Likelihood of defaulting credit card payments

Data Analytics BootCamp
June-2023

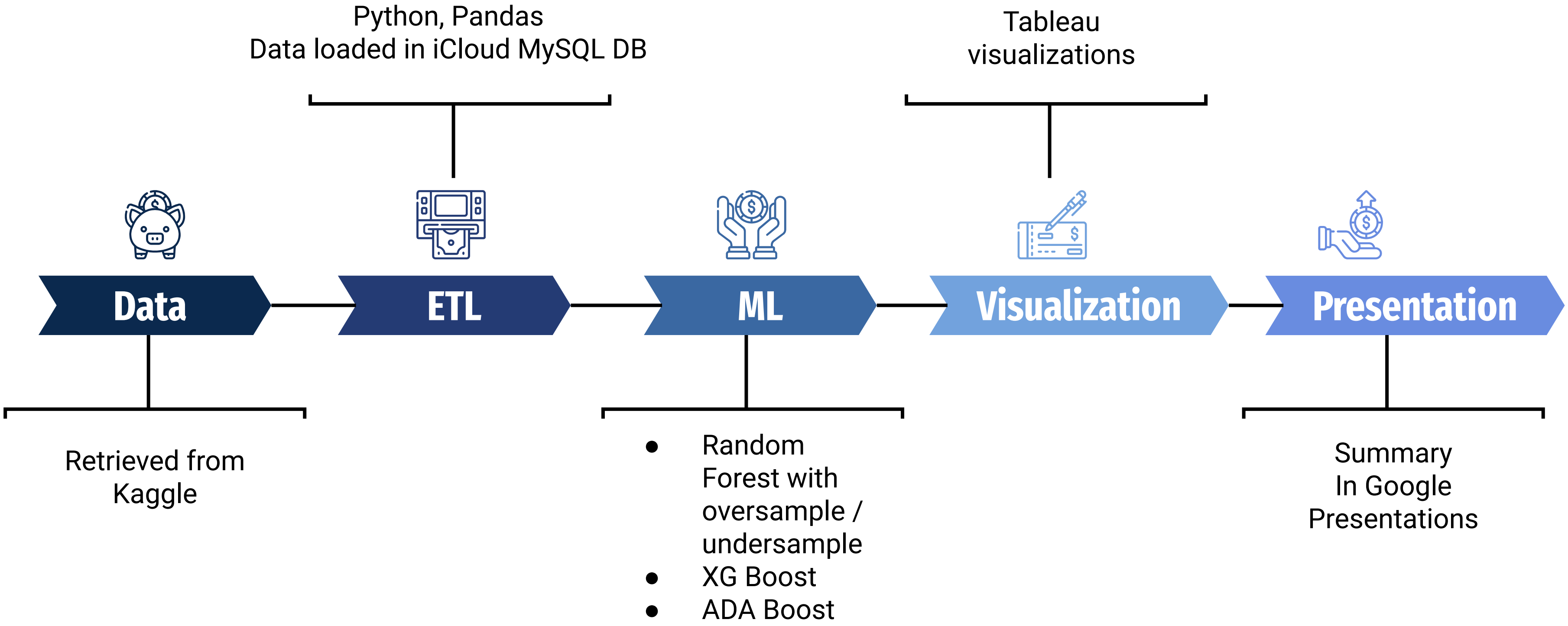
Alonso Lozano
Aldo Silva
Alejandra Espinosa
Marcela Maldonado

Project proposal

We have two goals:

1. Creating a set of visualizations in Tableau of the default and non-default credit card clients and the relationship that defaulting has with demographic features and the payment history of each customer.
2. Creating a supervised machine learning model to predict whether a credit card holder will be on default or not depending on their demographic profile and payment history

