المعهد الوطني للبريد والمواصلات
ⵓⵔⴰⵔ ⵏⵓⵟⵟⴰⵏ ⵏ ⵜⵓⵙⵙⵓⵔⵜ ⴰ ⵜⵉⵎⵓⵙⵙⵉⵍⵉⵏ
Institut National des Postes et Télécommunications

SCORING PROJECT

Submitted by :
AZIZ YOUNESS
AJABBOUR ILYASS
BALAH ILIAS

Guided by :
MS AKKAOUI ZINEB

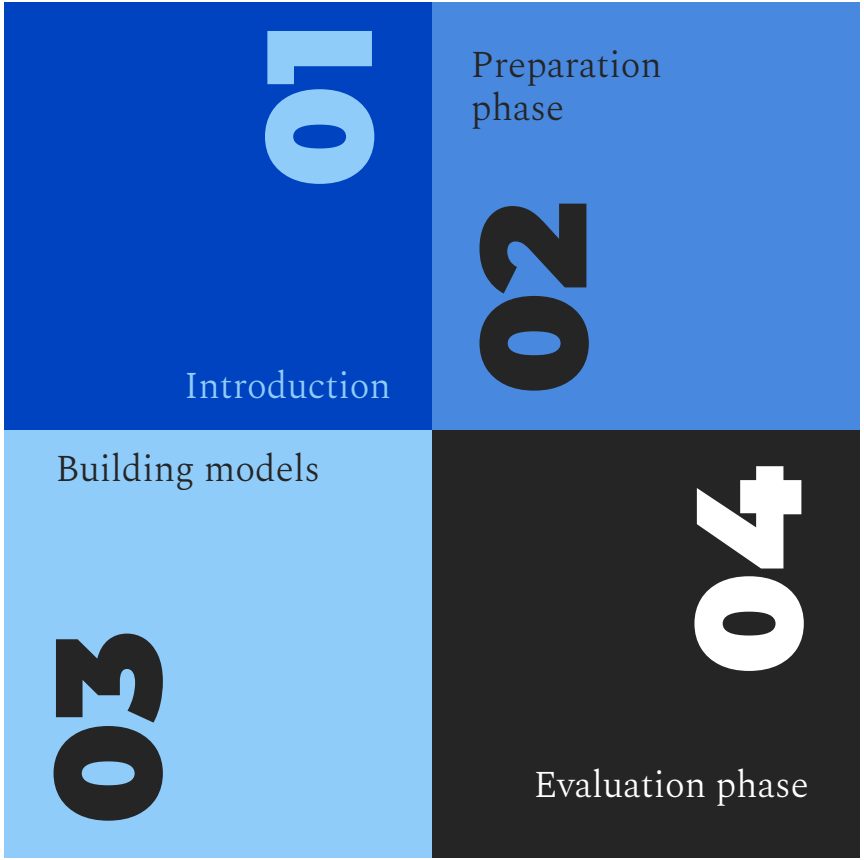
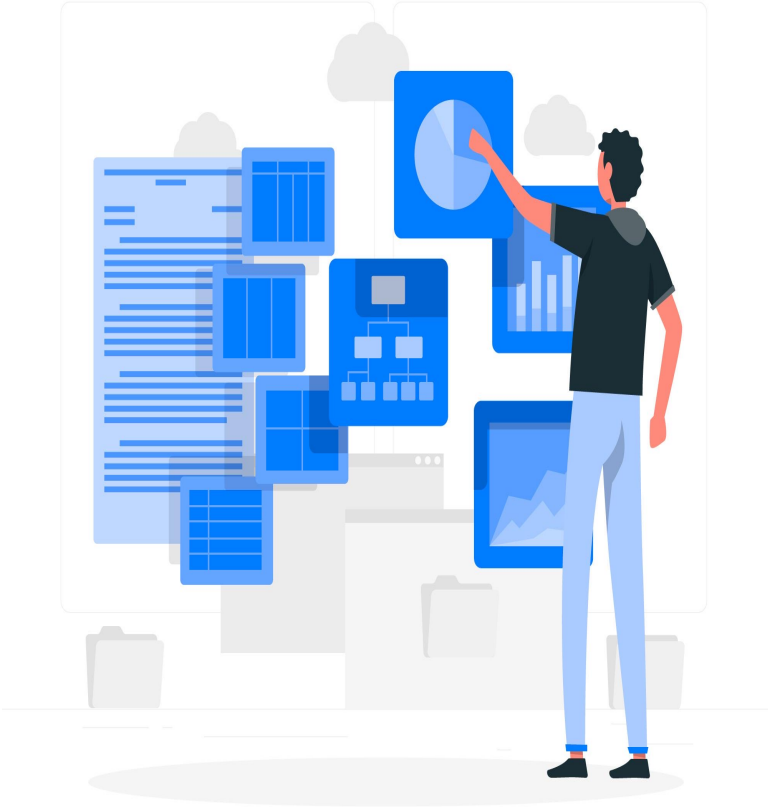


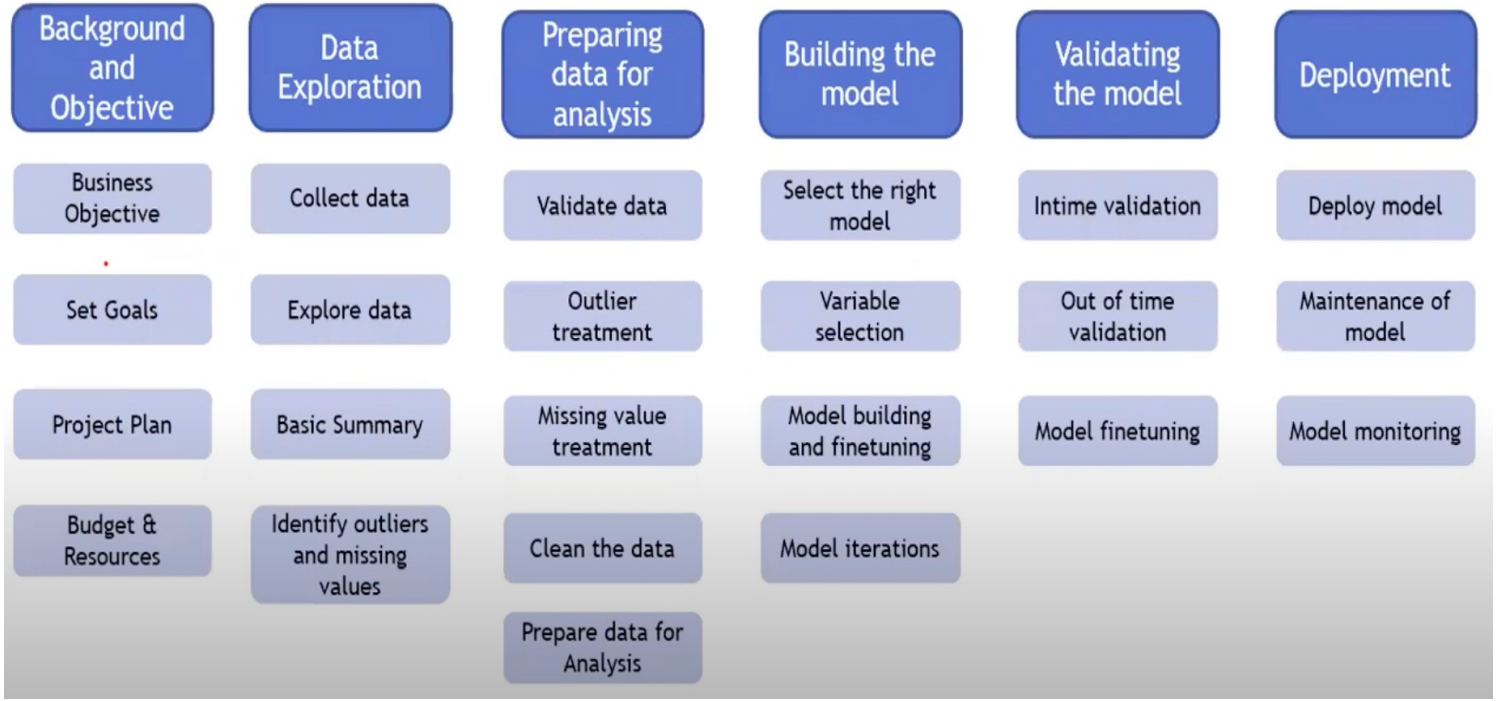
Table of Contents

1. Project objective

Credit is a very important product in banking and financial institutions. There is always a customer in need of a loan. Since loans are always accompanied by risks, it is important to identify suitable applicants, and there have to be a means to determine and separate the good applicants from the bad. To solve this issue, financial institutions such as banks started developing credit scores. Using the customer's credit scores, lenders can define the risk of loan applicants.



