ETHICS FOR ARTIFICIAL INTELLIGENCE

- Analyse the dataset
- Feature importance
- Scoring the test set
- Analyse some test records
- Remove the sensitive variables
- Understanding the problem

Group 3 | Olimpia Sannucci | Cosimo Poccianti | Miro Confalone | Hanna Carucci Viterbi | Carlo Ardito

#1 Focus on variables that you consider prone to ethical discussions

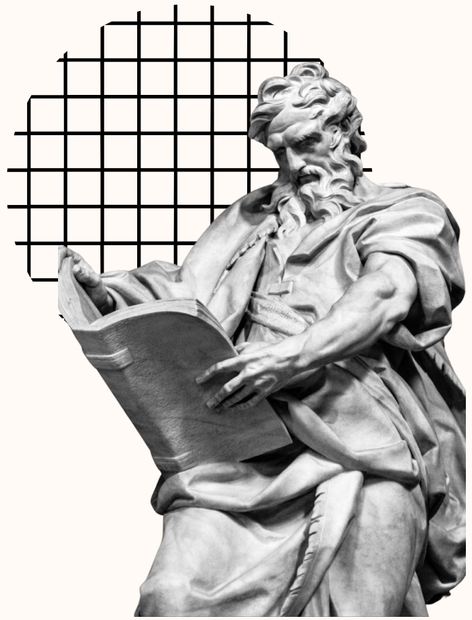
As a team, we have decided to focus on variables that could be unnecessary for an economic evaluation, too sensitive and so may lead to biases.

- CODE_GENDER
- CNT_CHILDREN
- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- Housing variables
- REGION_POPULATION_RELATIVE
- DAYS_BIRTH
- DAYS_EMPLOYED
- OWN_CAR_AGE
- OCCUPATION_TYPE
- ORGANIZATION_TYPE_CITY
- DEF_60_CNT_SOCIAL_CIRCLE
- EXT_SOURCE 1,2,3
- AGE

#2 In the training set are the interesting variables at #1 related to the target variable?

Of the 61.501 clients with different education types the ones with payments difficulties in % are:

- Secondary special: 8.93%
- Higher education: 5.35%
- Incomplete higher education: 8.48%
- Lower secondary: 10.92%
- Academic degree: 1.82%