Project breakdown



Data

Retrieved from Kaggle

ETL

Python, Pandas
Data loaded in iCloud MySQL DB

ML

- Random Forest with oversample / undersample
- XG Boost
- ADA Boost

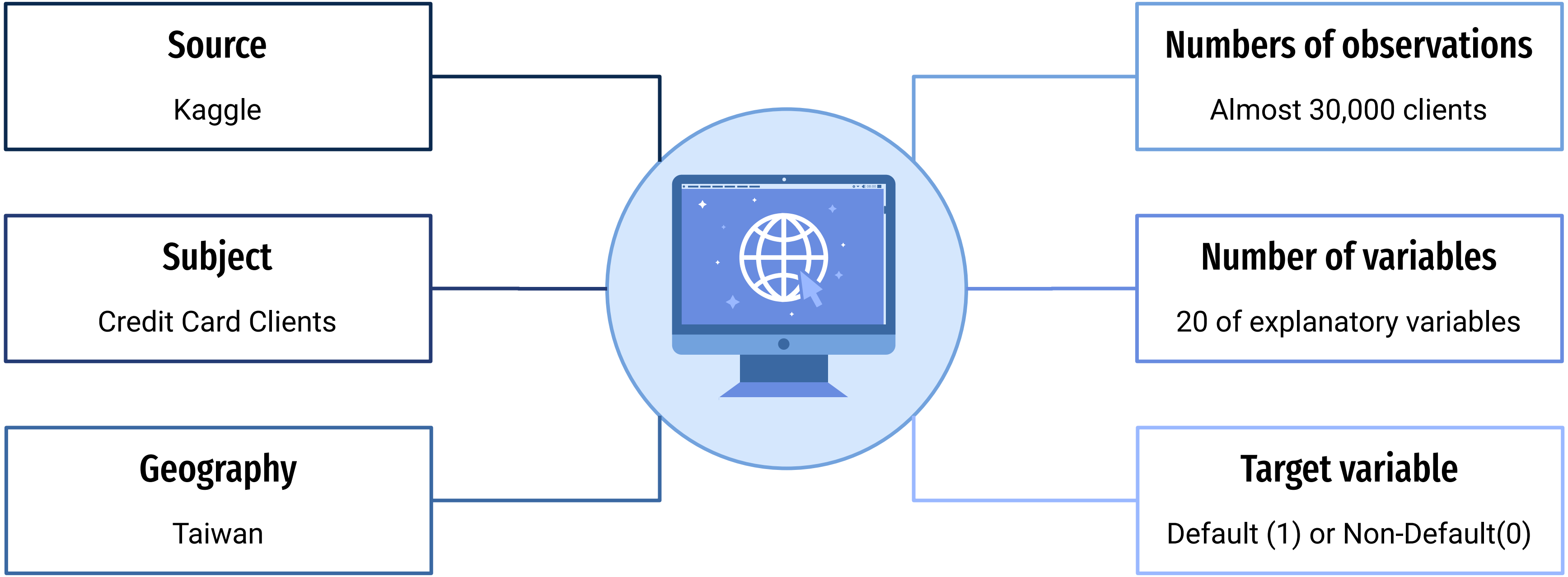
Visualization

Tableau visualizations

Presentation

Summary
In Google Presentations

Origin of the database



Source

Kaggle

Subject

Credit Card Clients

Geography

Taiwan

Numbers of observations

Almost 30,000 clients

Number of variables

20 of explanatory variables

Target variable

Default (1) or Non-Default(0)

