Credit default prediction

Omar LAOUAN

Agenda

01

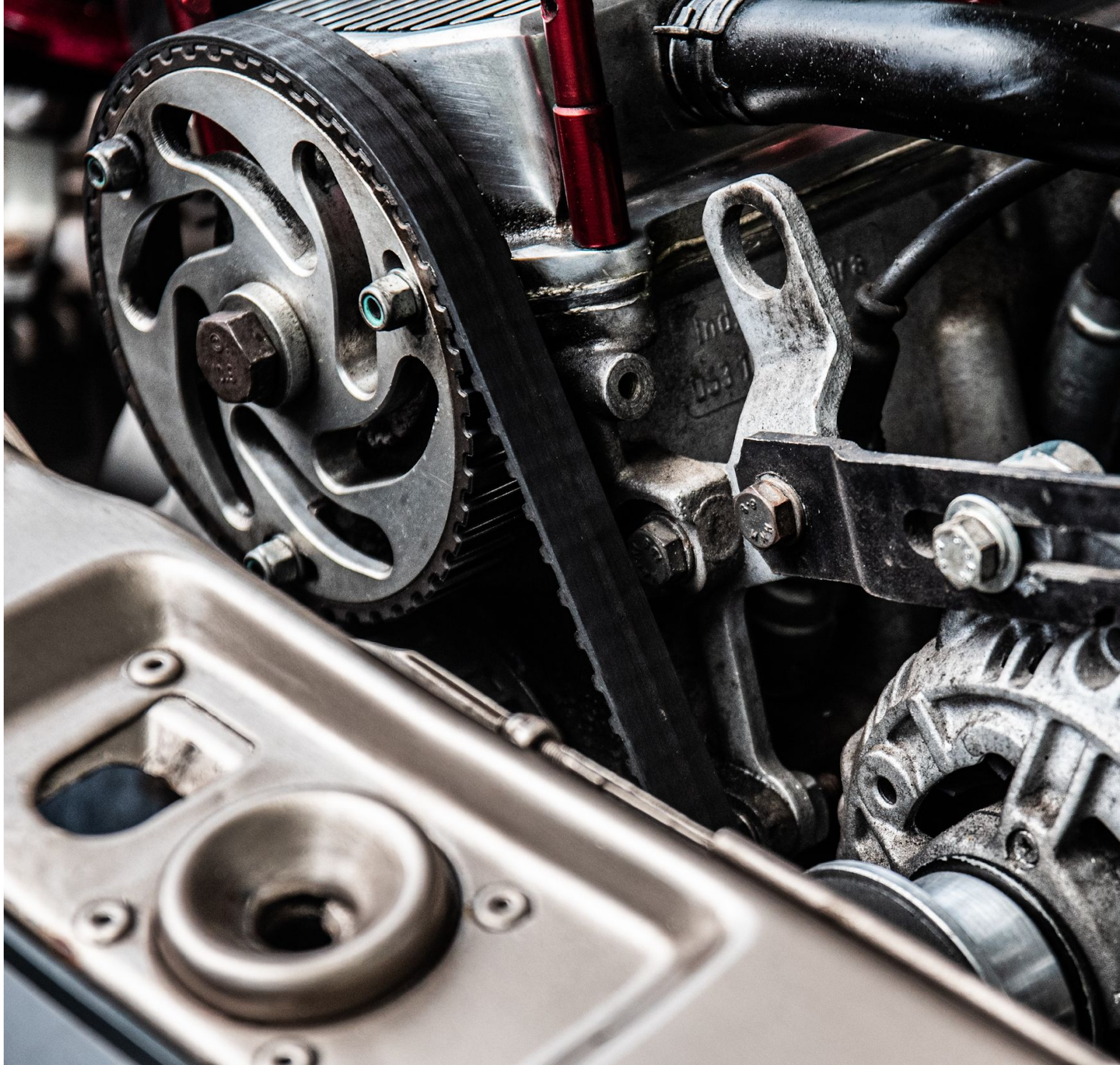
Context and objective

Context and objective

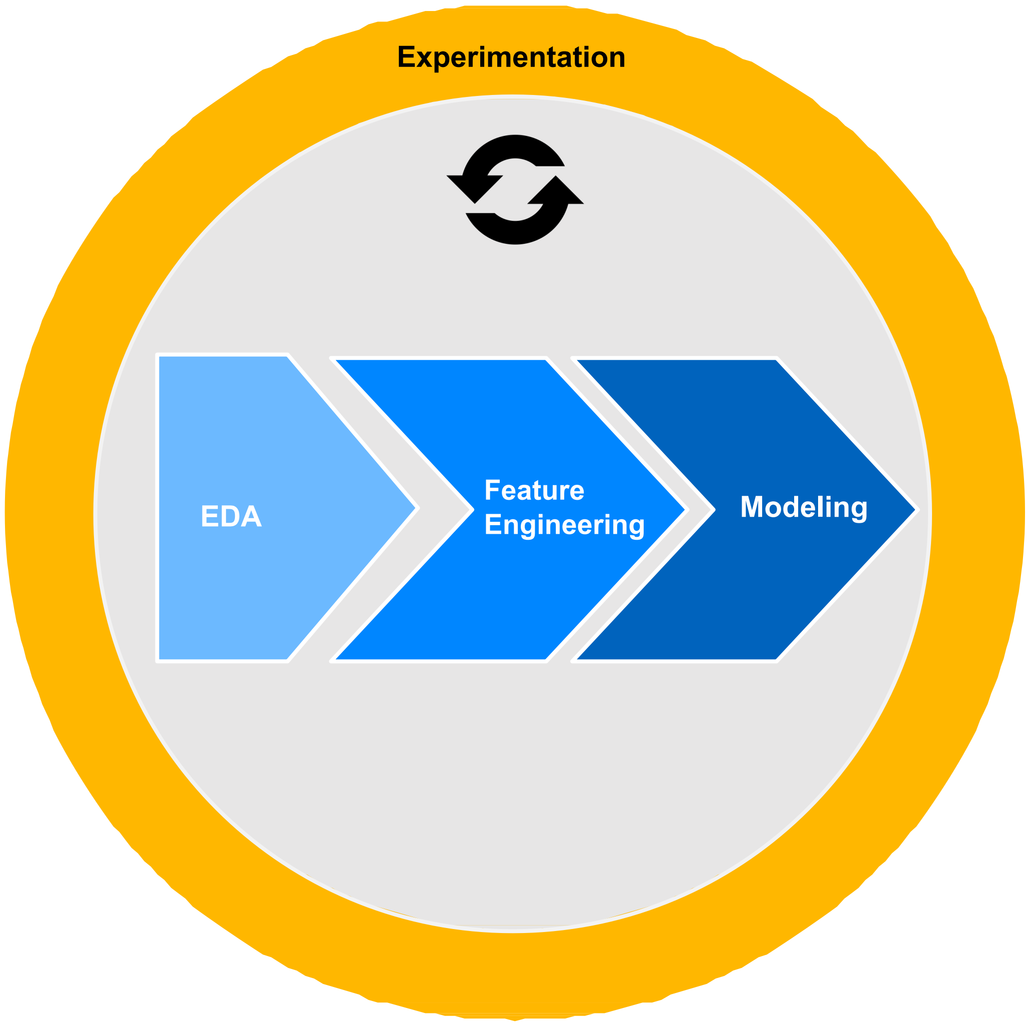
- This is a machine learning project that aims to predict whether a customer will default on a loan within 60 days of disbursement.
- The project uses historical customer financial data to train and evaluate several machine learning models and select the best-performing one.
- The target we aim to predict has two values:
  - 0 - Non default
  - 1- Default
- So it is a binary classification problem

02

Methodology overview



Methodology overview



EDA
- Checking for irregularities (null values, outliers)
- Understanding the distribution of the data
- Looking for relationship between variables

Feature engineering:
- Creating new features
- Extracting the most important

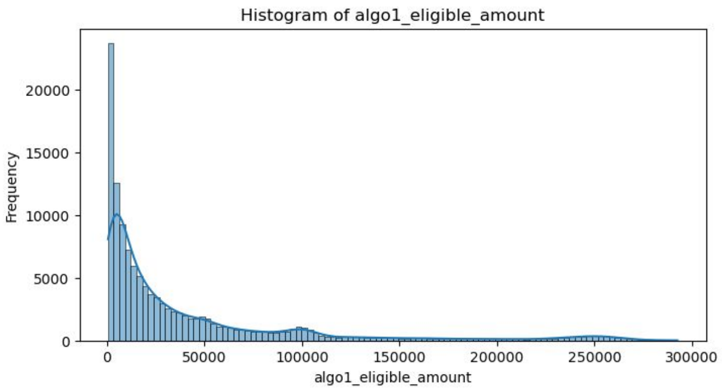
Modeling:
- Testing several algorithms
- Tuning the best performing algorithm