2. Process Aproach



3. Data sources and variables

Data source

Our data exists in the .csv files attached with this document: ScoringTraining, ScoringTest, and SampleScoring and the description predictive variables Data Dictionary file. These data will help us to learn and validate the constructed score prediction models, as well as evaluate their performance.

Data variables

The target "SeriousDlqin2yrs" is a binary variable that determines whether or not someone defaulted on their bank loan. There are 10 predictors in the data set:

"RevolvingUtilizationOfUnsecuredLines"
"Age"
"DebtRatio"
"MonthlyIncome"
"NumberOfOpenCreditLinesAndLoans"
"NumberRealEstateLoansOrLines"
"NumberOfTime30-59DaysPastDueNotWorse"
"NumberOfTime60- 89DaysPastDueNotWorse"
"NumberOfTimes90DaysLate"
"NumberOfDependents"

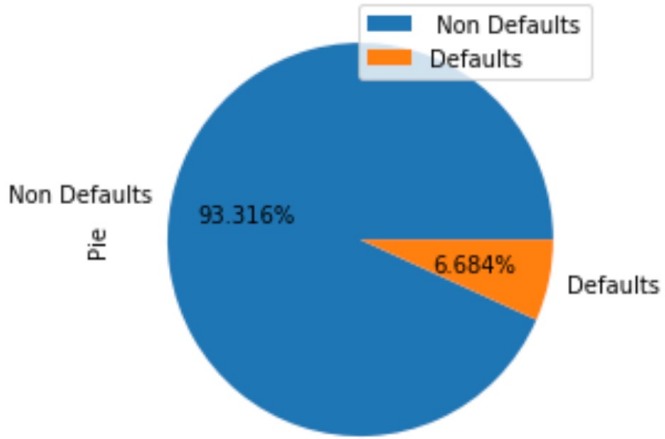
Chapter Two
Preparation phase / Data preprocessing

1. The proportion of defects

- **The Problem with Class Imbalance:**

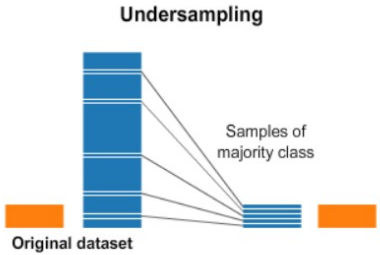
Most machine learning algorithms work best when the number of samples in each class are about equal.
This is because most algorithms are designed to maximize accuracy and reduce errors.

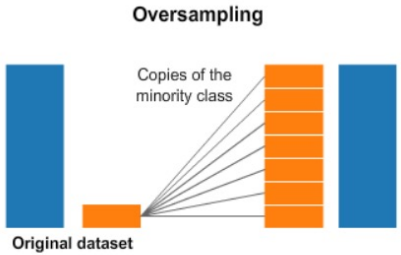
However, if the data set in imbalance then In such cases,
you get a pretty high accuracy just by predicting the **majority class**,
But you fail to capture the **minority class**, which is most often the point of creating the model in the first place.



Resampling Technique

A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling).





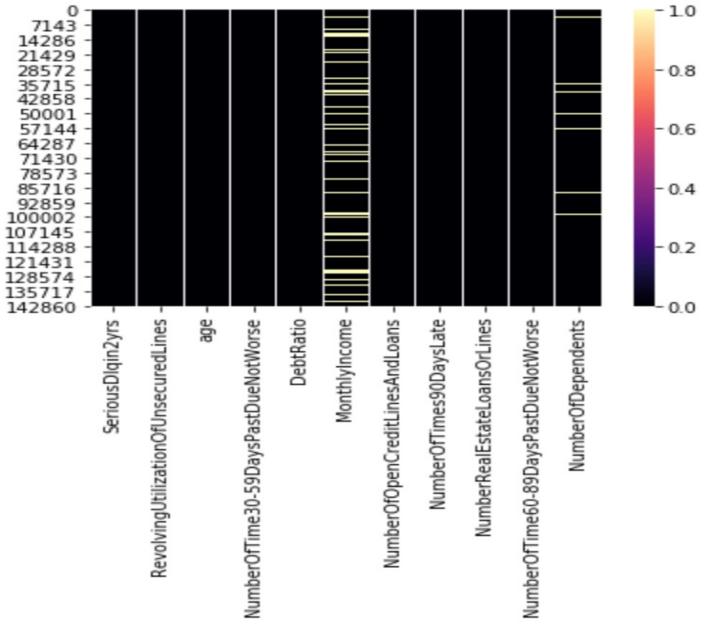
Undersampling :
There is a big loss of data as the total number of rows will be reduced to the number of rows for event rate (Defaults). So, there is a high possibility of High Bias.

Oversampling:
There will be synthetic data added to make the Defaults observations equal to the Non-Defaults. This can increase the variance in our model which leads to overfitting. And increase the model run time if we are dealing with large number of Non-Defaults.

2. Missing values

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

	Column Name	Missings	Missing_pct
0	MonthlyIncome	20103	0.2
1	NumberOfDependents	2626	0.0



Method used to treat missing values :

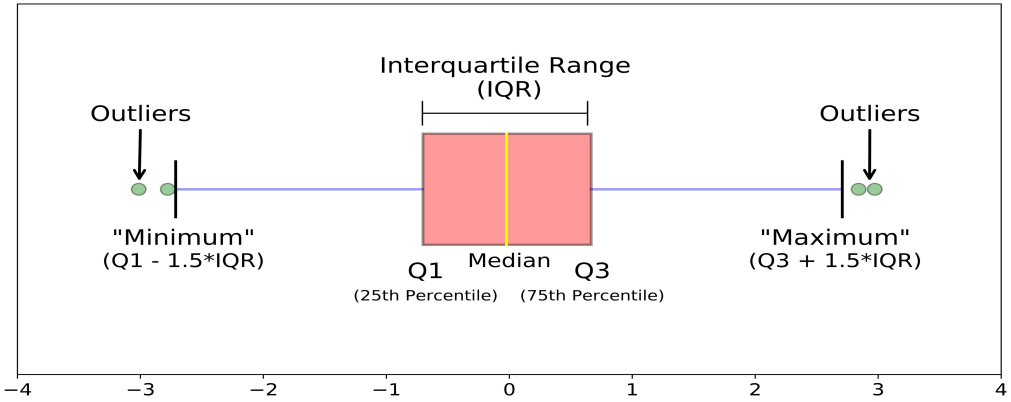
Since the missing percentage is really small it is better to replace the missing values with the mean / median.

