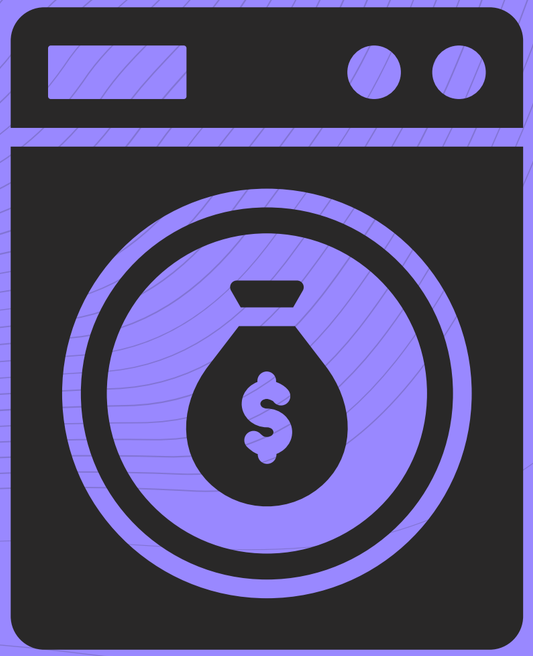
Università di Genova
DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

BUSINESS ANALYSIS PROJECT : BANK ACCOUNT FRAUD

by
Nicolò Guainazzo

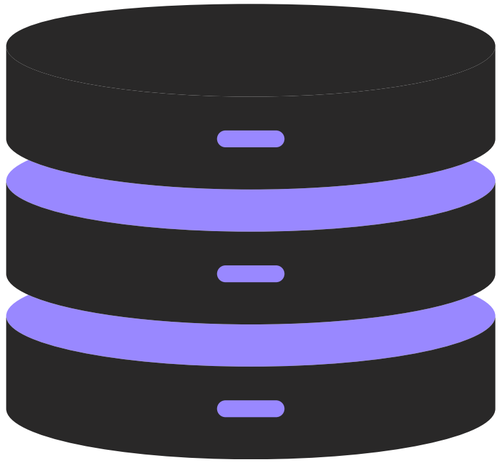
Understand the problem: money laundering

The problem lies in deciding whether a bank account opening request is legal or fraudulent. It is of paramount importance, in my opinion, for our domain (banking domain) to have the least number of false positives.



UniGe | DIBRIS

Bank Account Fraud Dataset



UniGe | DIBRIS

OVERVIEW

- Size: 1 milion rows x 32 features
- No missing values (only columns with all values equal to zero)
- Highly imbalanced (approx. 1% fraud labels)
- Low correlation between the features
- Found on Kaggle

ROWS

Each row in the dataset represents a request ,for a new bank account, labeled as good or fraudulent.

COLUMNS

5 categorical feature(*payment_type*, *source*, *employment_status*, *housing_status*, *device_os*) 28 binary or numerical feature (e.g. *income*, *foreign_request*, *session_lenght_in_minute*).

Problem definition in Machine learning terms

- The problem is a **supervised problem** and in general is a **unbalanced binary classification** problem.
- Many types of model can be use (*Gradient boost*, *ANN*, *One-class, isolation forest*).

UniGe | DIBRIS

GIVEN OUR PROBLEM:

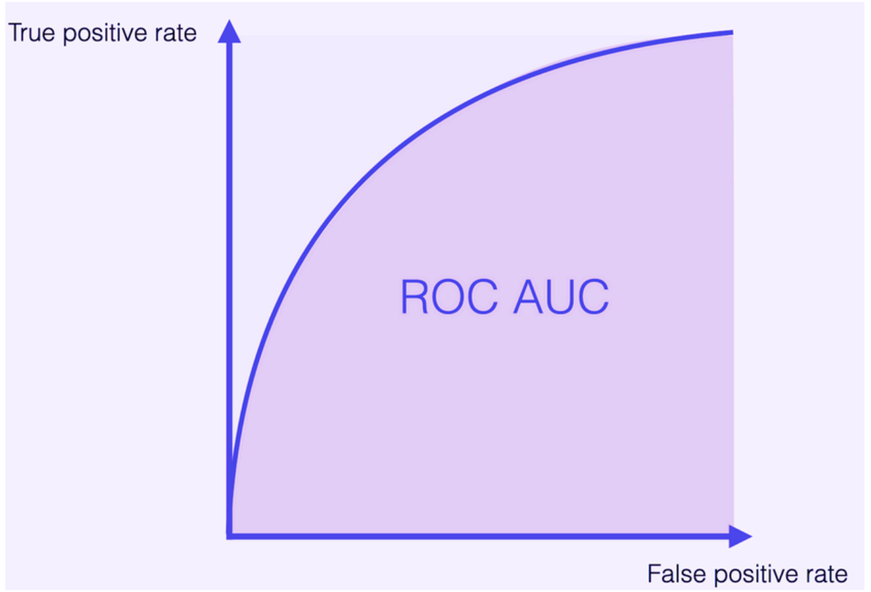
I chose the Random Forest model because, first of all, it does **not require data preparation** and also it is **robust and efficient**. In general, it is known to **work well with unbalanced data** as in our case.

UniGe | DIBRIS

Metrics

Confusion Matrix, Roc curve, Auc

Because of the imbalance of the data set, we cannot use the usual metrics (e.g. *accuracy*) to evaluate the model.



UniGe | DIBRIS

RESULTS

1st ATTEMPT

Try using the raw data

Achieve Auc = 0.85 and a very large number of false positives

2nd ATTEMPT

Try scaling, delete categorical features and weight of classes

Achieve Auc = 0.88 and a better balance between false and true positives

3rd ATTEMPT

Try with resampling

Undersampling done using NearMiss method.
Achieve Auc = 0.94 and detect more true positives than false.

4th ATTEMPT

Try with oversampling

Oversampling to rebalanced the dataset done with SMOTE method.
Achieve Auc = 0.87 but not a good number of false positives

UniGe | DIBRIS

Are the results as I expected ?

In general, yes, I did not expect to have striking results.

UniGe | DIBRIS

What's Next? How to develop my project

Finding other ways to decrease false positives.

Understand which features are more important for the prediction.

Increase the number of trees in the Random Forest model

Compare the results with other results obtained with different model.

UniGe | DIBRIS

Time spend

for every main steps in hours