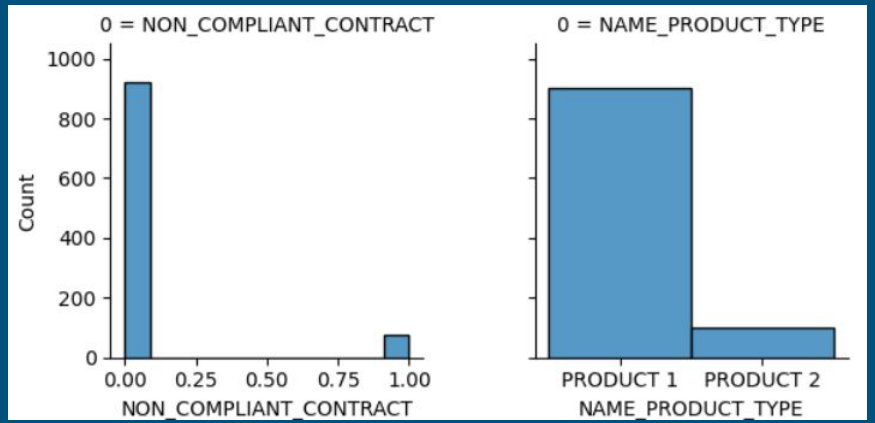
Challenge
BBVA AI Factory

Ignacio Vellido Expósito

Data Analysis

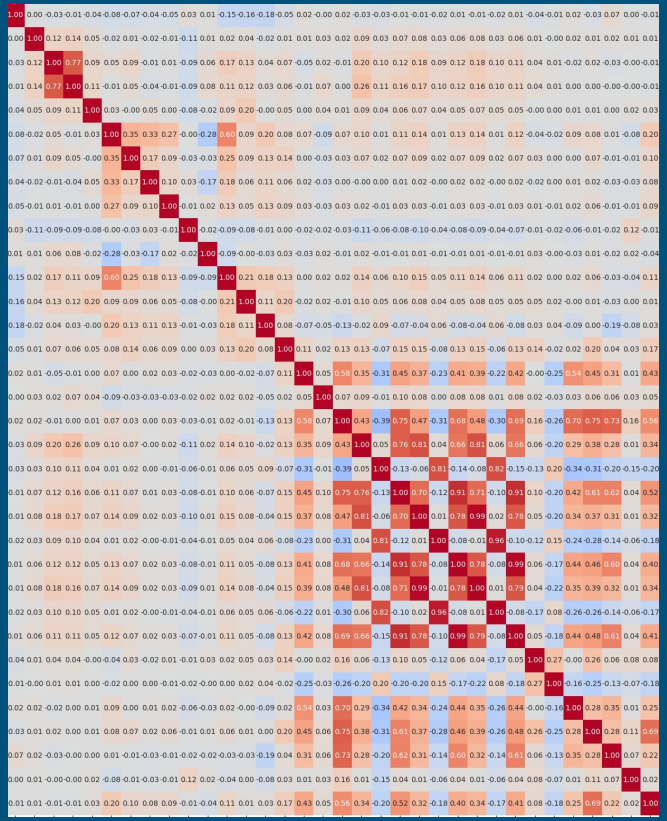
- CLIENTS
  - Imbalanced target class: ~90% of non-compliant
    - Also on categorical features
  - Numerical variables equally distributed across both classes
  - Lack of normality

- BEHAVIOURAL
  - ⅓ of the clients
  - Time indexed



Data Analysis

- Anomaly detection
  - Could be more than one dimensional
  - 1% identified with a KNN-based model
- Correlation
  - Low correlation with target, but high in measure variables (e.g. LOAN_ANNUITY_MIN/MAX/SUM)
- Missing data
  - High for some features
    - CAR_AGE 66%
    - REACTIVE_SCORING 56%



Q1: Can you design a code that detects those clients who are **not** going to fulfill the contractual conditions?

Could help the marketing experts design the best strategy for the products.

