Credit Card Predictive Analysis

# Introduction

Credit card bankruptcy occurs when you have fallen far behind on your credit card payments. At the same time, regardless of their repayment abilities, most cardholders misused their credit cards for consumption and accumulated large amounts of credit and debt. This project aims to apply supervised machine learning algorithms to identify the important drivers that influence the chance of credit card default while emphasizing the mathematical features of the approaches used..

The objective is to create an automated model for detecting critical determinants and forecasting credit card defaults based on client information and recorded transactions. The supervised machine learning paradigm's general principles are later described, along with a full description of the methodologies and algorithms utilized to generate the models.

# Business Understanding

Credit cards are a specific risk management tool in the financial business. It predicts the likelihood of future defaults and credit card borrowings based on personal information and data given by credit card applicants. The bank can determine whether or not to offer the applicant a credit card. Credit ratings can estimate the level of risk objectively.

Credit scorecards are often based on past data. When confronted with significant economic volatility. Previous models may lose their predictive ability. A logistic model is a typical credit scoring approach. Logistics is appropriate for binary classification jobs and can determine the coefficients of each feature. The score card will multiply the logistic regression coefficient by two to aid comprehension and operation.

Each time there is a hard inquiry, your credit score suffers. This project predicts whether or not a credit card application will be accepted. This program predicts your chances of approval without hurting your credit score. Applicants who wish to check out if they will be approved for a credit card without impacting their credit score can use this.

# Business Objectives

Each time there is a hard inquiry, your credit score suffers. This project predicts whether or not a credit card application will be accepted. This program predicts your chances of approval without hurting your credit score. Applicants who wish to check out if they will be approved for a credit card without impacting their credit score can use this.

# Data Description

The dataset includes credit card transactions performed by European cardholders in September 2013. This dataset contains 492 frauds out of 284,807 transactions over two days. The dataset is very uneven, with positive transactions accounting for 0.172% of all transactions. It only has numerical input variables that result from a PCA transformation. We cannot give the original features and further background information about the data owing to confidentiality concerns. V1, V2,... are the features. The primary components derived by PCA are V28; the only features that have not been converted using PCA are 'Time' and 'Amount.' The 'Time' feature contains the number of seconds between each transaction and the first transaction in the dataset.

There're two tables could be merged by ID:

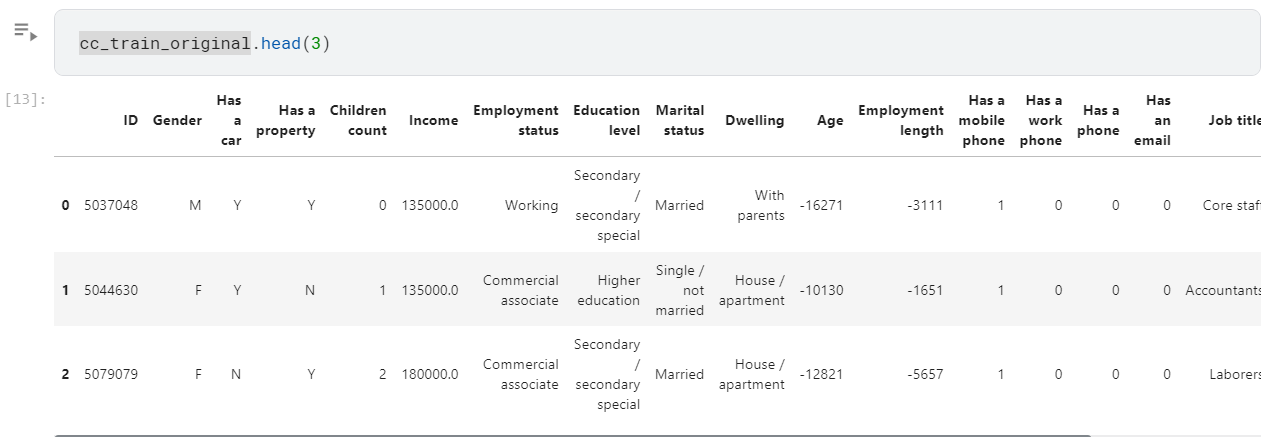
| application\_record.csv |  |
| --- | --- |
| Feature name | Explanation |
| ID | Client number |
| CODE\_GENDER | Gender |
| FLAG\_OWN\_CAR | Is there a car |
| FLAG\_OWN\_REALTY | Is there a property |
| CNT\_CHILDREN | Number of children |
| AMT\_INCOME\_TOTAL | Annual income |
| NAME\_INCOME\_TYPE | Income category |
| NAME\_EDUCATION\_TYPE | Education level |
| NAME\_FAMILY\_STATUS | Marital status |
| NAME\_HOUSING\_TYPE | Way of living |
| DAYS\_BIRTH | Birthday |
| DAYS\_EMPLOYED | Start date of employment |
| FLAG\_MOBIL | Is there a mobile phone |
| FLAG\_WORK\_PHONE | Is there a work phone |
| FLAG\_PHONE | Is there a phone |
| FLAG\_EMAIL | Is there an email |
| OCCUPATION\_TYPE | Occupation |
| CNT\_FAM\_MEMBERS | Family size |