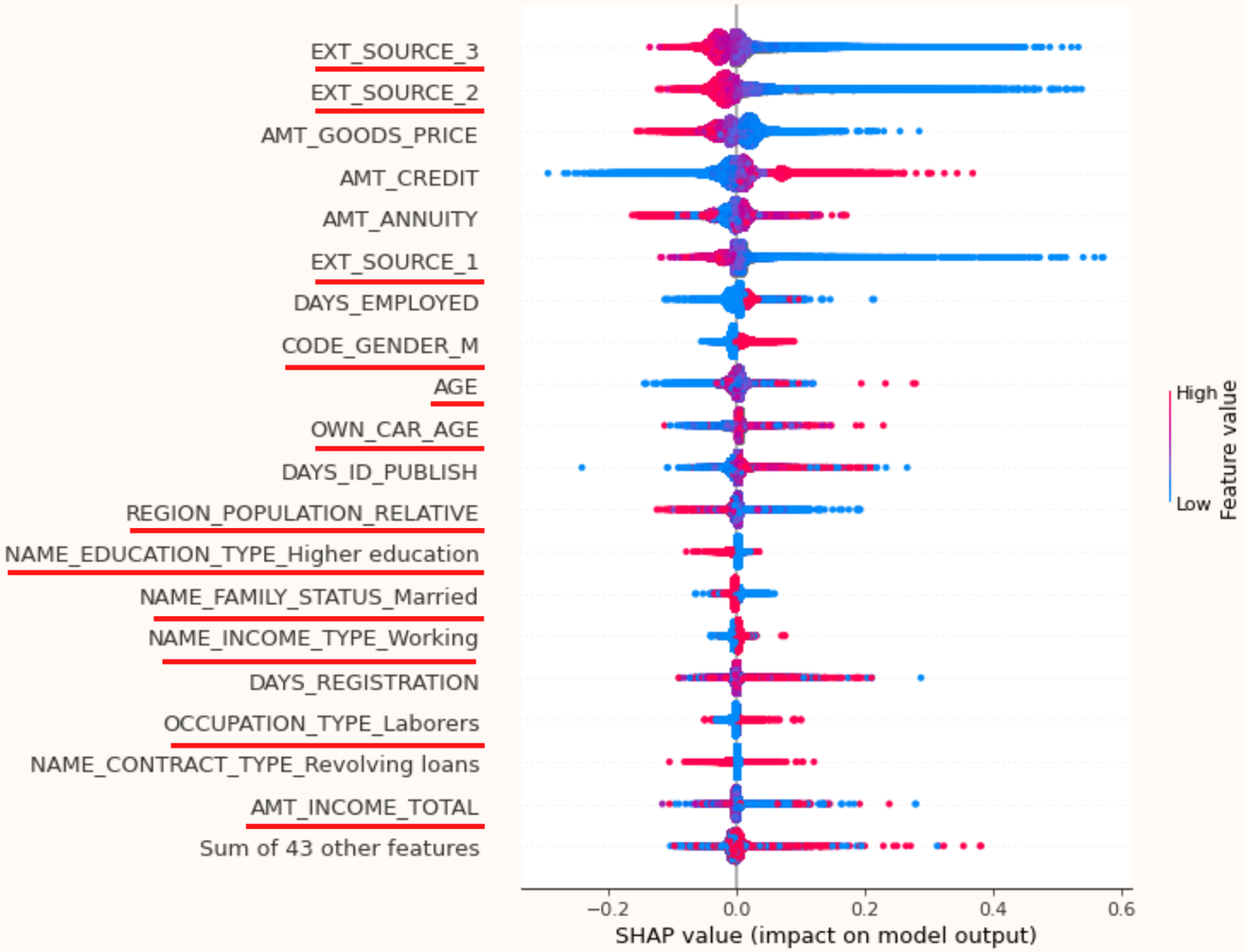


Analyse the dataset

Of the 40.561 females and the 20.940 males the ones with payments difficulties in % are:

- Females 22.55 %
- Male 37.67 %

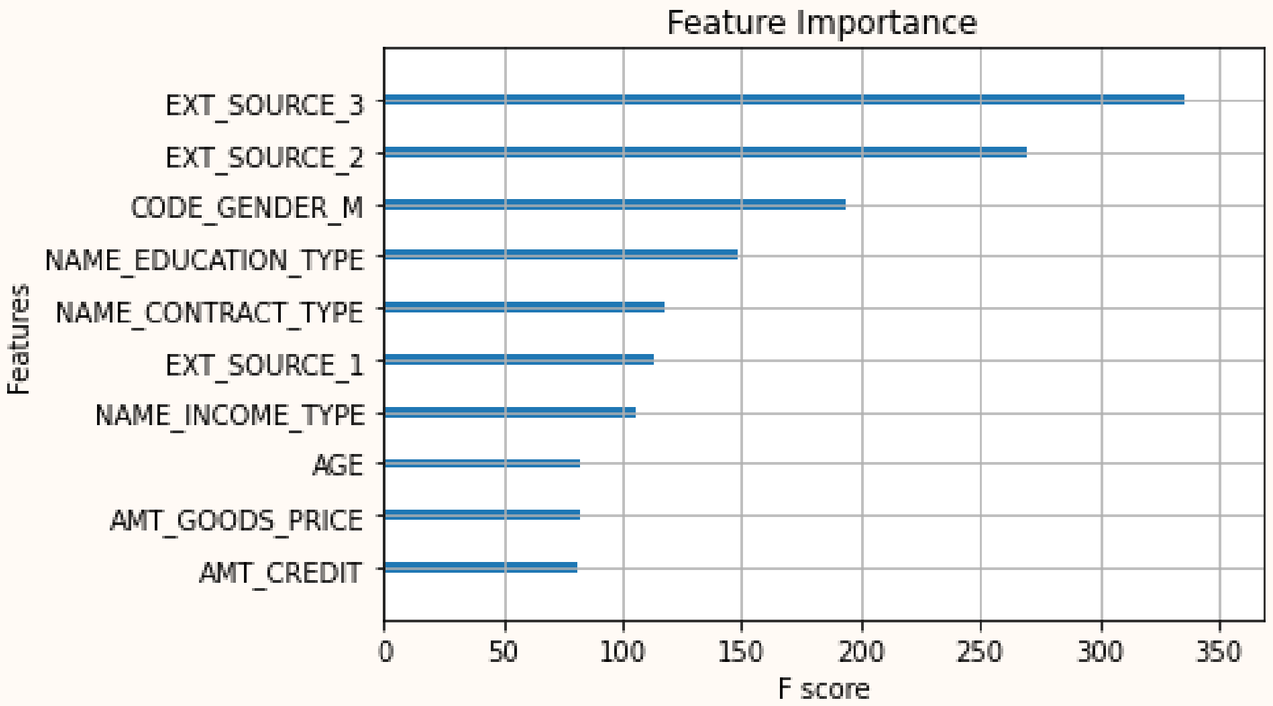
#3 Are the features you found in #1 considered relevant by the trained ML model?