Opens the door for some other interesting analyses:

- What client segments are more likely to fulfil the contract?
- What is our target market? How big it is?
- What makes a client valid? Could some other product be defined for those who not?

Preprocessing

- Outliers
  - Removed, but should be analyzed
- Feature selection
  - High dimensional space with low data → Multiple functions that fit the data
- Normalization
  - Not normally distributed → min-max more appropriate
- Missing values imputation
  - Using simple statistical based methods. Models could be trained
- Treatment of class imbalance
  - Using class weight properties of the models
- Data augmentation
  - Appropriate, but not carried out

Training

Logistic Regression

- Simple model
- Interpretable
- With regularization for feature selection

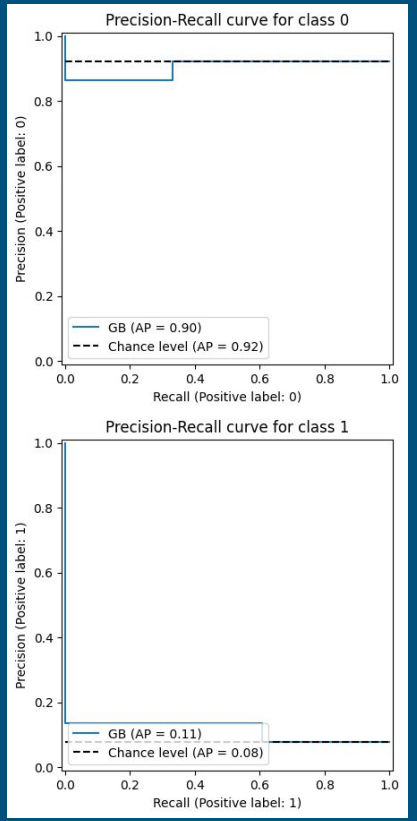
Gradient boosting

- Complex model
- Somehow interpretable using other techniques

Evaluation

- With imbalance some metrics could be misleading
  - Global metrics (F1, accuracy…)
  - AUC
- In-class precision-recall more suitable

- Same results for both models. Something could be wrong in the process
  - No better than baseline for non-compliant



How can you prove that any model that you develop is not capturing noise?

- Metrics: Good F1-scores (classification) or MSE/MAE (regression). Precision/recall over the desired thresholds
- Is the model generalizing?
- How outliers are affecting the model?
- Sensitivity analysis
- How are the predictions distributed? Is there any bias?
- If time is a variable, are the estimates (or the errors) stationary?

If you were auditing your own model from an outsider perspective, how would you determine whether or not the model is robust?

- Model Foundations
  - Is the best model for the problem?
  - Are assumptions fulfilled?
- Data
  - Do they encode the necessary information?
  - Preprocessing
- Performance
  - Backtesting. Has been data or concept drift?
  - Global/local/segment biases?
  - Counterfactuals
- IT Infrastructure
  - Documented, defined?
  - Dependencias identified and managed?

Q2: We think that it is viable to create a salary estimator for customers for whom we do not have demographic and financial data. Is this possible?

Accurately estimating the salary would increase the number of potential clients, but:

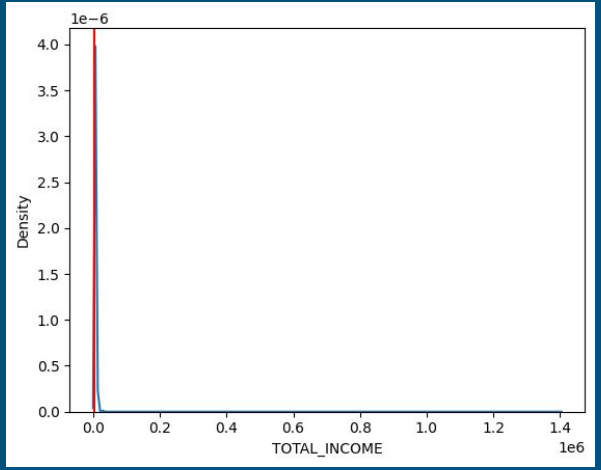
- Many variables missing. Although the BEHAVIOURAL dataset could be useful (e.g. trying to make an estimate from balance and drawings).