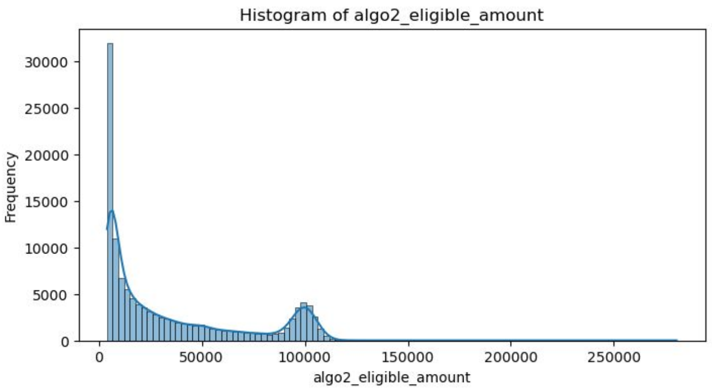
Experimentation loops:
-Experiment differents approaches and learn

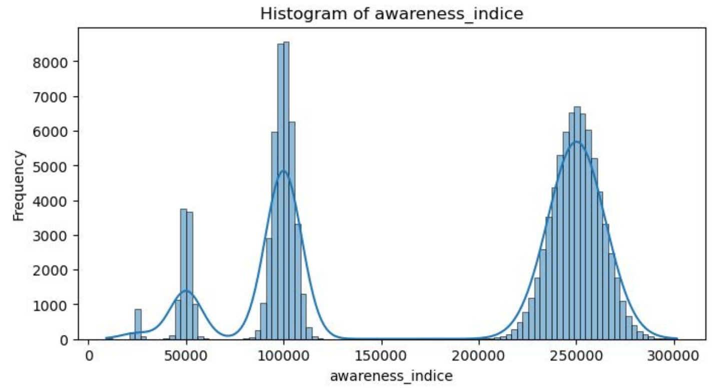
03

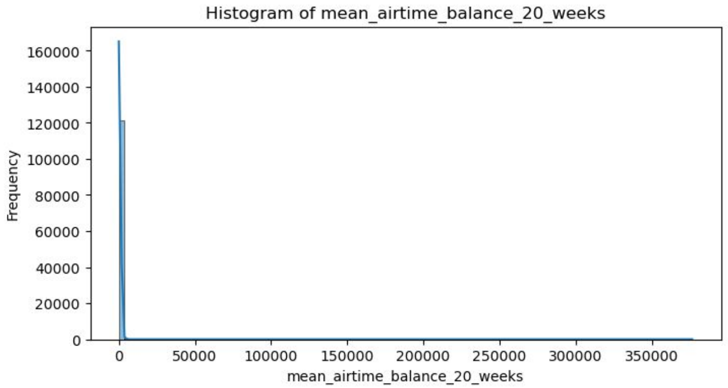
Exploratory Data Analysis

EDA - distributions



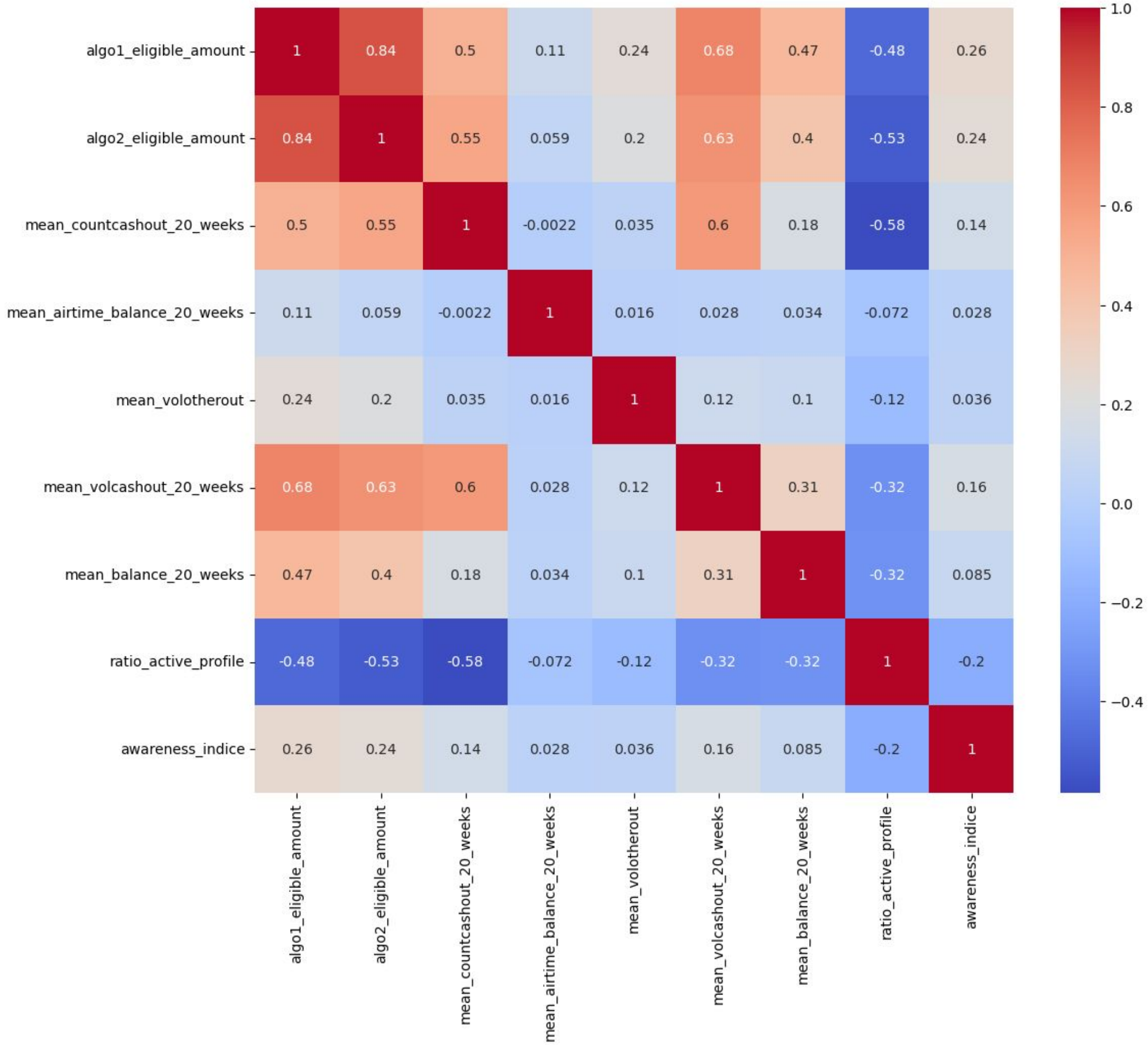






- Several features are right skewed => there may be a significant number of outliers towards the higher end of the range.
- Regarding the awareness_indice feature, we can observe that it has multiple peaks centered around 50,000, 100,000 and 250,000. This suggests that there may be different groups of customers with different levels of awareness of the offer.