Feature importance

The SHAP analysis indicates the combined relationships among variables that create the model's output. It is notable how the variables underlined in red contribute to the model; those are the ones we selected because they are prone to ethical discussion.

#3 Are the features you found in #1 considered relevant by the trained ML model?



Feature importance

As shown in the feature importance plot, EXT_SOURCE 1,2 and 3 are more relevant to the model, but we are unaware of what they represent. Could this model be defined efficient if is based on unknown variables? Could they bring more biases within the model?

#4 Ignore the original test target and just take the predictions of the ML model

After applying the model, the relationship between some variables and the test target significantly changes. Some predictors outstand significantly:

- GENDER: running the model shows that female loaners have lower payment difficulties.
- CNT_CHILDREN: The number of children it's a factor that influences the risk score; the more the number of children in a family, the higher the payment difficulties (the number of expenses in a family could be a liability)
- NAME_FAMILY STATUS: considering the family status, an interesting result is that the categories with less payment difficulties are married and widows.
- EDUCATION: having higher education decreases payment difficulties. Having higher degrees creates a sense of credibility.
- NAME_HOUSING_TYPE: the model considers the loaner's living situation (almost 50% of those who live with parents have a higher risk of not paying the loan back)
- OCCUPATION_TYPE: the loaner's occupation is another variable that the model examines; accountants and managers have a significantly lower risk percentage of not paying the loan compared to low-skill labourers

Scoring the test set

#5 Try and manually alter the sensitive variables (those at #1) and score the new, altered records: do the predictions change?

Features on the train set have been altered by shuffling categorical and numerical variables and increasing their value by 20/30%
By doing so, the AUC of the model (0.74) didn't change, but it did influence the predictions, which were slightly modified compared to the original ones. Therefore, ethical variables impact the prediction process and altering them makes it possible to perceive a change also on the risk score.

#6 You've analysed the overall behaviour of your ML model on the test set: what's your conclusion? Did the model learn the differences and biases?

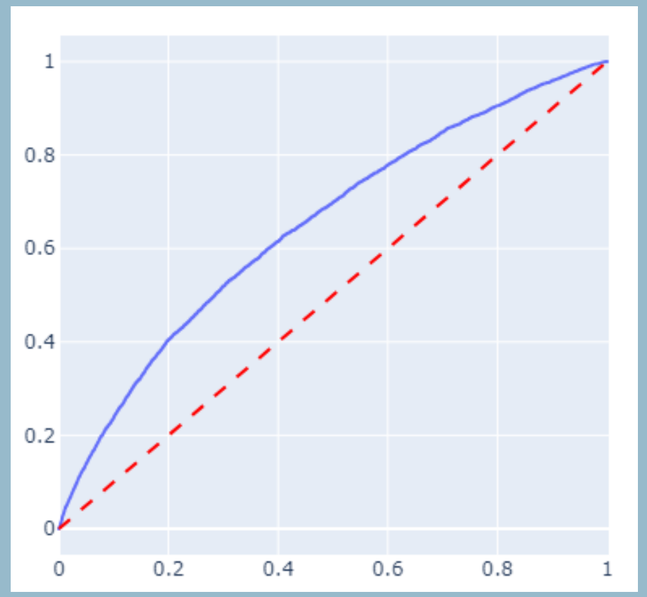
The model did not learn the differences and biases in the original dataset. This means the ethical variables have an influence on the risk score and, therefore, the model. Credit scores squeeze a range of socio-economic data, such as employment history, financial records, and purchasing habits, into a single number. Other than being critical factors in loan applications, credit scores are used to make many life-changing decisions, including insurance, hiring, and housing. The risk is that algorithms may create biases based on the fed data; thus, keeping the data objects inclusive is essential in building a non-biased ML model.