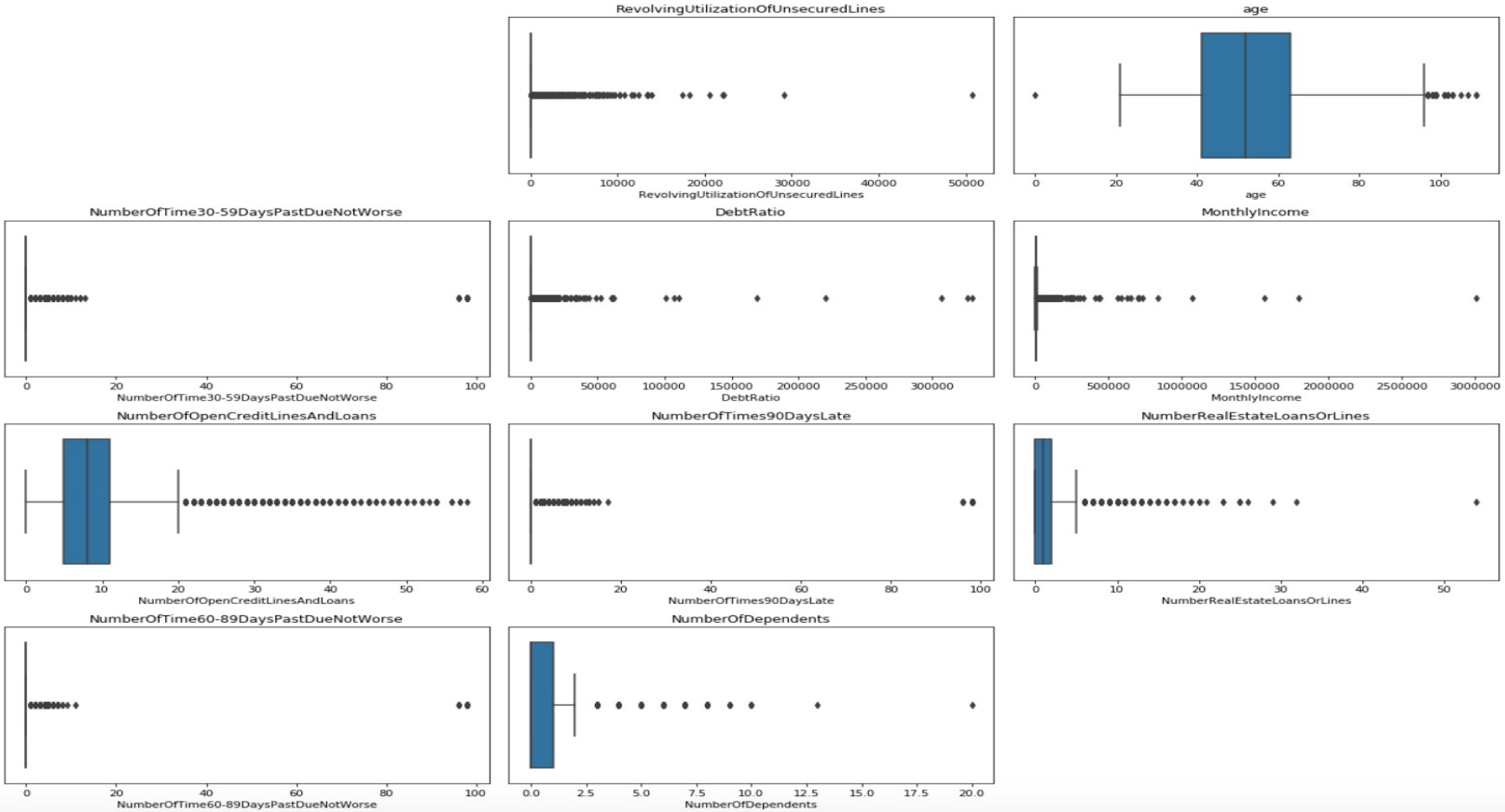
Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

3. Outliers

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:

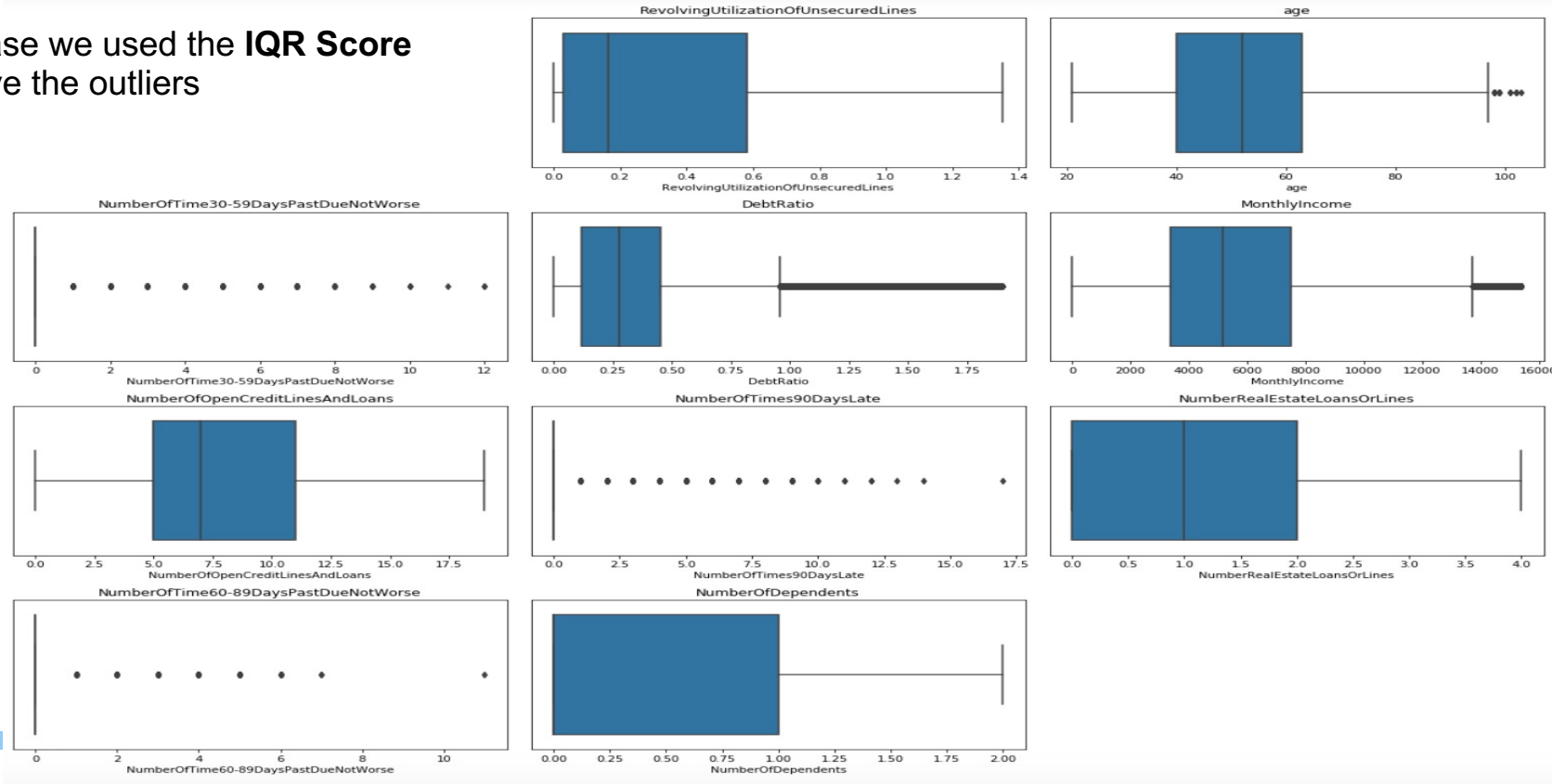
- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.



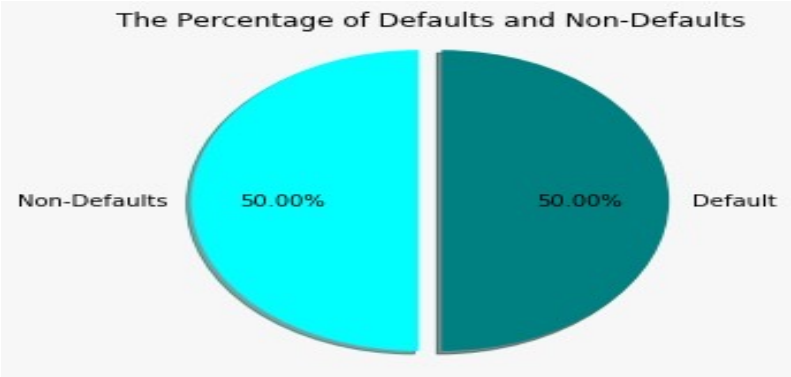

RevolvingUtilizationOfUnsecuredLines
0
10000
20000
30000
40000
50000
RevolvingUtilizationOfUnsecuredLines
age
0
20
40
60
80
100
age
NumberOfTime30-59DaysPastDueNotWorse
0
20
40
60
80
100
NumberOfTime30-59DaysPastDueNotWorse
DebtRatio
0
50000
100000
150000
200000
250000
300000
DebtRatio
MonthlyIncome
0
500000
1000000
1500000
2000000
2500000
3000000
MonthlyIncome
NumberOfOpenCreditLinesAndLoans
0
10
20
30
40
50
60
NumberOfOpenCreditLinesAndLoans
NumberOfTimes90DaysLate
0
20
40
60
80
100
NumberOfTimes90DaysLate
NumberRealEstateLoansOrLines
0
10
20
30
40
50
NumberRealEstateLoansOrLines
NumberOfTime60-89DaysPastDueNotWorse
0
20
40
60
80
100
NumberOfTime60-89DaysPastDueNotWorse
NumberOfDependents
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
NumberOfDependents