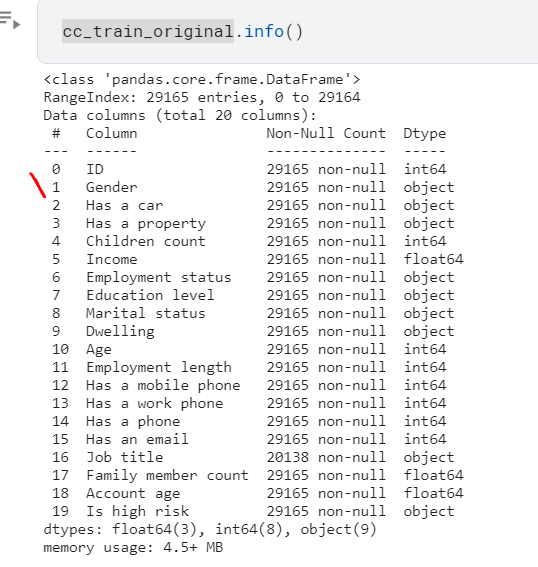
| credit\_record.csv |  |
| --- | --- |
| Feature name | Explanation |
| ID | Client number |
| MONTHS\_BALANCE | Record month |
| STATUS | Status |

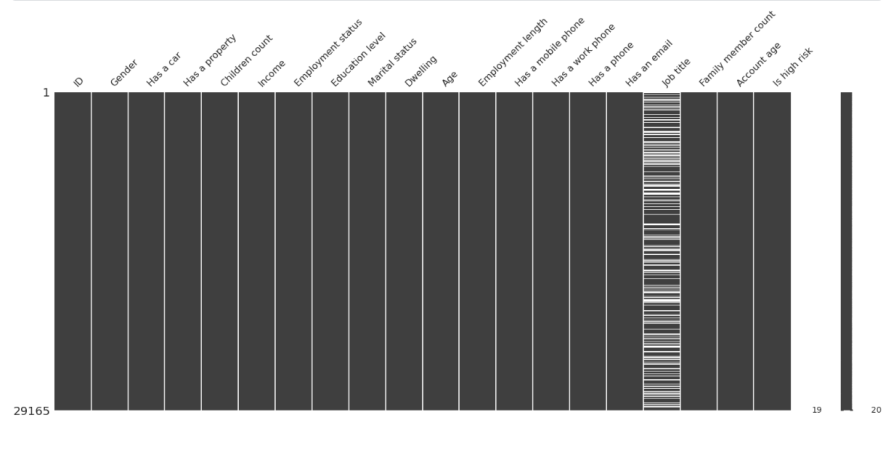
# EDA:

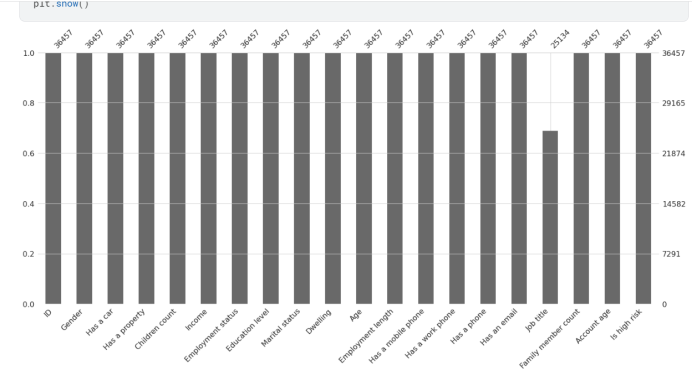
Exploratory Data Analysis (EDA) is a way of evaluating datasets to summarise their essential properties, frequently using visual approaches. EDA is used before modeling to see what the data can tell us. It is not easy to detect essential data properties by looking at a column of numbers or an entire spreadsheet. Looking at basic statistics may be time-consuming, uninteresting, and daunting. In this case, exploratory data analysis approaches have been developed to help.

## Quick glance at the data

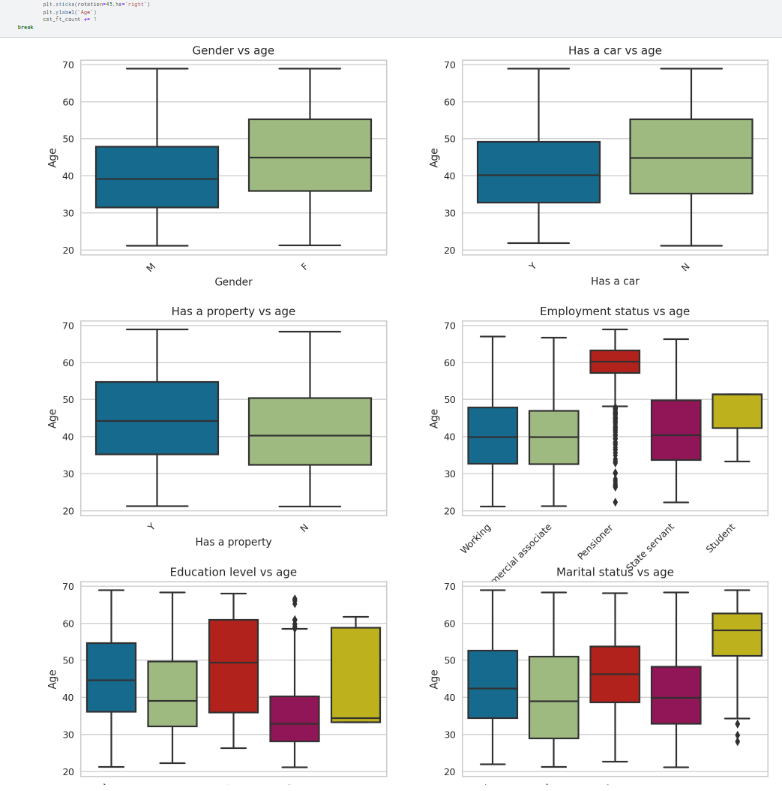


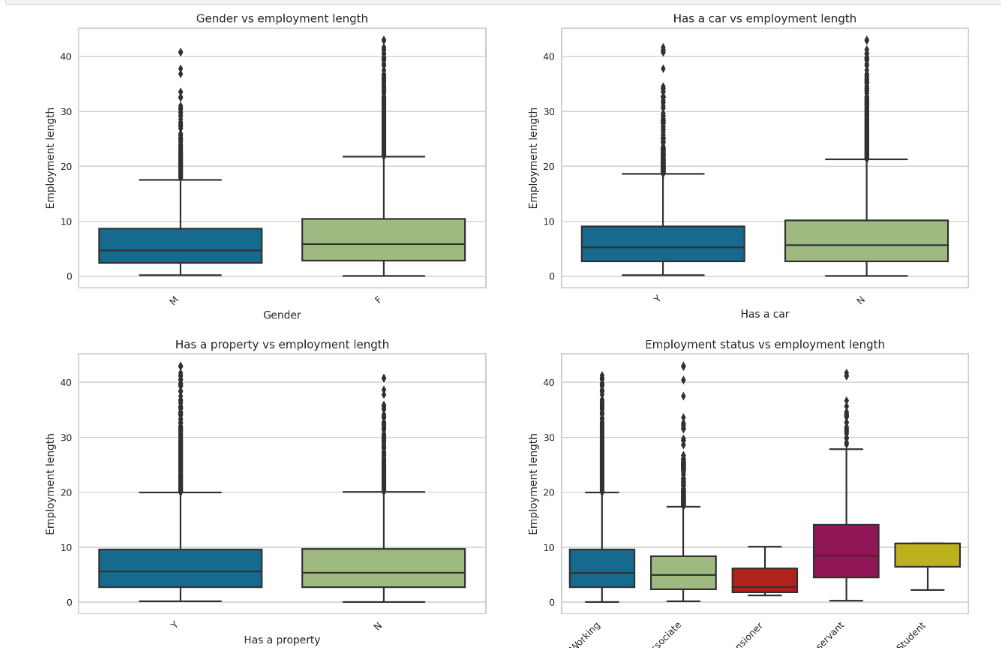


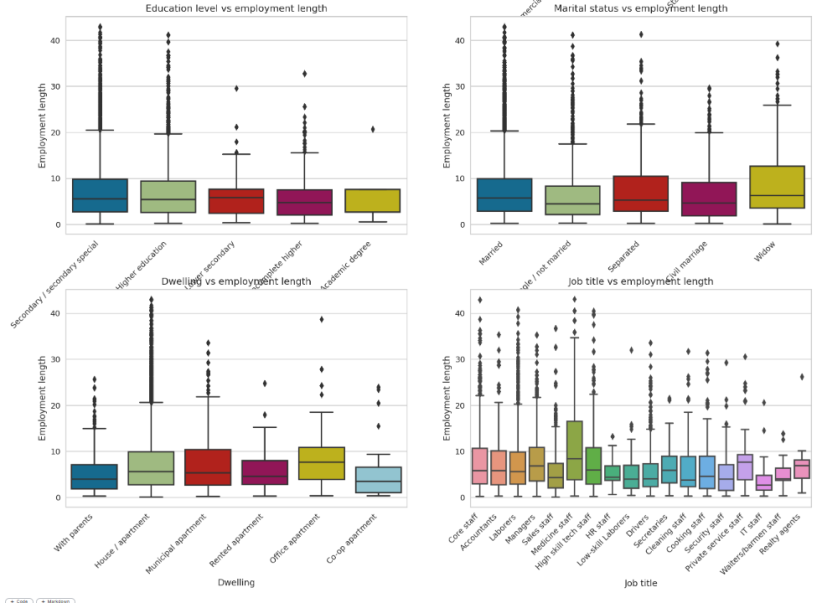


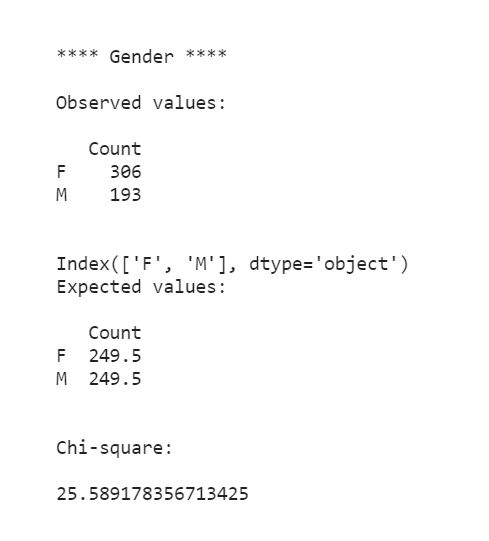


Now, I compare Age with gender. Has a car', 'Has a property', 'Employment status', 'Education level', 'Marital status', 'Dwelling', 'Job title'