Analyse some test records

#7 Retrain the ML model, score the test set and perform the analysis (#3). What's the new AUC performance (with respect to the original one)?

By removing all the ethical variables, the new AUC performance of the model is 64%, much lower with respect to the original at 75%.

ROC Curve

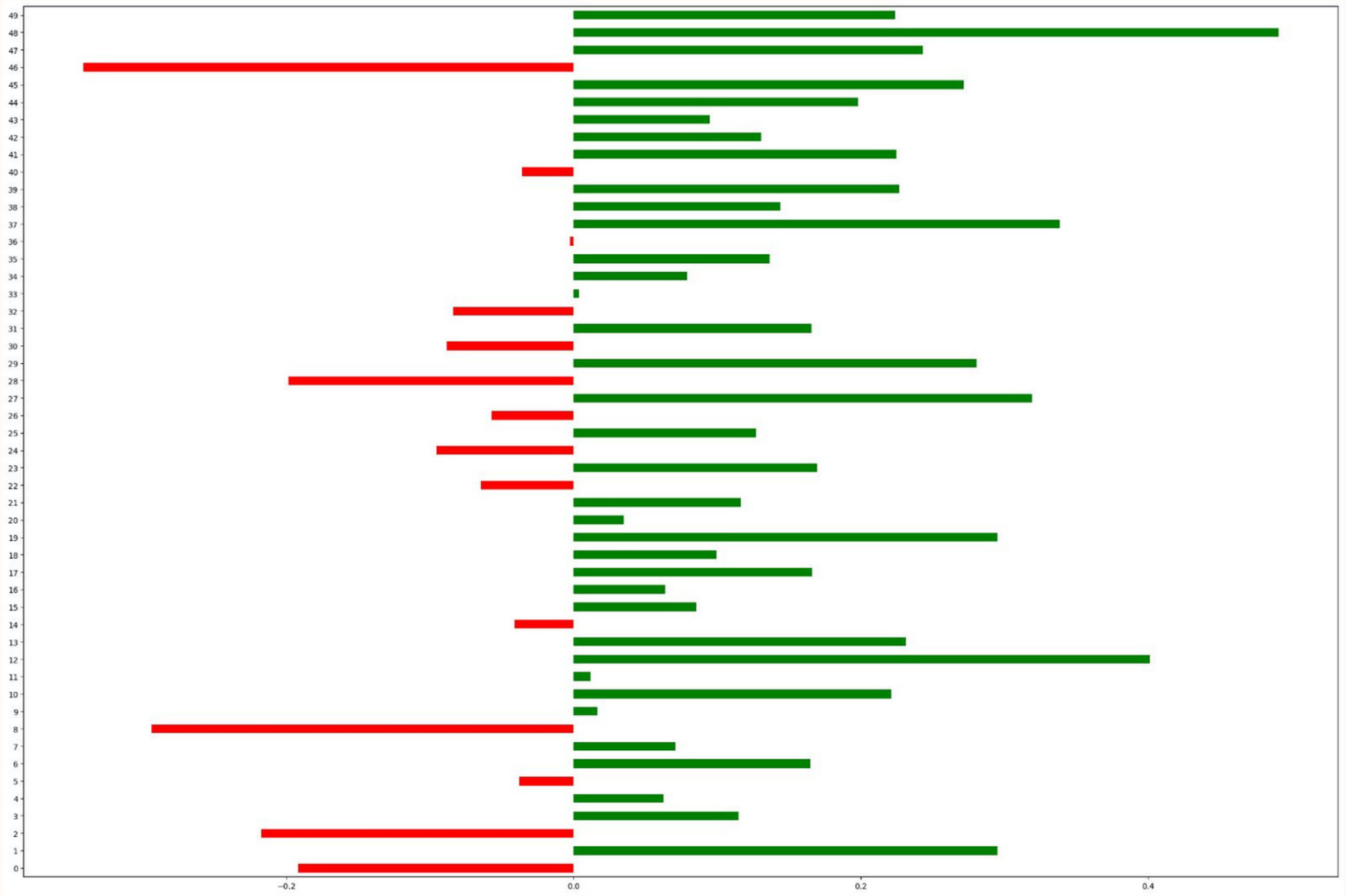


The ROC Curve shows the relationship between true-positive and false-positive. The closer the blue line is to the red one, it means the false-positives increase.
The result is a worst classification given by the algorithm.