4. After handeling these cases

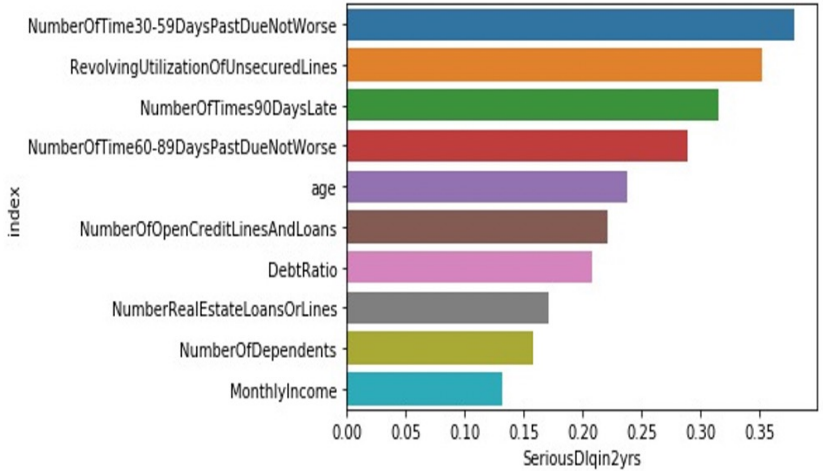
In our case we used the **IQR Score** to remove the outliers

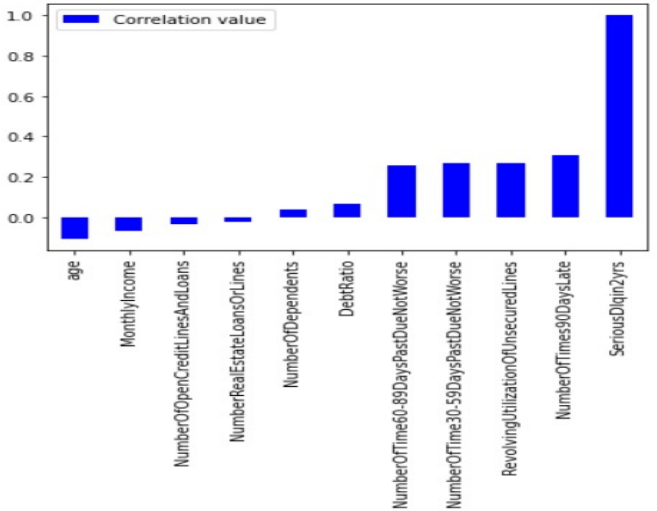


5. Trained data balancing



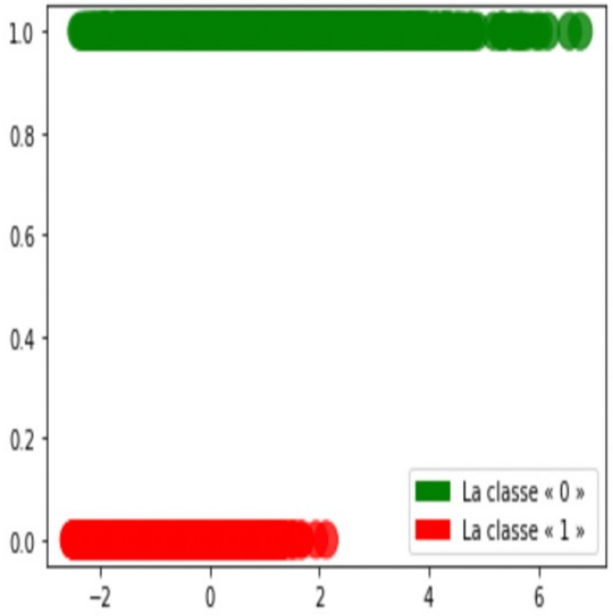
6. Identification of the best predictors among the variables