Variables

Credit Limit

Float

Gender

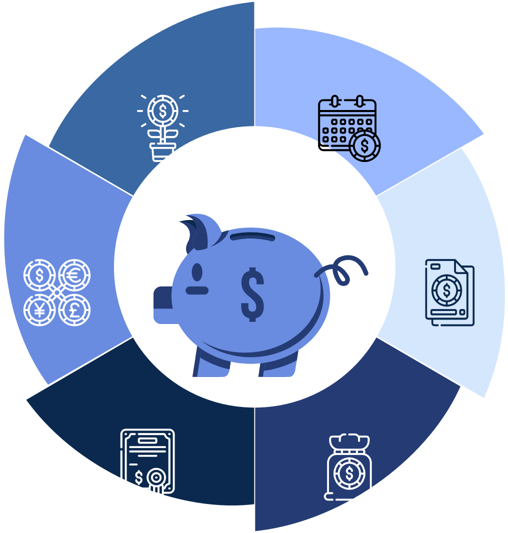
Categorical

Education

Categorical

Marital status

Categorical



Age

Categorical

Timeliness of past payments

Categorical

Amount of bill statement

Float

Amount of previous payment

Float

Default payment next month

Categorical

Data Cleanup, transform and preparation

Null Value Handling:
To ensure data integrity, null values are addressed by either removing or imputing missing values. ✓

Standardization:
Numeric features are standardized by scaling them to have zero mean and unit variance. Dummy variables are excluded from this process. ✓

Dummy Variable Creation:
Categorical variables are converted into binary variables, known as dummy variables, to effectively represent them in the dataset. ✓

Principal Component Analysis (PCA):
PCA is used to reduce the dimensionality of the dataset by transforming the original variables into uncorrelated principal components. This helps identify the most important features while minimizing information loss. ✗

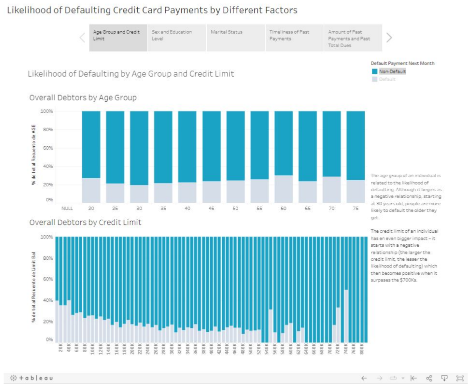
Client clusterization (Unsupervised machines learning):
We clusterized the dataset before running the model to confirm if this clusterizations could increase the accuracy of our models. ✗

Train-Test Data Split with Stratification:
The data is split into training and testing sets, stratified based on the target values. This ensures a balanced distribution of target values in both sets, reducing bias in subsequent analyses and model training. ✓

Tableau Story: Likelihood of defaulting credit card payments

1. Dashboard 1: By age group and credit limit
2. Dashboard 2: By sex and education level
3. Dashboard 3: By marital status
4. Dashboard 4: By timeliness of past payments
5. Dashboard 5: By amount of past payments and past total dues

Tableau Story: Likelihood of defaulting credit card payments


Age Group and Credit Limit
Sex and Education Level
Marital Status
Timeliness of Past Payments
Amount of Past Payments and Past Total Dues
Default Payment Next Month
Non-Default
Default
% de total Recuento de AGE
100%
80%
60%
40%
20%
0%
NULL
20
25
30
35
40
45
50
55
60
65
70
75
% de total Recuento de Limit Bal
+ a b | e a u

Likelihood of Defaulting Credit Card Payments by Different Factors

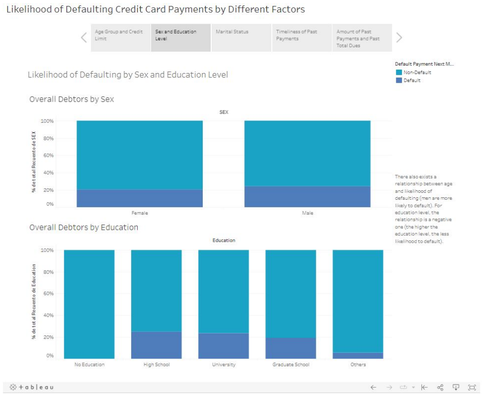
Likelihood of Defaulting by Age Group and Credit Limit

Overall Debtors by Age Group

The age group of an individual is related to the likelihood of defaulting. Although it begins as a negative relationship, starting at 30 years old, people are more likely to default the older they get.