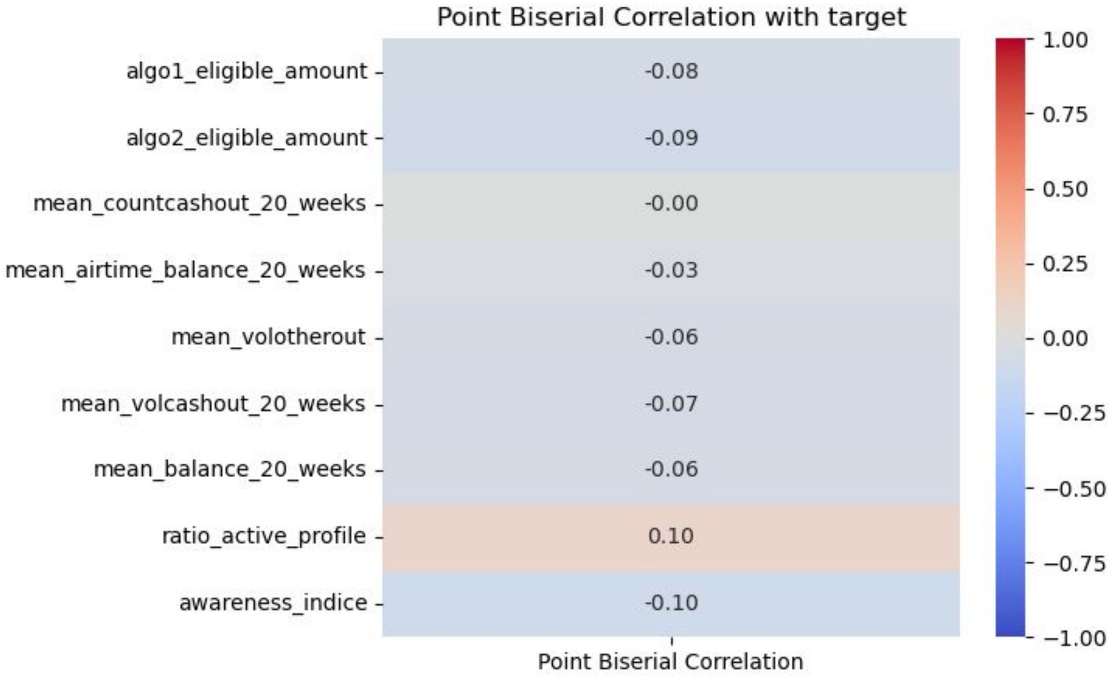
EDA - correlations between features



	algo1_eligible_amount	algo2_eligible_amount	mean_countcashout_20_weeks	mean_airtime_balance_20_weeks	mean_volotherout	mean_volcashout_20_weeks	mean_balance_20_weeks	ratio_active_profile	awareness_indice
algo1_eligible_amount	1	0.84	0.5	0.11	0.24	0.68	0.47	-0.48	0.26
algo2_eligible_amount	0.84	1	0.55	0.059	0.2	0.63	0.4	-0.53	0.24
mean_countcashout_20_weeks	0.5	0.55	1	-0.0022	0.035	0.6	0.18	-0.58	0.14
mean_airtime_balance_20_weeks	0.11	0.059	-0.0022	1	0.016	0.028	0.034	-0.072	0.028
mean_volotherout	0.24	0.2	0.035	0.016	1	0.12	0.1	-0.12	0.036
mean_volcashout_20_weeks	0.68	0.63	0.6	0.028	0.12	1	0.31	-0.32	0.16
mean_balance_20_weeks	0.47	0.4	0.18	0.034	0.1	0.31	1	-0.32	0.085
ratio_active_profile	-0.48	-0.53	-0.58	-0.072	-0.12	-0.32	-0.32	1	-0.2
awareness_indice	0.26	0.24	0.14	0.028	0.036	0.16	0.085	-0.2	1