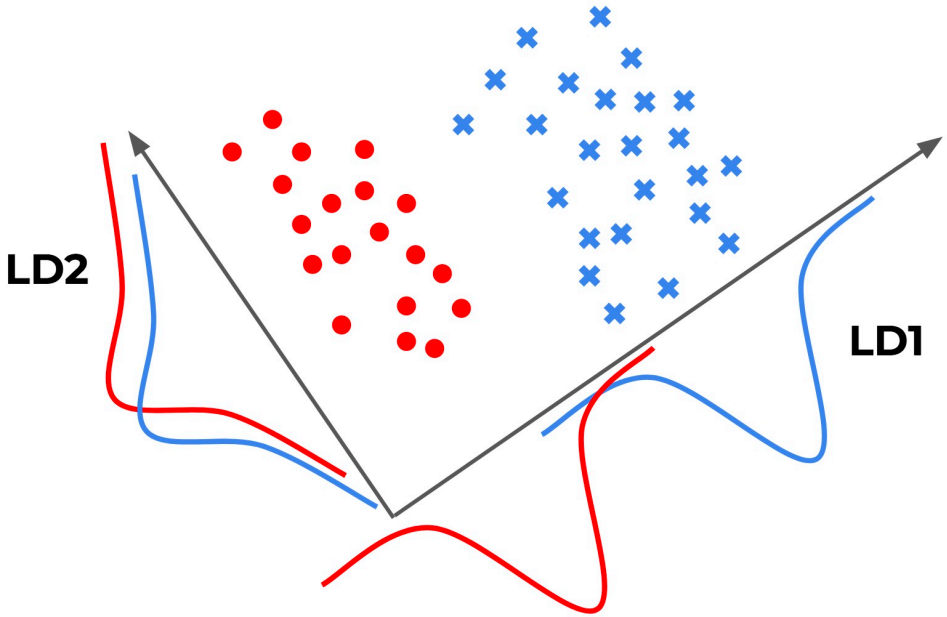


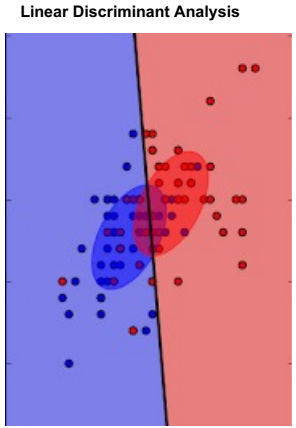
Chapter Three
Building models

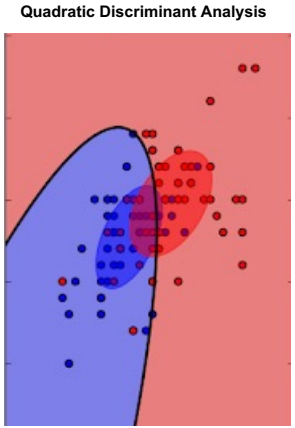
1. AFD



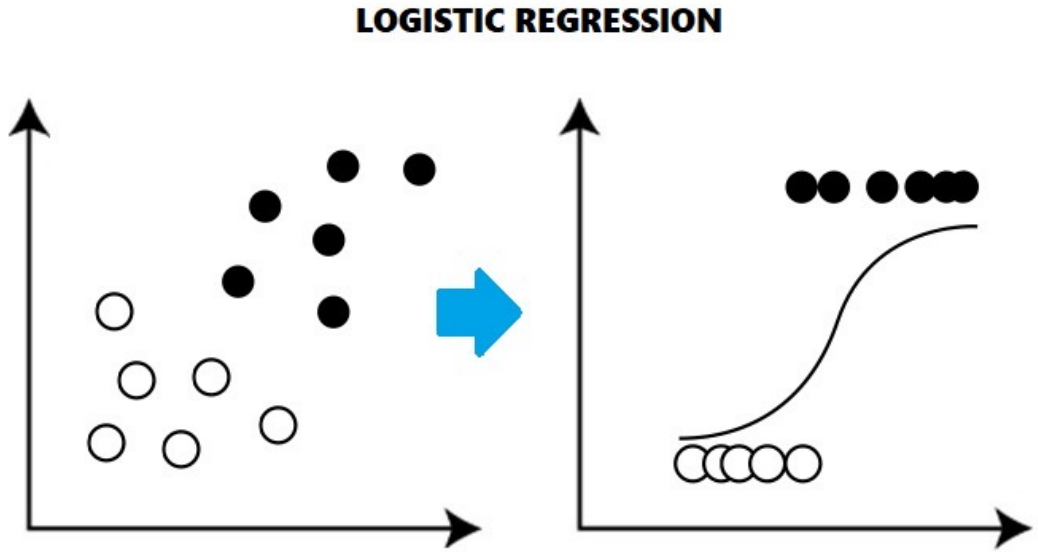
2. LDA and QDA


LD2
LD1

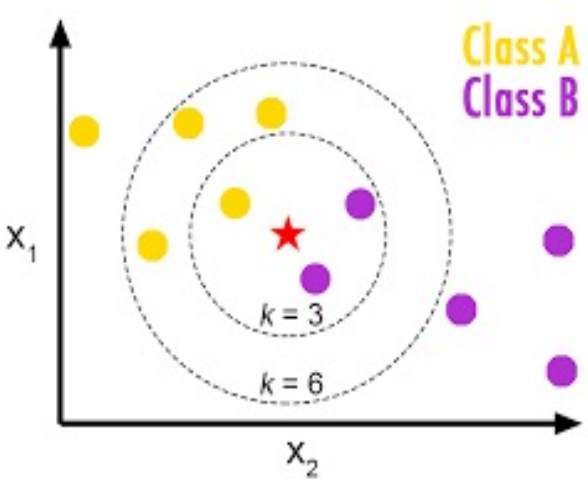

Linear Discriminant Analysis

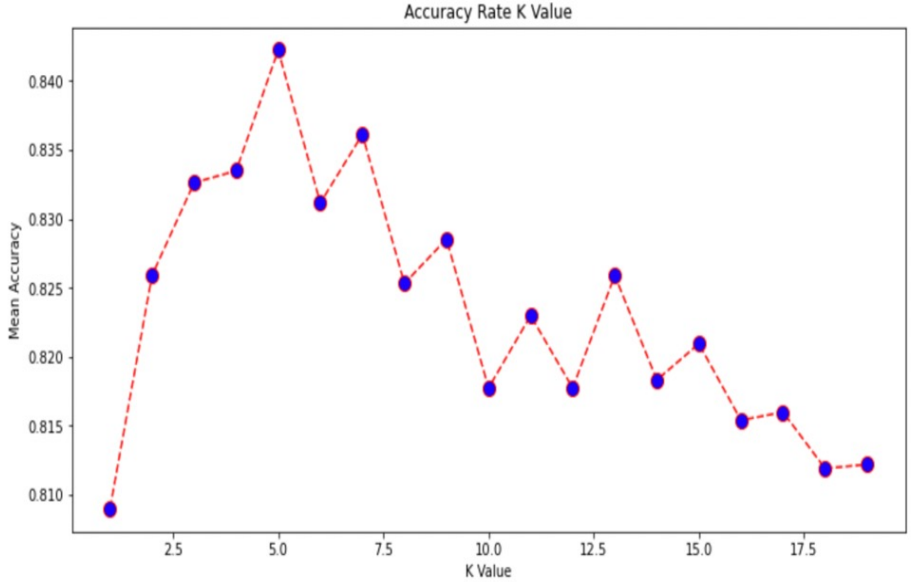

Quadratic Discriminant Analysis

3. Logistic regression

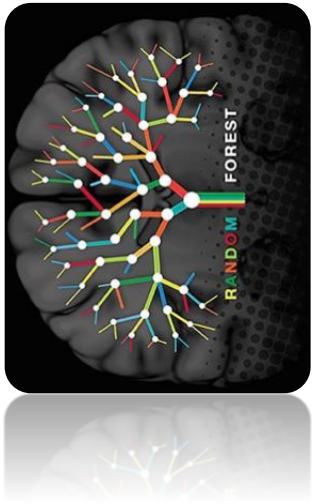


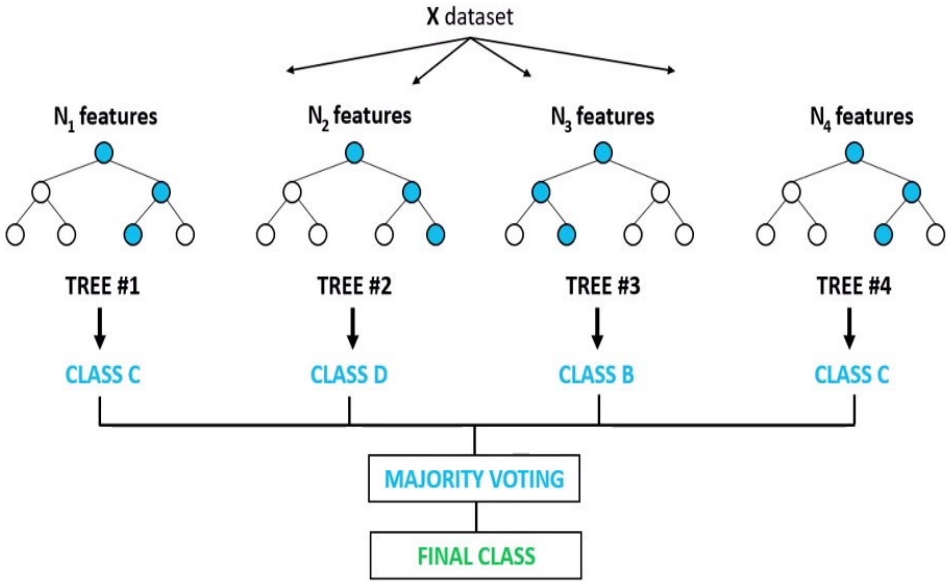
4. Nearest neighbors classifier



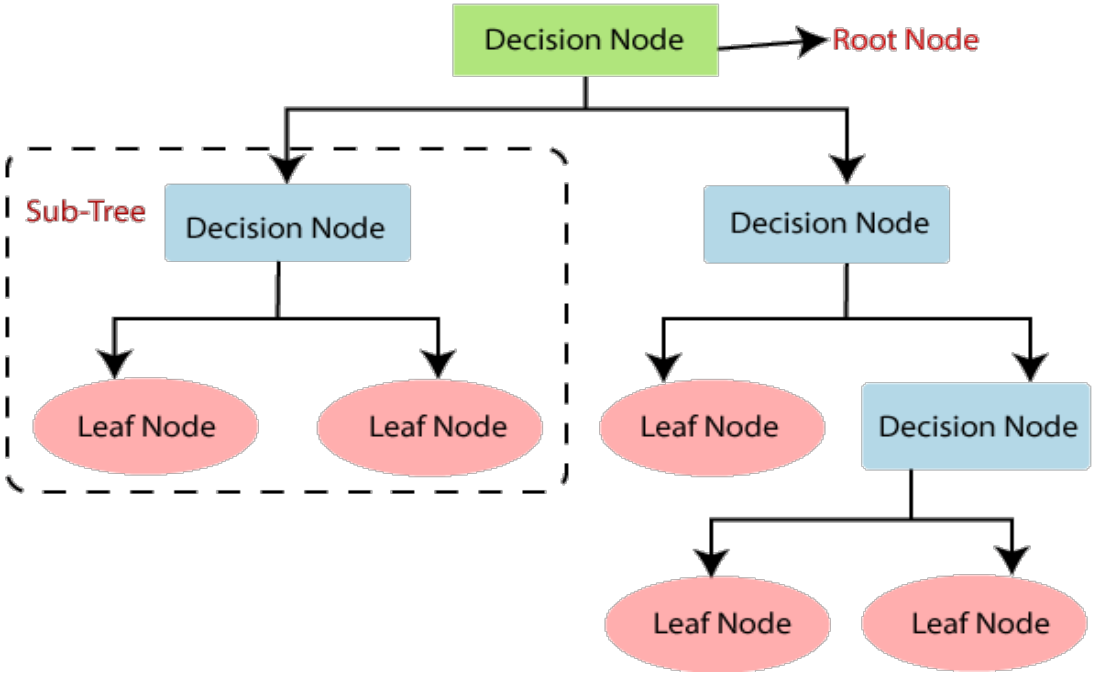


5. Random Forest Classifier

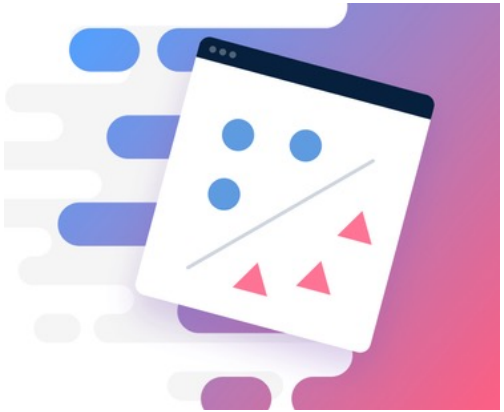


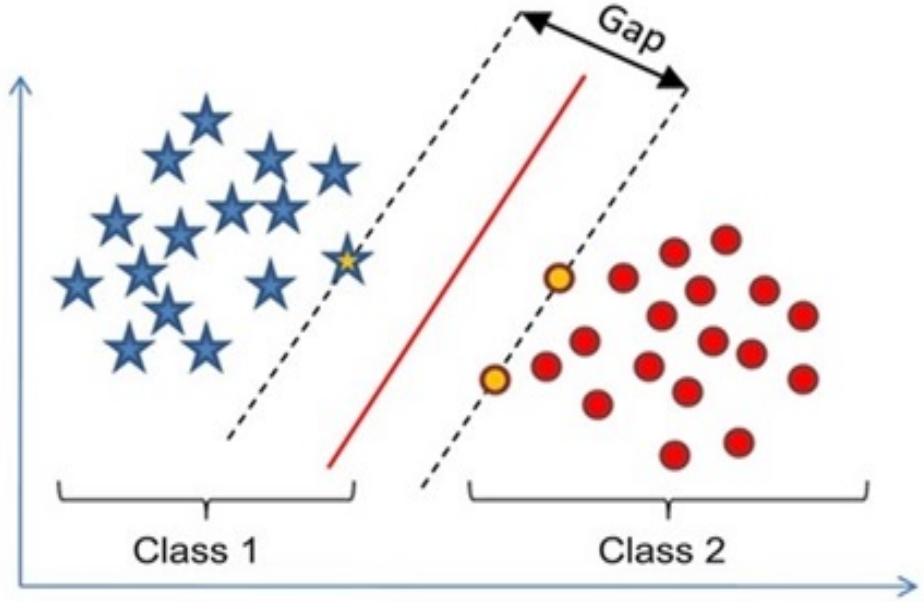


6. Decision Tree Classifier

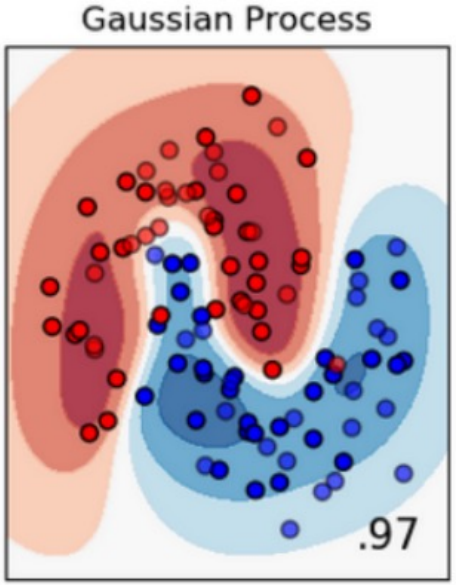