- Task prone to bias. We should think if we want to represent the data as it is or fix any possible bias in it.

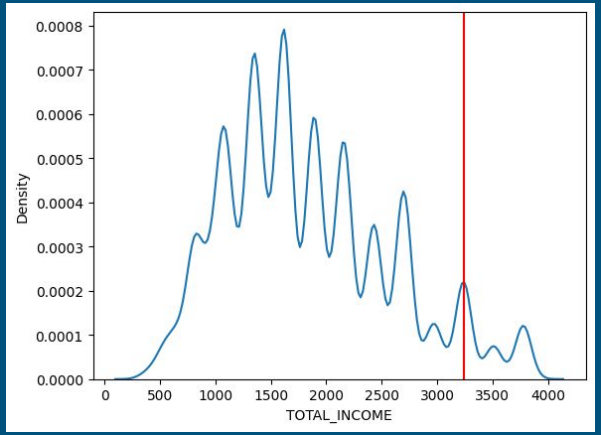
Preprocessing

- Only numerical features available
- Median imputation for missing values
- Min-Max scaling to [0,1] range
- Correlated variables kept
- Outliers for the target variable INCOME removed

From this



To this

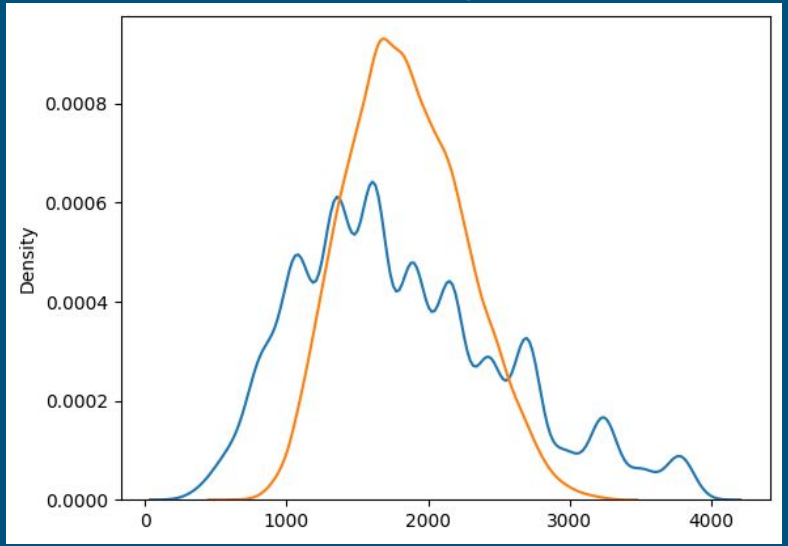


Model - Random Forest

- Most relevant variables
  - INSTALLMENT
  - LOAN_ANNUITY_PAYMENT_MAX
  - PROACTIVE_SCORING

- Results
  - Out-of-bag R^2
    - 0.2 (0 = Average, 1 = Best)
  - Mean Absolute Error
    - 525.80€ (High for a monthly income)

Distribution of test predictions



Q3: Can you propose new data, procedures or methodologies to apply with the idea of giving continuity to our project?

Data

- Account/card balance evolution through time
- Mortgage related data
- Customer behaviour on the app/web to measure effects of marketing
- Identify patterns in customer's operations to ensure they qualify for the contract requirements

Models

- Segmentation, looking to normality with VAE and/or adversarial networks
- RLHF to improve client preferences
- Ensemble and DL models for more complex but least interpretable models
- Learning relationships between clients with transactional data and graph theory (GNN)
- Evolution of customer profiles over time and the impact of economic cycles or other country-level actions on them

Challenge
BBVA AI Factory

Ignacio Vellido Expósito