Overall Debtors by Credit Limit

The credit limit of an individual has en even bigger impact – it starts with a negative relationship (the larger the credit limit, the lesser the likelihood of defaulting) which then becomes positive when it surpases the $700Ks.


Age Group and Credit Limit
Sex and Education Level
Marital Status
Timeliness of Past Payments
Amount of Past Payments and Past Total Dues
Default Payment Next M...
Non-Default
Default
SEX
% de total Recuento de SEX
100%
80%
60%
40%
20%
0%
Female
Male
Education
% de total Recuento de Education
No Education
High School
University
Graduate School
Others
+ a b | e a u

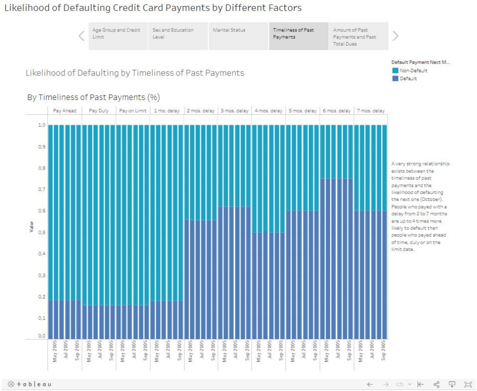
Likelihood of Defaulting Credit Card Payments by Different Factors

Likelihood of Defaulting by Sex and Education Level

Overall Debtors by Sex

There also exists a relationship between age and likelihood of defaulting (men are more likely to default). For education level, the relationship is a negative one (the higher the education level, the less likelihood to default).

Overall Debtors by Education

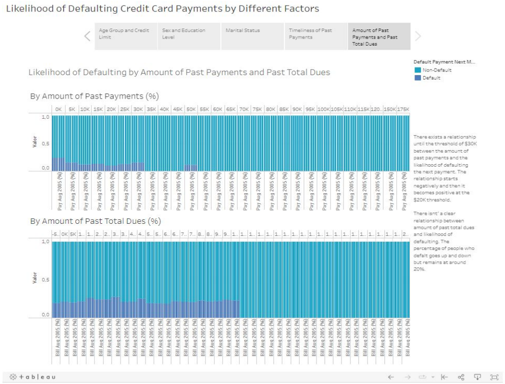


Likelihood of Defaulting Credit Card Payments by Different Factors

Likelihood of Defaulting by Timeliness of Past Payments

By Timeliness of Past Payments (%)

A very strong relationship exists between the timeliness of past payments and the likelihood of defaulting the next one (October). People who payed with a delay from 2 to 7 months are up to 4 times more likely to default than people who payed ahead of time, duly or on the limit date.



Likelihood of Defaulting Credit Card Payments by Different Factors

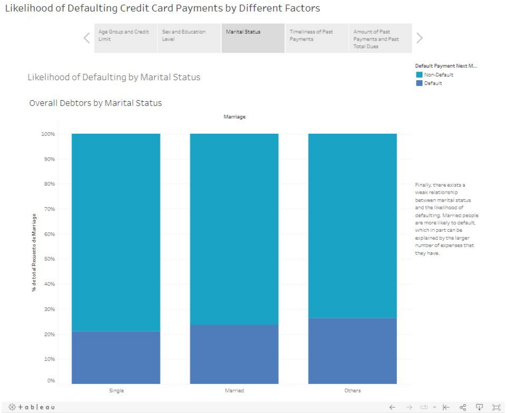
Likelihood of Defaulting by Amount of Past Payments and Past Total Dues

By Amount of Past Payments (%)

There exists a relationship until the threshold of $30K between the amount of past payments and the likelihood of defaulting the next payment. The relationship starts negatively and then it becomes positive at the $20K threshold.

By Amount of Past Total Dues (%)

There isnt' a clear relationship between amount of past total dues and likelihood of defaulting. The percentage of people who defalt goes up and down but remains at around 20%.