7. Support Vector Classifier





8. Gaussian Process Classifier



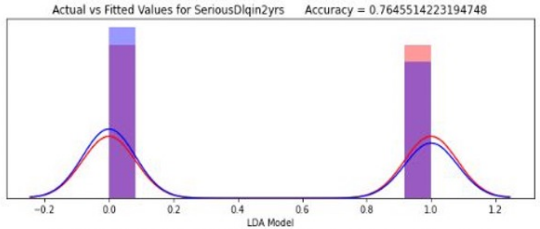
Chapter four
Evaluation phase

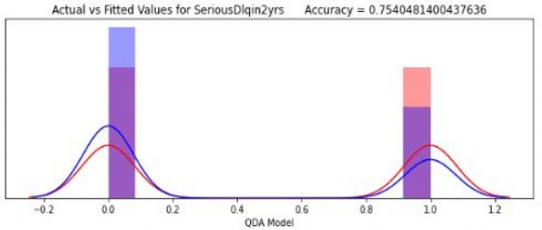
1. Comparison of Goodness of fit and the predictive power

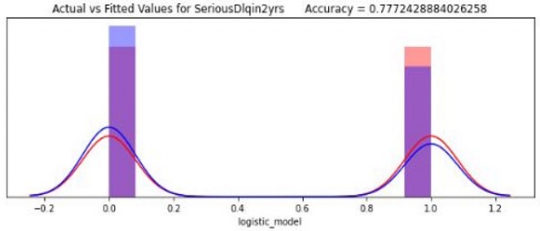
➔ Goodness of fit

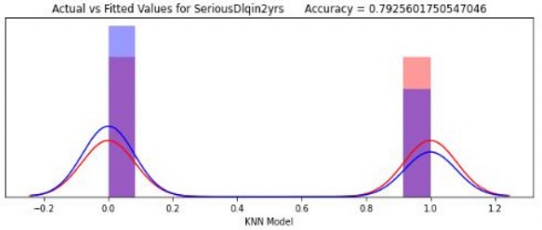
```
LDA model Mean Accuracy: 0.771 (0.001)
QDA model Mean Accuracy: 0.766 (0.003)
logistic regression model Mean Accuracy: 0.779 (0.012)
KNN model Mean Accuracy: 0.843 (0.004)
Random Forest model Mean Accuracy: 0.757 (0.005)
Decision Tree model Mean Accuracy: 0.770 (0.007)
Support Vector model Mean Accuracy: 0.810 (0.003)
Gaussian process model Mean Accuracy: nan (nan)
```