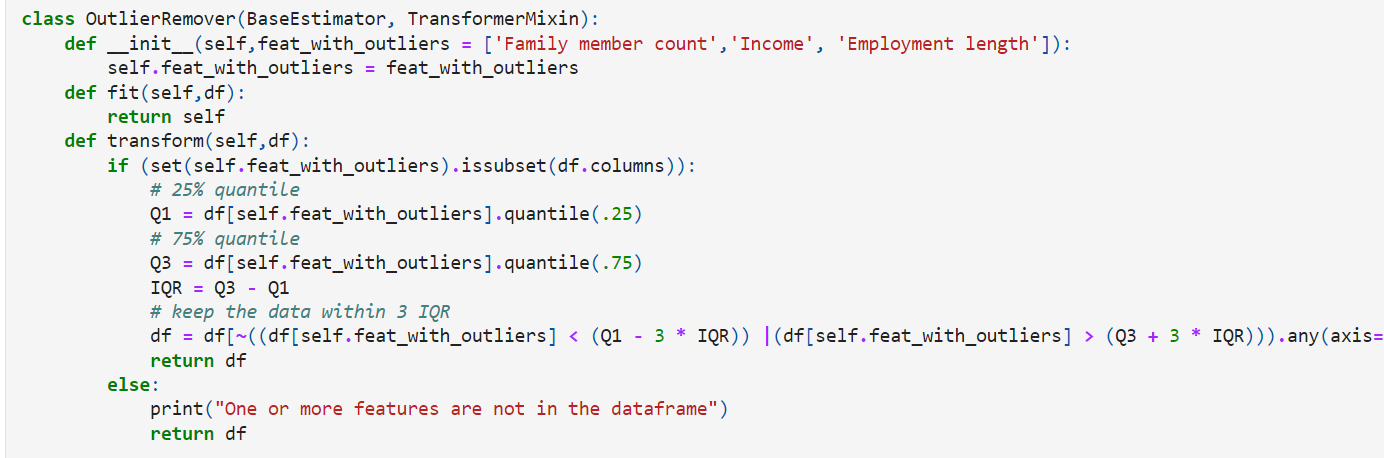






# Data PreProcessing

### Outliers handling



## ****Feature selection****

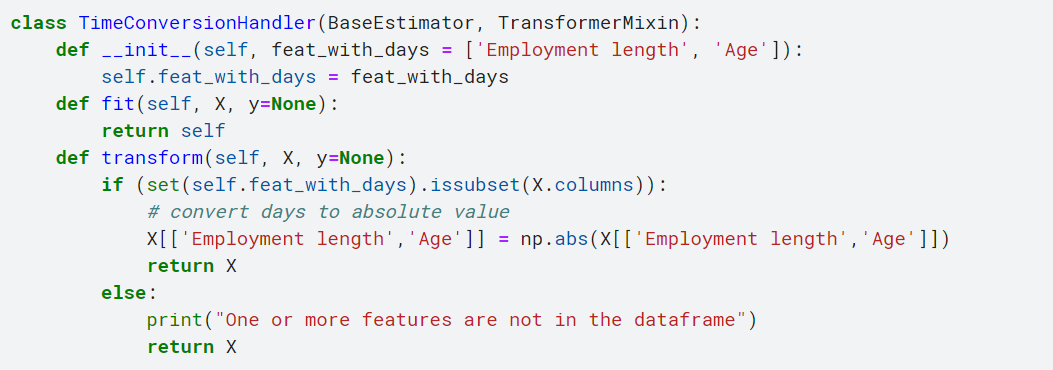
We drop extra feature.



Why are we getting rid of these characteristics?

* ID: ID is not relevant for prediction; it helped us when we integrated the two datasets, but there is no reason to preserve it beyond that.
* Has a mobile phone: Because everyone has a mobile phone, this attribute provides no information.
* Children count: is highly correlated with Family member count and to avoid multicollinearity, we drop it.
* Job title: Has some missing values and the count of each category is not very different to justify the use of mode. So we drop it.
* **Account age:** Because the account was used to create the target, reusing will make our model overfit. Plus, this information is unknown while applying for a credit card.