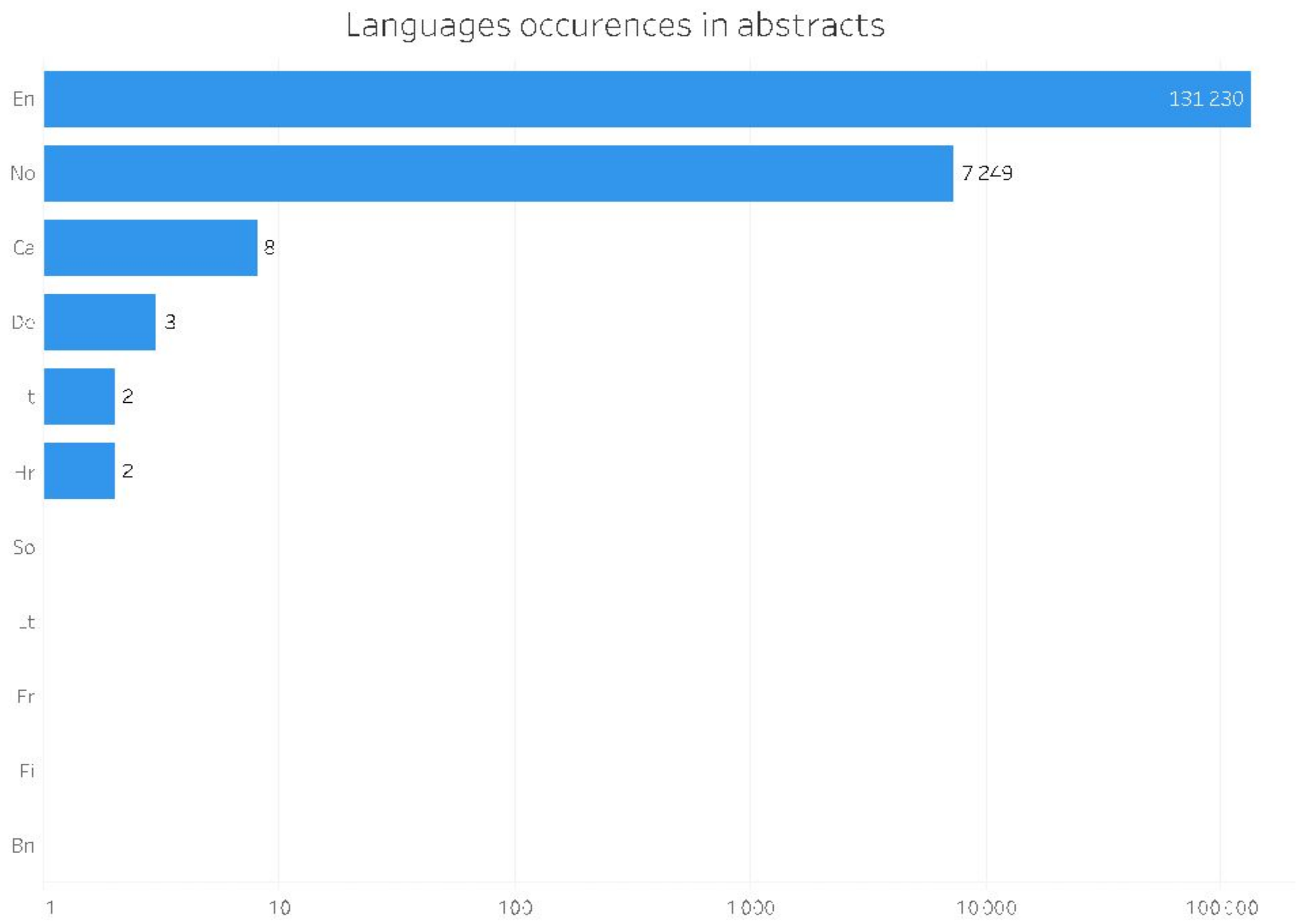
- algo1_eligible_amount and algo2_eligible_amount have a strong positive correlation of 0.84. This is indicating that the two algorithms generally agree on the eligible amount for a customer.
- mean_countcashout_20_weeks, mean_volcashout_20_weeks, and mean_balance_20_weeks are positively correlated with both algo1_eligible_amount and algo2_eligible_amount. This suggests that customers with higher average cashout transactions, cashout volumes, and balances in the last 20 weeks are likely to be eligible for larger loan amounts.
- ratio_active_profile has a negative correlation with both algo1_eligible_amount (-0.47) and algo2_eligible_amount (-0.53). This suggests that customers with more volatile transaction patterns may be considered less eligible for larger loan amounts.
- awareness_indice has a weak positive correlation with both algo1_eligible_amount (0.26) and algo2_eligible_amount (0.24). This indicates that customers with a better understanding of the offer might be slightly more eligible for larger loan amounts, but this relationship is not very strong.

EDA - correlation between features and target



- The majority of the features have a weak correlation with the target variable (default_60days).
- The low correlation values suggest that the individual features alone may not be sufficient for accurate prediction, and it is likely that a combination of features will be necessary.
- This can also indicate that the relationship between the features and the target is non-linear, thus sophisticated models that can capture such relationships may be needed for better performance.

Data exploration and processing



- 11 detected languages in the abstracts. English is the most represented
- Language-specific processing technics

04

Features engineering

Feature engineering

Dates

10

- request_year
- request_month
- request_day
- request_hour
- request_dayofweek
- reimbursement_year
- reimbursement_month
- reimbursement_day
- reimbursement_dayofweek
- date_diff

Amounts

8

- mean_cashout_to_balance_ratio
- mean_cashout_to_airtime_ratio
- mean_volotherout_to_balance_ratio
- algo_diff
- total_eligible_amount
- min_algo_eligible_amount
- max_algo_eligible_amount
- mean_algo_eligible_amount

