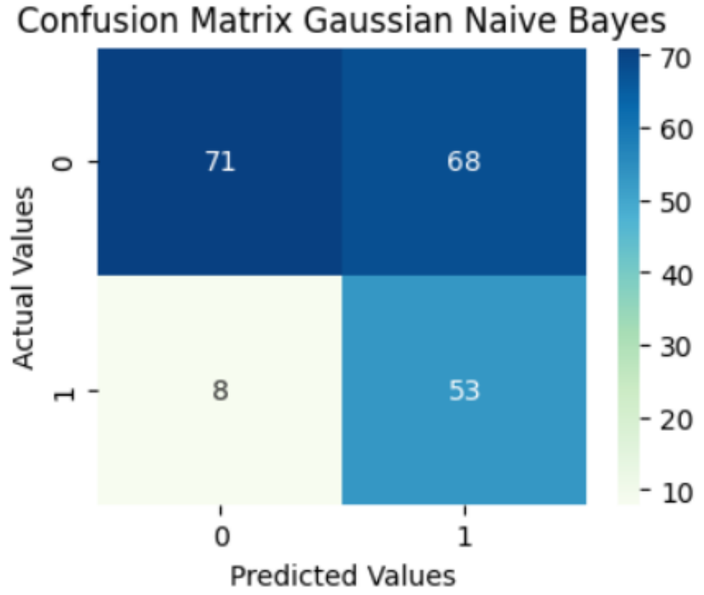
In the credit risk classification, it is more important for us to correctly detect a bad record.

Decision Tree, KNN and Naive Bayes are models with high TP

However, KNN and Naive Bayes also have high FP







	PRECISION	F1-SCORE	TEST RECALL	VALIDATION RECALL
Logistic Regression	0.75	0.73	0.72	0.69
K-Nearest Neighbors	0.75	0.66	0.65	0.68
Decision Tree	0.94	0.94	**0.94**	**0.92**
Random Forest	0.76	0.76	0.77	0.97
Gaussian Naive Bayes	0.75	0.64	0.63	0.78

Bad news: Random Forest, Gaussian Naive Bayes are overfiting models!

Good new is that Decision Tree Classifer have recall in test set and validation set roughly consistent (94% vs 92%) and much higher than results before optimization (68%)

DISCUSSION AND FUTURE WORK

The best-suited model for predicting credit risk is shown to be the **Decision Tree** model.

IMPLEMENT UNDERSAMPLING TECHNIQUE

Results of models after undersampling data can be different from oversampling

USE CREDIT SCORECARD FOR PREDICTION

credit scorecard is a formula that uses data elements, or variables, to determine a threshold of risk tolerance