Age Group and Credit Limit
Sex and Education Level
Marital Status
Timeliness of Past Payments
Amount of Past Payments and Past Total Dues
Default Payment Next M...
Non-Default
Default
Marriage
% de total Recuento de Marriage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Single
Married
Others
+ a b | e a u

Likelihood of Defaulting Credit Card Payments by Different Factors

Likelihood of Defaulting by Marital Status

Overall Debtors by Marital Status

Finally, there exists a weak relationship between marital status and the likelihood of defaulting. Married people are more likely to default, which in part can be explained by the larger number of expenses that they have.

Data Conclusions

Predicting credit card default is hard.

There are certain variables with a strong relationship, such as:

- Credit limit
- Timeliness of past payments

Some variables have a mild relationship, such as:

- Age group
- Sex
- Education level
- Amount of past payments

Some variables have a weak relationship or no relationship at all:

- Marital status
- Amount of past total dues

Despite of these challenges, we still managed to create a predictive model with acceptable accuracy.

Finding variables with more explanatory power might yield better results.

We applied 4 models to ensure the best performance

The accuracy of the models was compromised due to the exclusion of PCA and clusterization techniques, which were not incorporated into the analysis process.

Random forest with undersampler

Technique to address class imbalance by reducing the number of majority class instances.

XG Boost

Efficient gradient boosting algorithm that sequentially adds decision trees to improve prediction accuracy. It is known for its high performance and effectiveness in various machine learning tasks