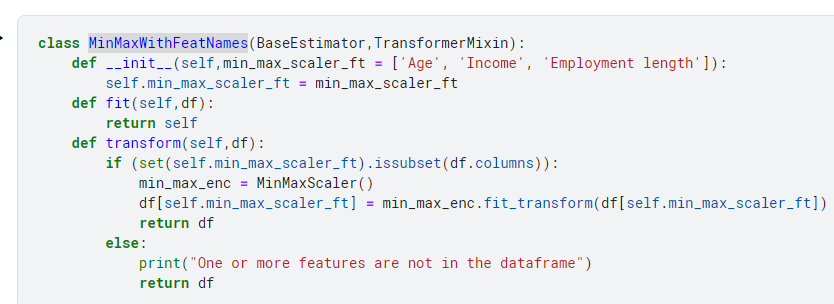
### Time conversion



### One hot encoding



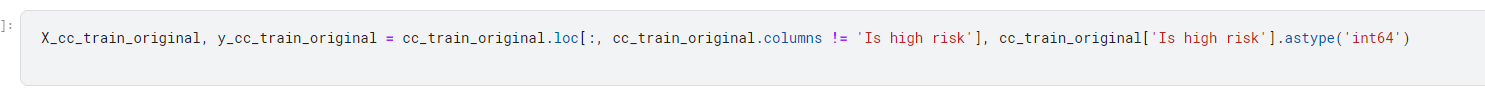
## MinMaxWithFeatNames



We are oversampling with SMOTE because the minority class (Is high risk = 1) is very rare in the data.

# Train and test

split the train data into X and y (target)



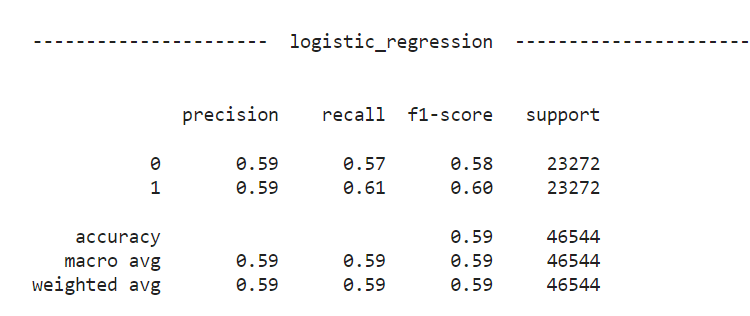
Modeling

In this I am using LogisticRegression, SVM, Decision tree, random forest, KNN, naïve bayes, gradient boosting, LinearDiscriminantAnalysis, NN, AdaBoostClassifier.