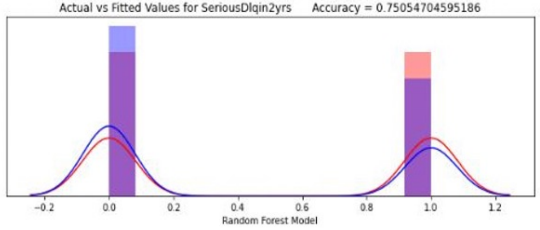
Prediction accuracy of seven models

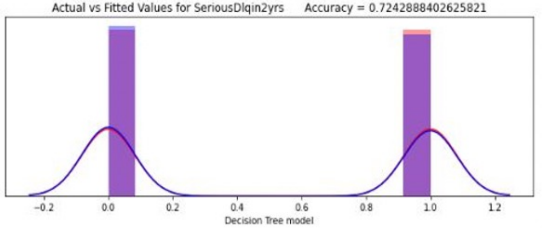


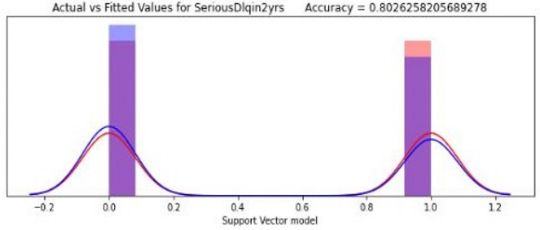






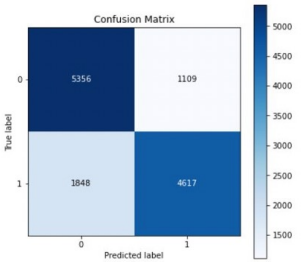




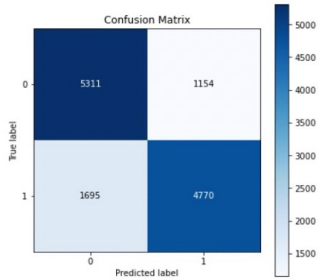


➔ Confusion Matrix

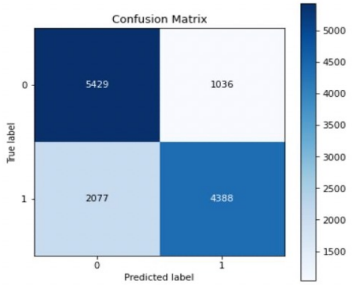
LDA Model



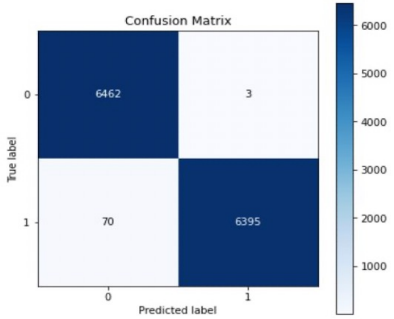
QDA Model



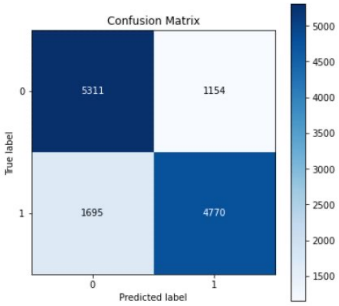
Random Forest Model



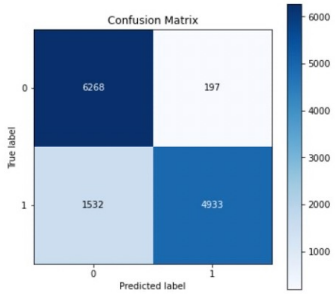
Decision Tree model



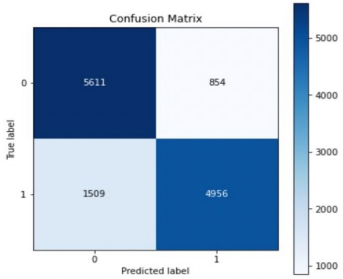
Logistic regression



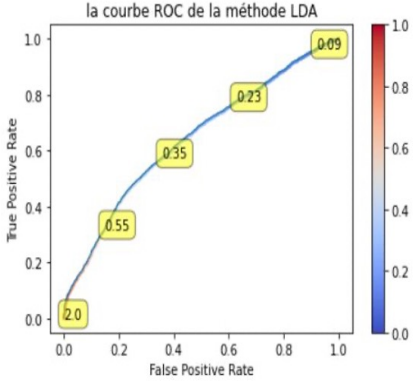
KNN Model

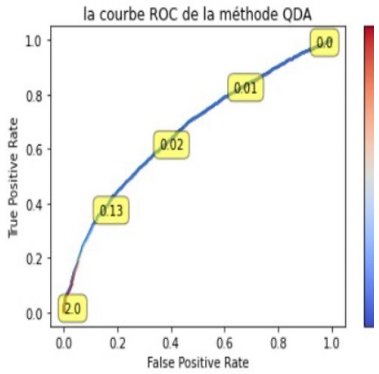


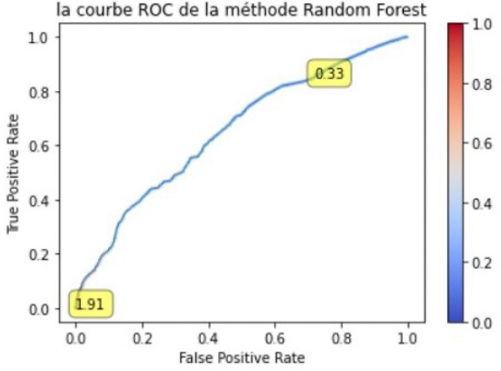
Support Vector model

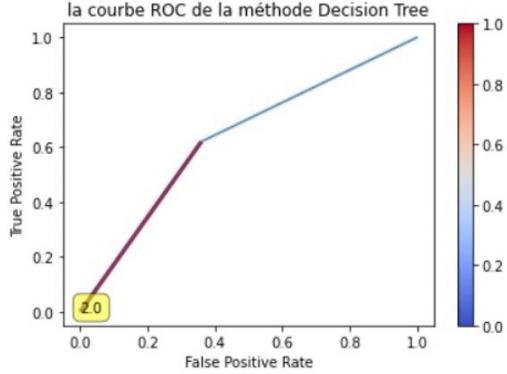


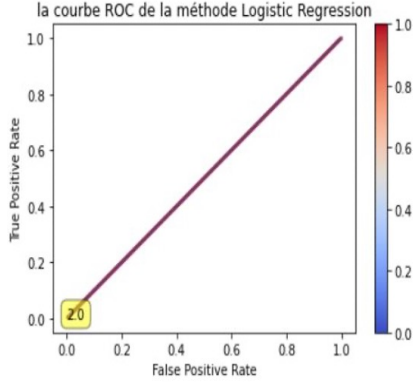
➔ Predictive Power

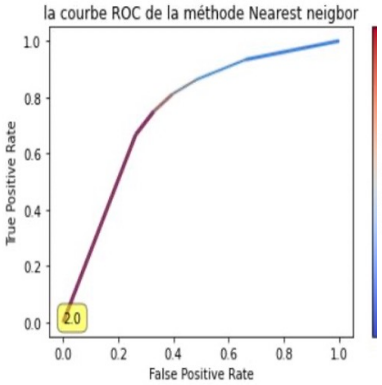


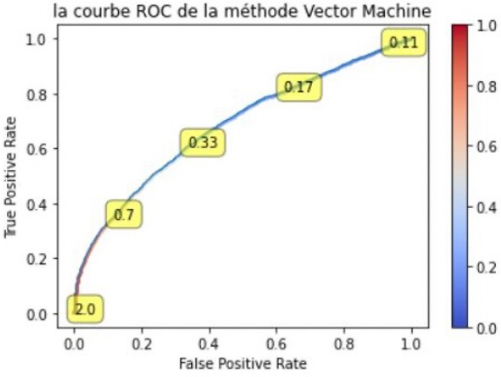


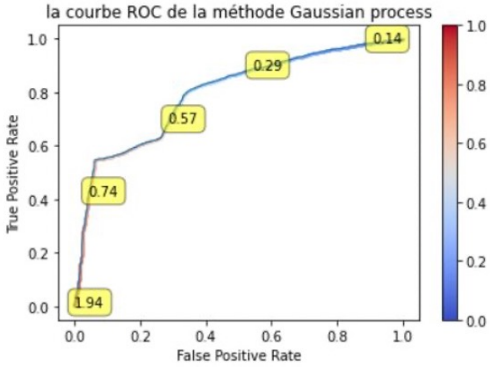






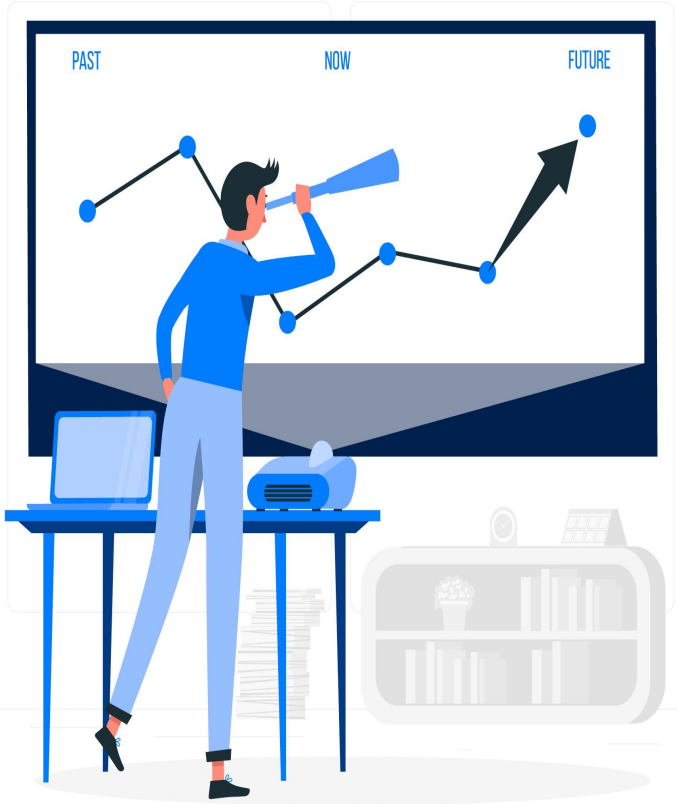






➔ AUC

```
le critére AUC pour la méthode LDA, LDA_AUC =  0.6438346639891717
le critére AUC pour la méthode QDA, QDA_AUC =  0.6703792285432792
le critére AUC pour la méthode Logistic Regression, LgRg_AUC =  0.50080054894785
le critére AUC pour la méthode KNN, KNN_AUC =  0.7472101115055552
le critére AUC pour la méthode Random Forest, RFS_AUC =  0.650905647809524
le critére AUC pour la méthode Decision Tree, DST_AUC =  0.6302016324460156
le critére AUC pour la méthode Support vector machine, SVM_AUC =  0.682953790192768
le critére AUC pour la méthode Gaussian process , GPC_AUC =  0.7965828770626352
```