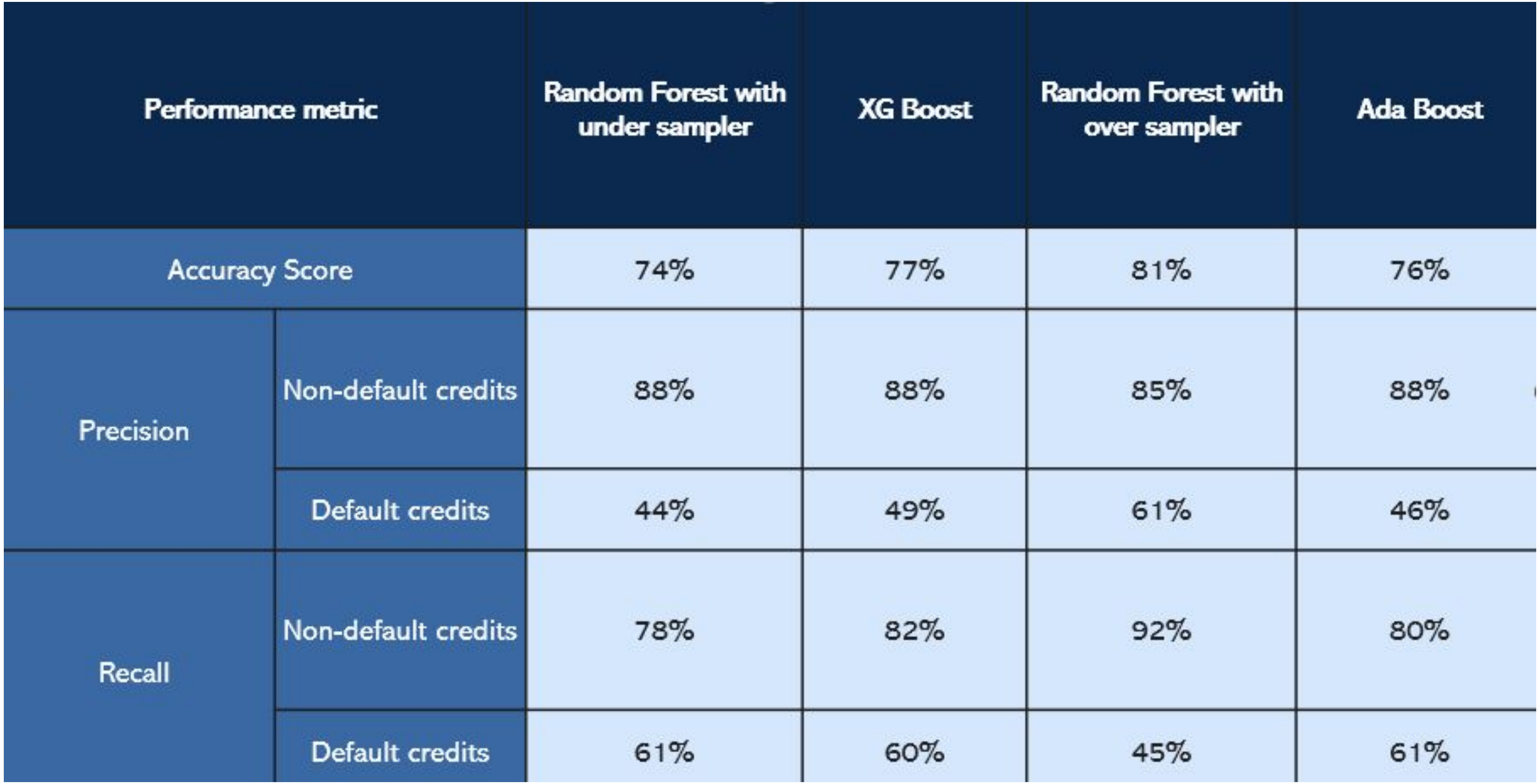
Random forest with oversampler

Similar to the first model, this approach uses the Random Forest algorithm, but instead employs an oversampling technique to increase the number of minority class instances and balance the class distribution.

ADA Boost

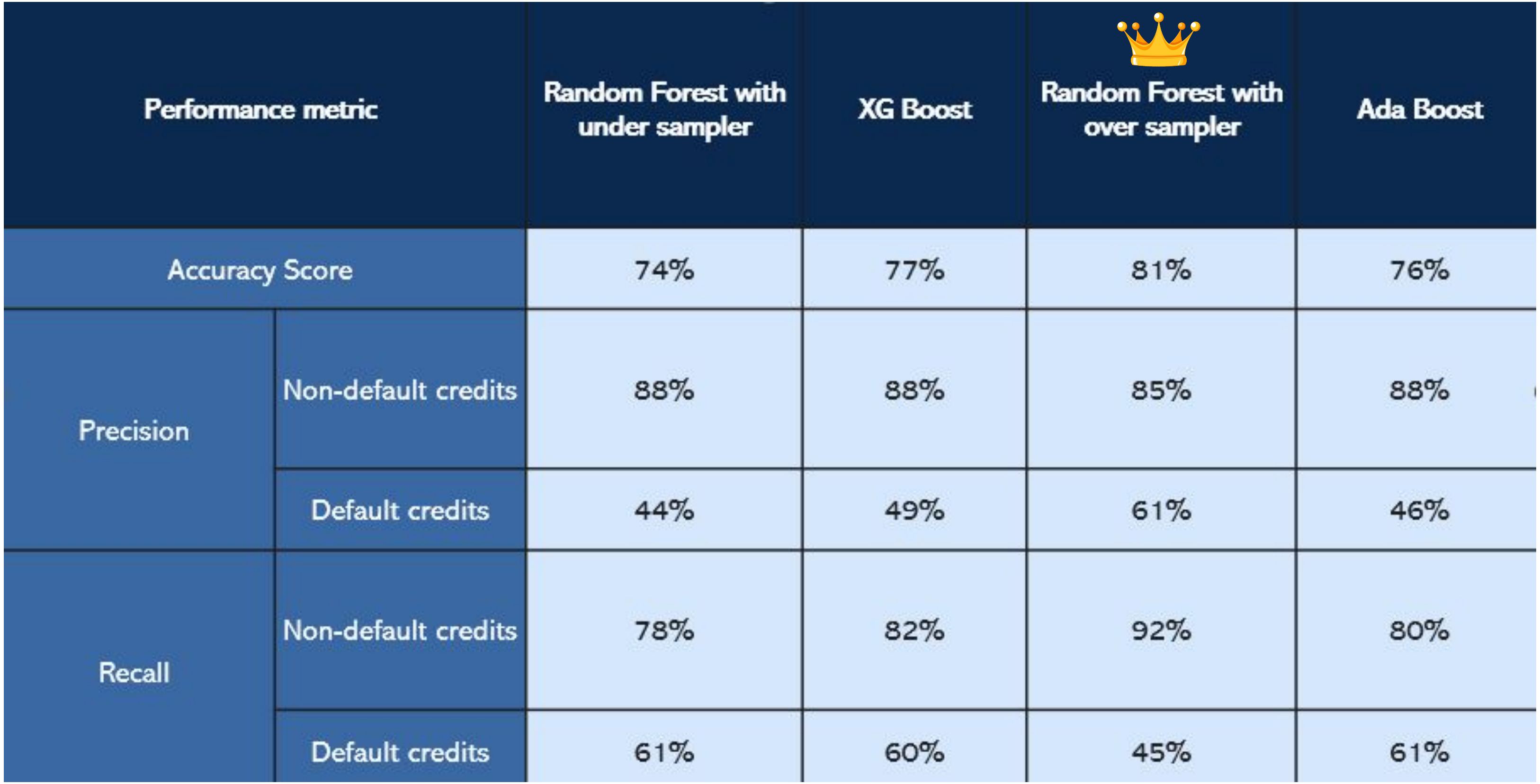
Ensemble learning method that combines weak learners, such as decision trees, to create a strong learner. It iteratively adjusts the weights of misclassified instances to focus on difficult samples and improve overall model performance.