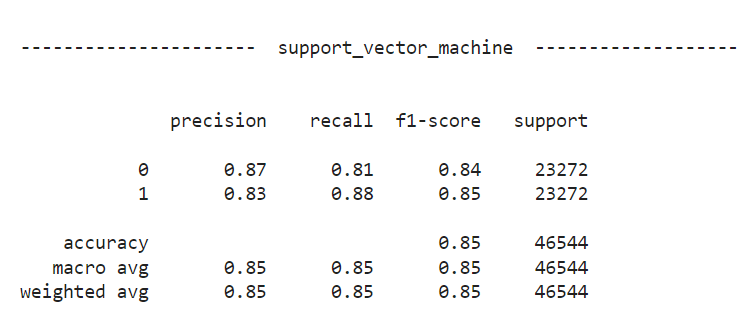
****

## Outputs

### Logistic regression



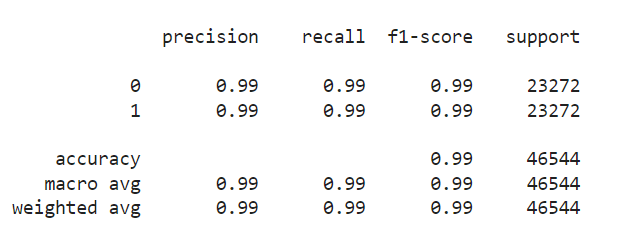
### SVM



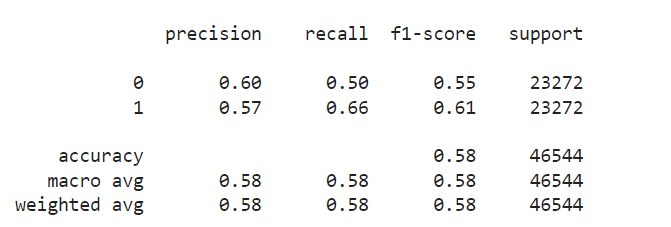
### Decision tree

### 

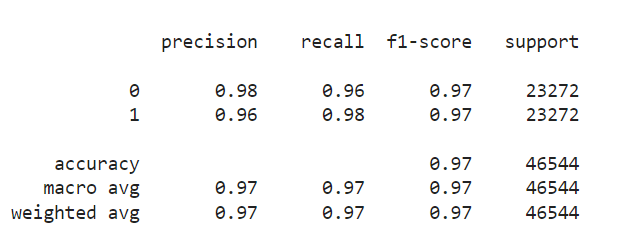
### Random forest



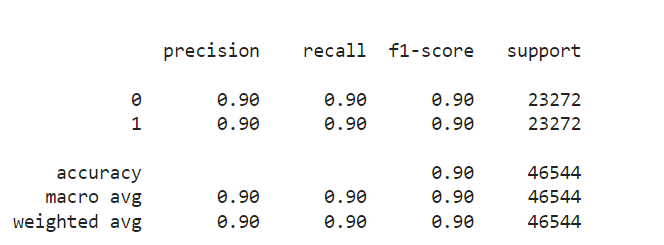
### Naïve bayes



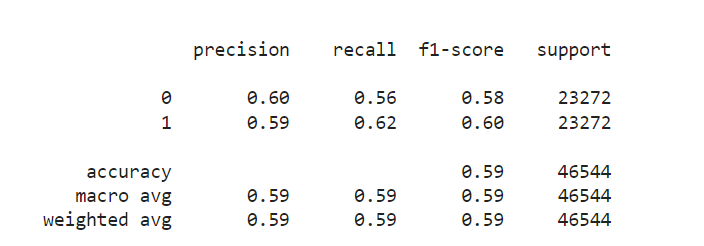
### KNN



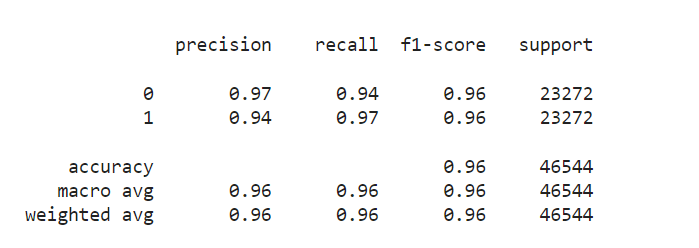
### Gradient boosting



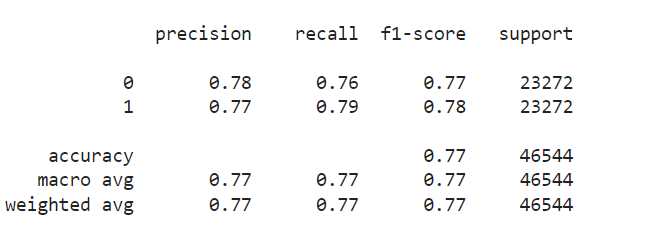
### Linear discriminant analysis



### NN



### Adaboost



# Conclusions:

In this, we will utilize recall as our metric because we are in the longest bull market . we can that conclude that our top model is Gradient boosting classifier when we check Roc and recall.

Because the goal of this challenge is to reduce the risk of loan default for the financial institution, the indicators to employ are determined by the present economic situation:

* People feel prosperous and are frequently employed during a bull market (when the economy is expanding). Money is often inexpensive, and the danger of default is low. Because the financial institution can handle the risk of default, it is not too cautious regarding credit. As long as most applicants are excellent clients, the financial institution can handle a few problematic clients (aka those who pay back their credit). A high recall (sensitivity) is ideal in this instance.
* People lose their employment and money in the stock market during a bear market (when the economy is declining). Many people have difficulty meeting their financial commitments. As a result, financial institutions are more cautious about extending credit or making loans. The financial firm cannot afford to extend credit to consumers who will be unable to repay it. The financial institution would prefer fewer excellent clients, even if it means that some good clients are denied credit than to have any bad clients. In this instance, high accuracy (specificity) is preferred.