Ids

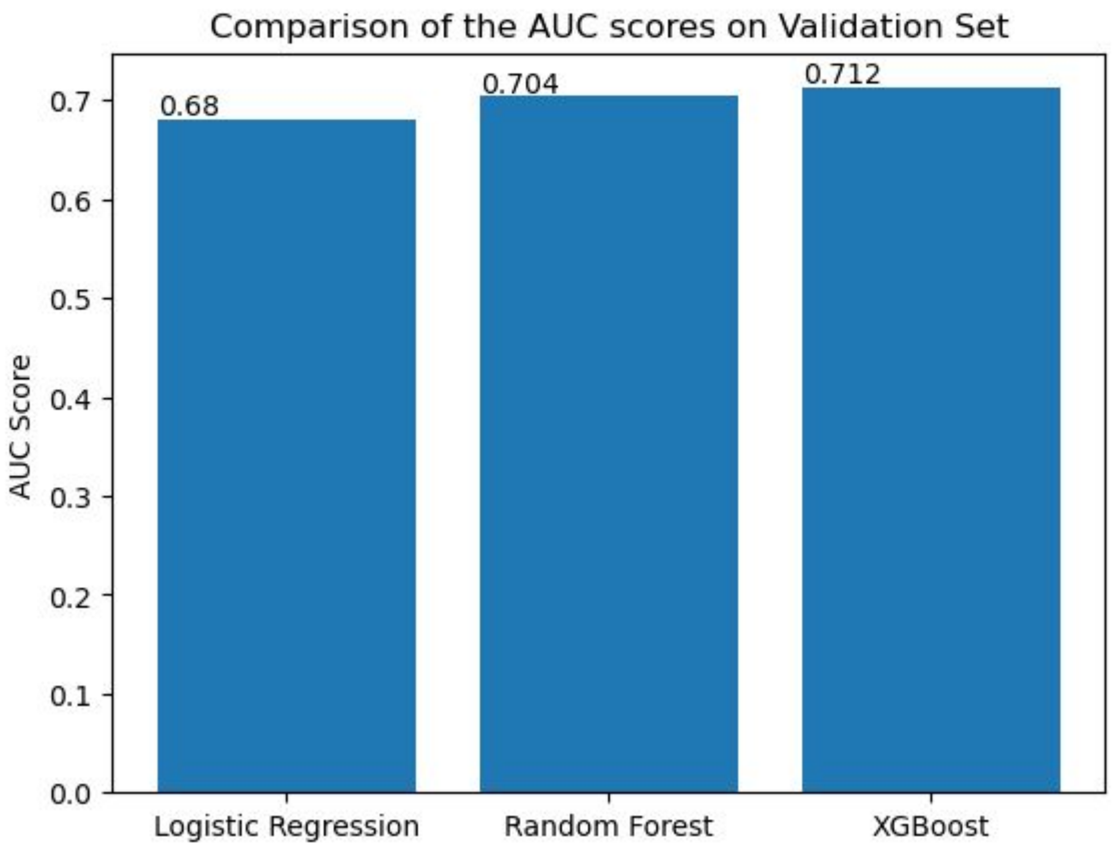
1

- customer_id X
- simulation_id X
- loan_id X

05

Modeling

Trying differents models



06

Tuning

Tuning XGBoost

Validation set

```
Best parameters: {'learning_rate': 0.1, 'max_depth': 6, 'min_child_weight': 3, 'n_estimators': 100}
```

Classification Report:

	precision	recall	f1-score	support
0.0	0.71	0.81	0.76	7474
1.0	0.60	0.47	0.53	4669
accuracy			0.68	12143
macro avg	0.66	0.64	0.64	12143
weighted avg	0.67	0.68	0.67	12143

AUC Score:
0.7201627883638364

Test set

Classification Report:

	precision	recall	f1-score	support
0.0	0.70	0.81	0.75	7357
1.0	0.61	0.48	0.54	4787
accuracy			0.68	12144
macro avg	0.66	0.64	0.64	12144
weighted avg	0.67	0.68	0.67	12144

AUC Score:
0.7214909018435736

Handling class imbalance with SMOTE

Classification Report:

```
              precision    recall  f1-score   support

         0.0       0.73      0.73      0.73      7357
         1.0       0.58      0.58      0.58      4787

    accuracy                           0.67     12144
   macro avg       0.65      0.65      0.65     12144
weighted avg       0.67      0.67      0.67     12144
```

AUC Score:

0.7168164543550068

07

To go further

To go further

- Retrain the XGBoost model using only the top 10 most important features identified from feature importance analysis. This may help to simplify the model and reduce overfitting.
- Explore the use of SMOTE (Synthetic Minority Over-sampling Technique) as a hyperparameter in the XGBoost tuning. This technique may help to balance the class distribution and improve the model's ability to predict defaults.
- Gather additional information about the loan amount and use it as a feature in the model. This could potentially improve the model's performance by capturing the relationship between loan amount and default risk.