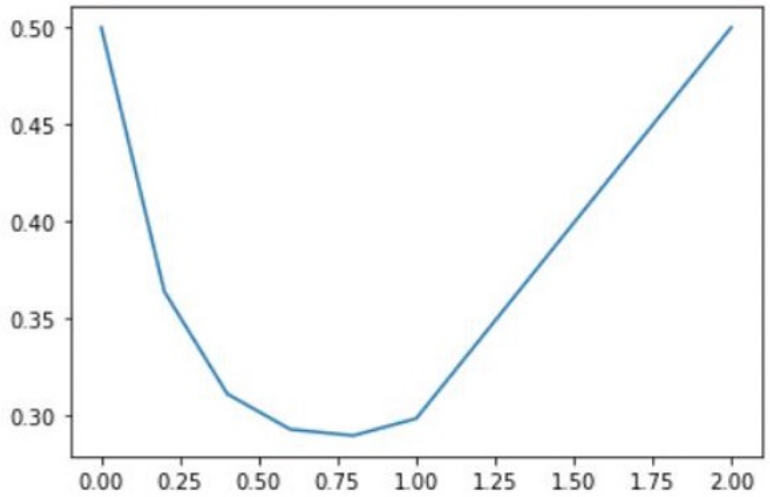


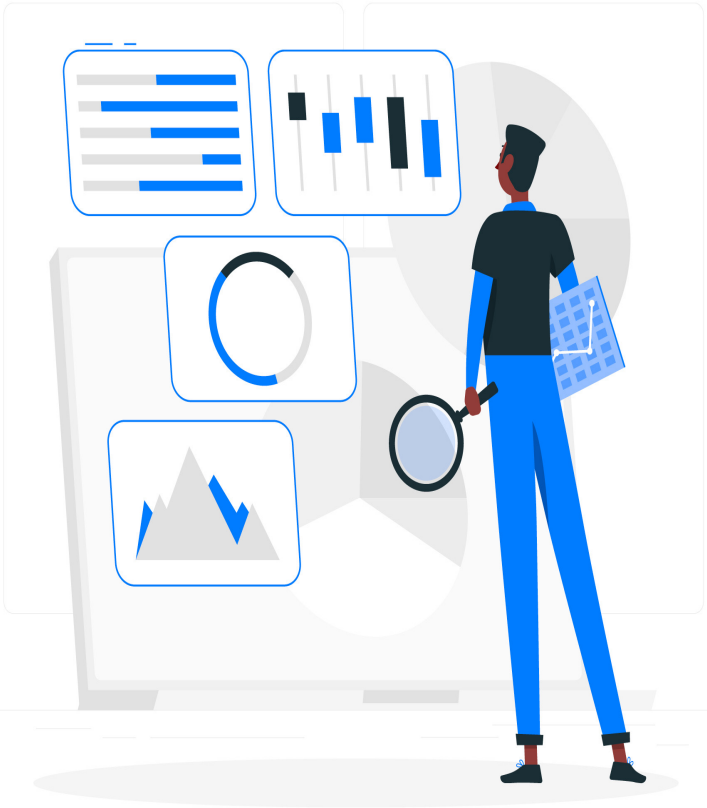
2. Most efficient model

Credit scoring using different machine learning algorithms are used by many lending organizations, to control and mitigate the credit risks arising out of a default. In this data analysis, Nearest Neighbors performed best for classification while the Regression model was the least helpful among all the models to classify customers into default and non-default set.

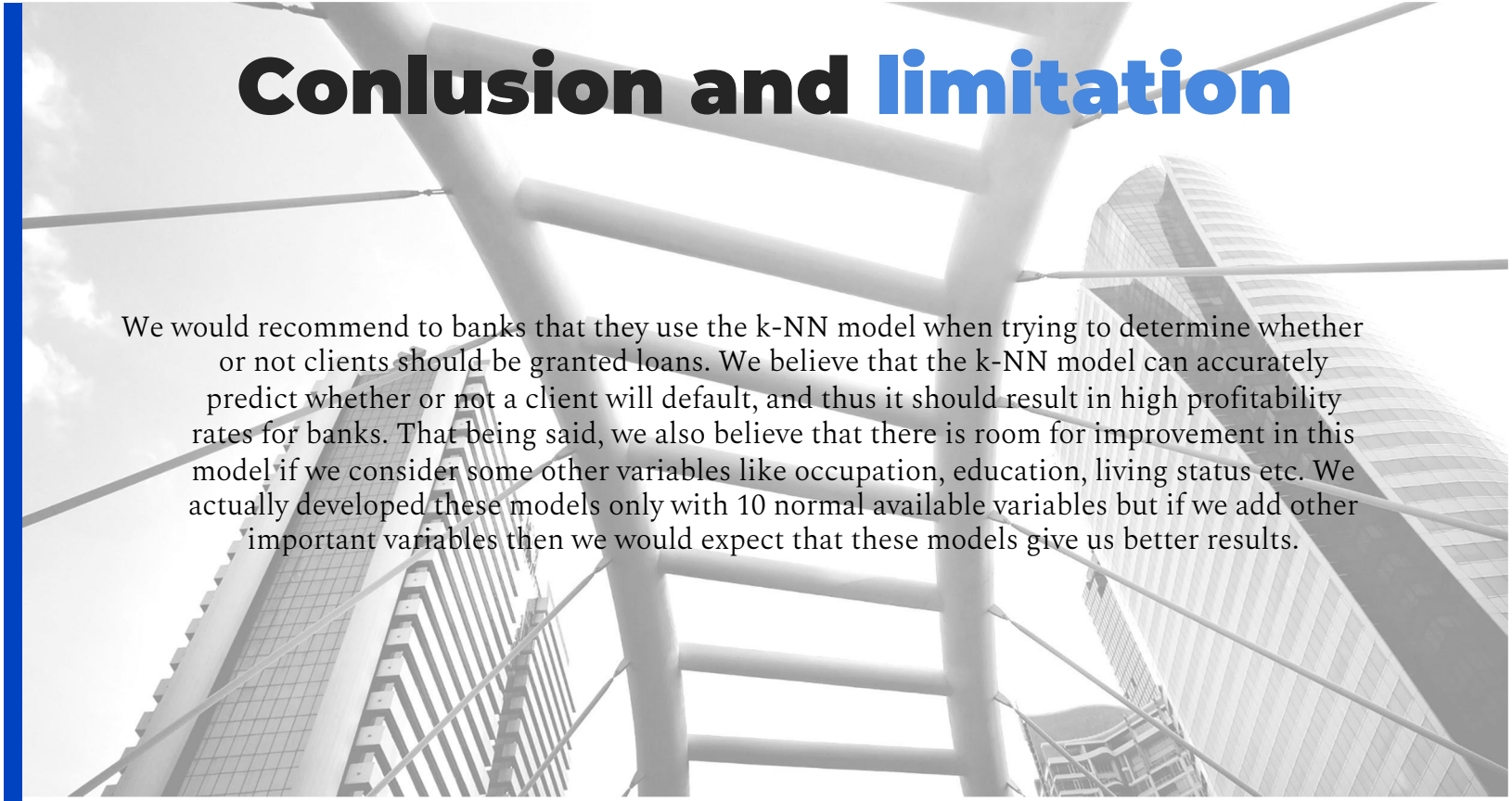
For the final model we chose to use K-Nearest Neighbors (k-NN). This model had the closest AUC-Score to 1.

3. The decision rule





Conlusion and limitation



We would recommend to banks that they use the k-NN model when trying to determine whether or not clients should be granted loans. We believe that the k-NN model can accurately predict whether or not a client will default, and thus it should result in high profitability rates for banks. That being said, we also believe that there is room for improvement in this model if we consider some other variables like occupation, education, living status etc. We actually developed these models only with 10 normal available variables but if we add other important variables then we would expect that these models give us better results.