Model performance evaluation



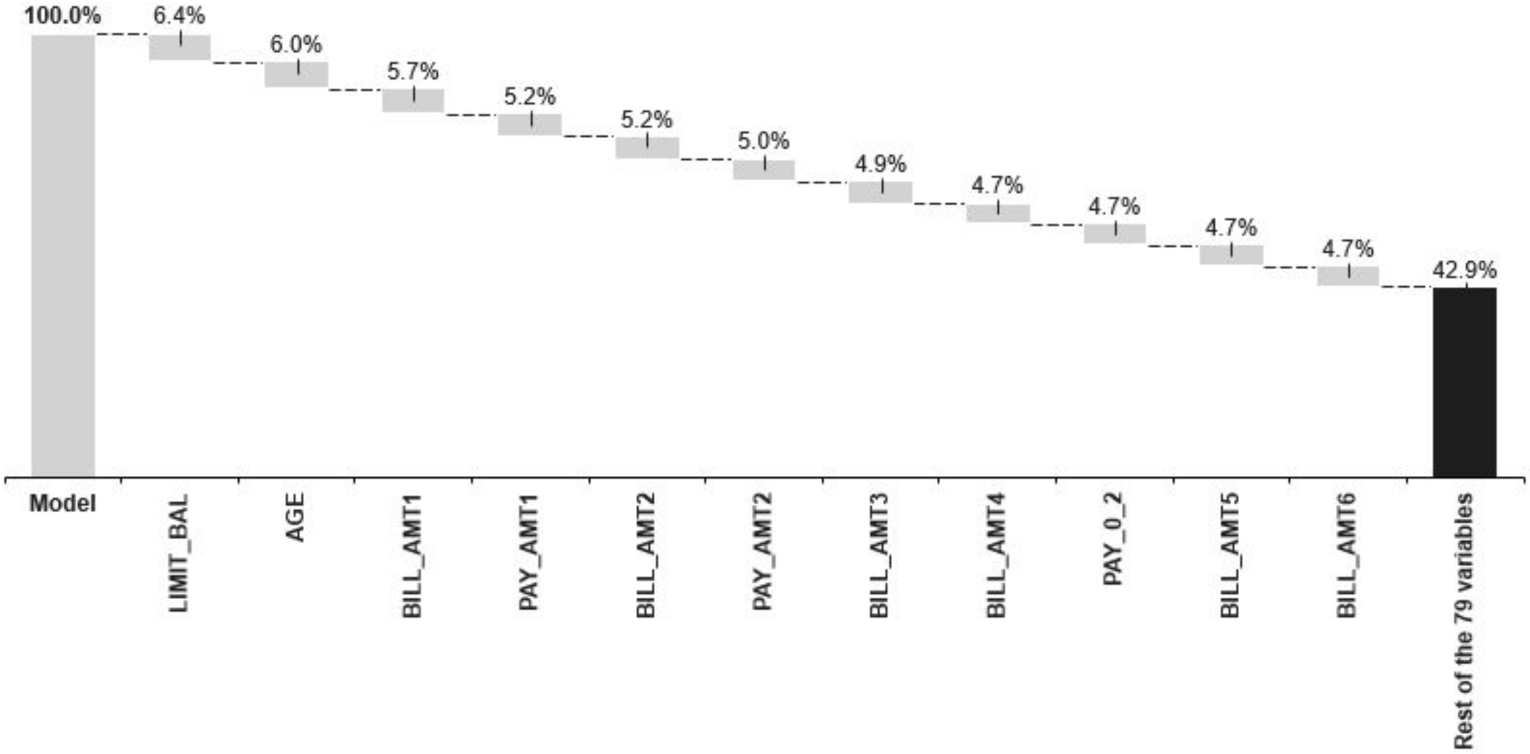
Performance metric		Random Forest with under sampler	XG Boost	Random Forest with over sampler	Ada Boost
Accuracy Score		74%	77%	81%	76%
Precision	Non-default credits	88%	88%	85%	88%
	Default credits	44%	49%	61%	46%
Recall	Non-default credits	78%	82%	92%	80%
	Default credits	61%	60%	45%	61%

