UNIQA Data Science Challenge 2022

Tadeusz Dziarmaga
tadeusz.dziarmaga@student.uj.edu.pl +48 538 214 535 Poland
Jacek Kowalczyk
jk417691@students.mimuw.edu.pl +48 514 578 007 Poland

30th March 2022

1 Introduction

Building a XAI class (explainable AI) models differs from standard Machine Learning tasks. Data Scientist not only has to deliver the best performing model, but also has to develop ability to interpret and alter its behaviour in a human understandable way. Out of plethora of established, well researched Machine Learning models, only subgroup of them are easily readable by human. That makes most of them unusable in a highly regulated field like banking or insurance, where every business decision has to be interpretable. Thus, well established GLM/GAMM models are still used in the trade.
To tackle the UNIQA Data Science Challenge 2022, we suggest a different approach to building GLM model. We use a standard logistic model, but feed train it on data, which is partly generated by SMOTE, an algorithm based on K Means. This way we reduce the inflation of 0's standard problem in the industry and achieve better model's performance, especially for the riskier clients.

2 Data preparation

Our work was based on a training data set provided by UNIQA. Working on a data of almost 4 million observation and 150 unique columns required some preparation, before a proper model could be built and verified.

2.1 Cleaning Data

We begin by standard methods of making data set more readable. We remove artifacts from data set (1, 000) and change them to standard Numpy NANs. Then we change the decimal separators from commas to points, so the REAL columns are readable to Pandas as float.

2.2 Dealing with Missing Values

The hypothesis we approached the given data set was: a lack of information is also an information. Thus, we were reserved in a careless and mechanical removal of columns and rows, which were only partially filled. We've decided to drop rows, that had more than 50% of missing values, but not to drop any columns, but try to fill them by mean/median (INT and REAL) or dominant (NOM). In one case (CAR 10 NOM) we've found out, that making NAN as a separate nominal category was very informative.

2.3 Categorical Variables and Dummies

As the dataset provided by UNIQA was lacking any description, we assumed every nominal variable was categorical. We took standard way of using dummy, binary variables for every category. We included NAN as a category in a single columns (CAR 10 NOM).

2.4 Scaling Data

We scale all INT and REAL columns to mean 0 and standard deviation 1, so the gradient descent converges fast enough. This changes the interpretation of logit's parameters.

3 Method

3.1 Dealing with 0 Inflation K means algorithm and SMOTE

The challenge of working with imbalanced datasets is that most machine learning techniques will have poor performance on the minority class. To address that problem, we used oversample technique called SMOTE (Synthetic Minority Oversampling Technique). This technique is an improvement on just duplicating examples from the minority class. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (we use k=5 as by default). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space. We oversample the dataset to make it balanced (50/50).

3.2 Logistic Model

A logistic regression classifier was implemented using the library sklearn. The solver used to optimize the model and thereby minimize the loss in the cost function was ,,liblinear". Regularization, as implemented by default in the model, was used to penalize large class weights and thereby prevent overfitting. The maximum amount of iterations was increased from the default amount of 100 to 2500 in order for the optimization to converge.

4 Verification, Interpretation and Important Features

After the model had been trained and verified, its feature weights were extracted and sorted by their absolute values. Positive coefficients indicate a decrease in creditworthiness and negative indicate an increase, why they were sorted by

their absolute values. This would explain what features had the most impact on the decisions made by the model.

4.1 Somer's D and Gini Coefficient

As the dependent variable is 0 inflated, we use a standard way of verifying our predictions on test set Somer's D order statistic, known in the trade as Gini Coefficient. We use our own implementation, but crosscheck it with built in Somer's D function.

4.2 Interpreting the Coefficients

The feature weights extracted from the trained logistic regression classifier are global explanations as they are computed with respect to the whole dataset and are describing a measure of how much an increase of a feature would change a prediction. A positive feature weight indicates that an increase of the feature value will increase the probability that the applicant is bad. In contrast, a negative feature weight indicates the opposite.

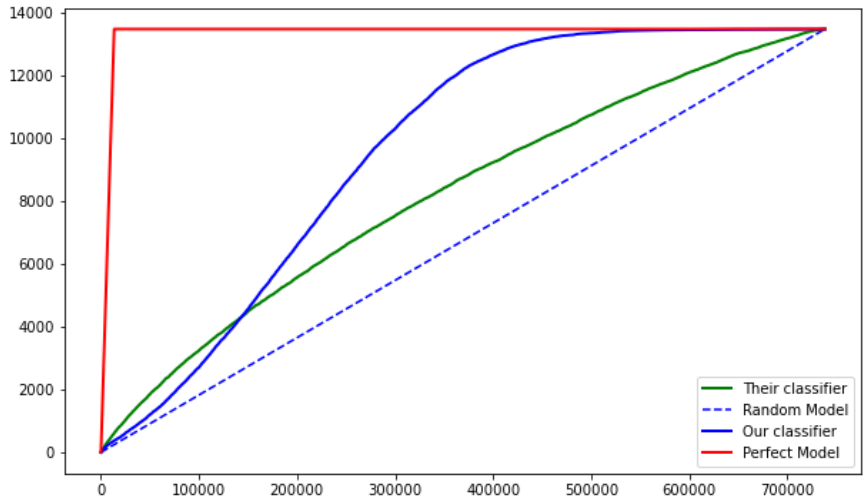

Figure 1: Performance of our model vs CURRENT MODEL REAL

4.3 Important Features

Table 1: Logistic regression coefficients:

Feature	Feature coefficient
CAR 11 NOM 6	-1.889
PERSON 1 NOM 4	-1.879
INS HISTORY 31 NOM 16	-1.688
OTHER 3 NOM 3	-1.681
PERSON 10 NOM 3	-1.378
CAR 11 NOM 4	-1.338
INS HISTORY 31 NOM 1	-1.275
INS HISTORY 31 NOM 8	-1.167
CAR 14 NOM 9	-1.117
OTHER 4 NOM 9	-1.110
INS HISTORY 31 NOM 0	-1.093
INS HISTORY 31 NOM 12	-1.078
INS HISTORY 31 NOM 14	-1.074
INS HISTORY 31 NOM 17	-1.024
CAR 11 NOM 3	-1.002
INS HISTORY 31 NOM 6	-0.997
INS HISTORY 31 NOM 13	-0.996
INS HISTORY 31 NOM 5	-0.988
CAR 14 NOM 6	-0.985
CAR 14 NOM 5	-0.976
CAR 14 NOM 21	-0.966
INS HISTORY 31 NOM 9	-0.964
OTHER 4 NOM 6	0.753
CAR 8 NOM 2	0.536
PERSON 9 NOM 3	0.328
CURRENT MODEL REAL	0.269

Last four features are the highest ranked positive features.

5 Summary and Suggestion for Model's Improvement

We've found out, that given the data it is much easier to recognise features of non-risky clients, than those risky. It can be seen in our model's performance compared to current UNIQA's model. It can be seen, that SMOTE improves performance of the model, compared to the benchmark. We suggest that best features of our our model to be incorporated into future modeling. Link to results of our model: Google Drive