Remove the sensitive variables

#8 Analyse both the overall behaviour and individual one (#4 to #6). What's going on? Do we still see differences in the average prediction of different groups?

We saw changes in the model and its predictions by removing the sensitive variables. Considering the average predictions, omitting from the model the variable gender on average increases the risk score for female clients. If the model knows the client's age, the risk score decreases as the latter increases. Looking at the single cases, some changes in their risk score have been observed. General variations of the predictions can be observed in the plot.



Remove the sensitive variables

#8 Analyse both the overall behaviour and individual one (#4 to #6). What's going on? Do we still see differences in the average prediction of different groups?

Original Results	**0.295304**	**0.92504**	**0.462637**	0.189093	0.617748
Manually Altered Results	0.6352069	0.6166435	0.5886503	0.2598523	0.38365341
Without Ethical Variables	0.6352069	0.6166435	0.5886503	0.2598523	0.3836534

The results show that ethical variables have a weight in the model in determining the client's risk score. Their removal or alteration causes a change in the latter. By observing alterations in the predictions, we can deduce that the model still considers the ethical variables as relevant.

Remove the sensitive variables

#9 Did results at #8 surprise you? Can you explain why just removing the variables wasn't enough?

Removing the variables was not enough because:

- The model's accuracy has decreased, with an increase of the false positives.
- Since the model is trained on historical data, the biases may be inherent in other variables. More than removing the sensitive ones may be required for removing bias because, in any case, an individual with target 1 remains with that target even after the removal of sensitive variables, and it is precisely on that target that the training of the model is based.
- Data Collection: The process is designed and constructed by human beings, who decide which variables and why to collect them. To eliminate bias, one would have to restructure the whole process and thus act at the base of data collection: collecting less, more, or different types of data.
- Vicious circles: biases lead banks to distrust certain groups of people, particularly minorities, leading to low credit scores for minorities, reinforcing the bias in the algorithms.
- Is the problem really in the data? They reflect inherent societal biases, which are then reflected in the algorithms through the people who deal with the data. Is it then helpful to focus on the algorithm when it is the society that is biased?
- Minority data are often less than those of majorities, thus leading to a lack of precision and inequality, not necessarily related to the present bias.

The results sometimes confirm the assumptions made but sometimes disprove them, as in the case of gender: one would expect that if the subject is female, their credit score will decrease, when in fact, it increases.

#10 Any ideas on different ways to reduce the bias in this specific problems?

Possible Solutions:

- Informing users a priori about ethical risks
- To be aware and to indicate to the machine, which variables are ethically problematic to have a sort of 'bias alert' when the score is significantly based on these ethically relevant variables, and thus require human intervention and not leave the decision on access to credit to the machine alone.
- The lack of spent accuracy comes from noise in the data, not so much from the model, so one has to consider whether it is important to focus on improving the model or on thoroughly cleaning the data.
- Constantly update the model with up-to-date and accurate data. Outdated data can distort the evolution of the society
- Ensure that ethical principles such as benevolence, non maleficience, justice, faireness and explainability are kept in mind.
- Rely on open-source softwares which are validated by third parties

Understanding the problem

#10 What are the difficulties and the tradeoffs we could encounter?

TRADE-OFF & DIFFICULTIES

- Accuracy-Bias Trade-off: Keeping variables considered ethically problematic in the model significantly improves its accuracy, particularly on certain variables such as level of education or gender.
- Accuracy-Privacy Tradeoff: The achievement of accuracy should not infringe on the individual's privacy concerning his or her data, but there is also the legitimate interest of the bank to secure itself.
- Basic limitation of the algorithms, they cannot see the difference between correlation and causation.
- AI methods may lack of explainability
- One can fall into a few statistical pitfalls that mislead even those who have expertise in the matter (let alone those who do not), especially on heartwarming issues.
- The supporting costs of implementation and consumed time to have a harmless performance

Understanding the problem

THANKS