Model performance evaluation



Performance metric		Random Forest with under sampler	XG Boost	👑 Random Forest with over sampler	Ada Boost
Accuracy Score		74%	77%	81%	76%
Precision	Non-default credits	88%	88%	85%	88%
	Default credits	44%	49%	61%	46%
Recall	Non-default credits	78%	82%	92%	80%
	Default credits	61%	60%	45%	61%

12 features represent 62% of the feature importance of the model

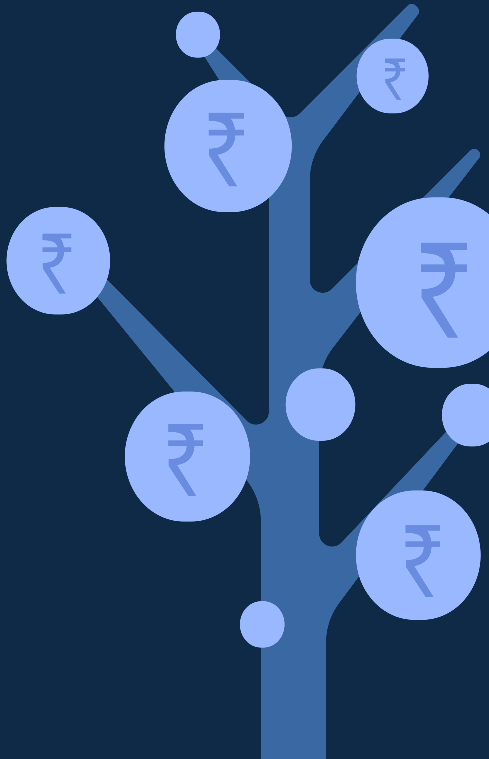


Overall Conclusions

Predicting credit card default is hard.

Despite the data challenges, we still managed to create a predictive model with acceptable accuracy.

Finding variables with more explanatory power might yield better results.



References

Bansodesandeep. (2022). Credit Card Default Prediction. *Kaggle*. https://www.kaggle.com/code/bansodesandeep/credit-card-default-prediction/notebook