A Machine Learning Model for Lapse Prediction in Life Insurance Contracts

M. Azzone

Joint work with E. Barucci, G. Giuffra, D. Marazzina.

Aim of the research

- We use the Random Forest methodology to predict the lapse decision of life insurance contracts by policyholders.

- We use global and local interpretability tools to investigate how the model works.
  - We show that non economic features (time passed from the incipit of the contract and the time to expiry, as well as the insurance company) play a significant effect in determining the lapse decision while economic/financial features (except the disposable income growth rate) play a limited effect.
  - The analysis shows that linear models, such as the logistic model, may not be adequate to capture the heterogeneity of financial decisions.

- Barucci, E., Colozza, T., Marazzina, D., & Rroji, E. (2020) *European Actuarial Journal*

The determinants of lapse rates in the Italian life insurance market.

- Azzone, M., Barucci, E., Giuffra, G., Marazzina, D. (*submitted*)

A Machine Learning Model for Lapse Prediction in Life Insurance Contracts

Life Insurance Contract – Our dataset

- The dataset considered in our analysis comes from one of the Italian largest insurance companies: it covers the 2008 - 2016 time interval with over one million life insurance contracts.
- Insurance contracts in the dataset belong to two broad families: 60% are standard contracts with a guarantee (non-negative minimum guarantee rate) and 40% are unit-linked contracts.

  - **Traditional**

  Traditional Insurance plans usually offer guaranteed maturity proceeds and they invest in low risk return options.

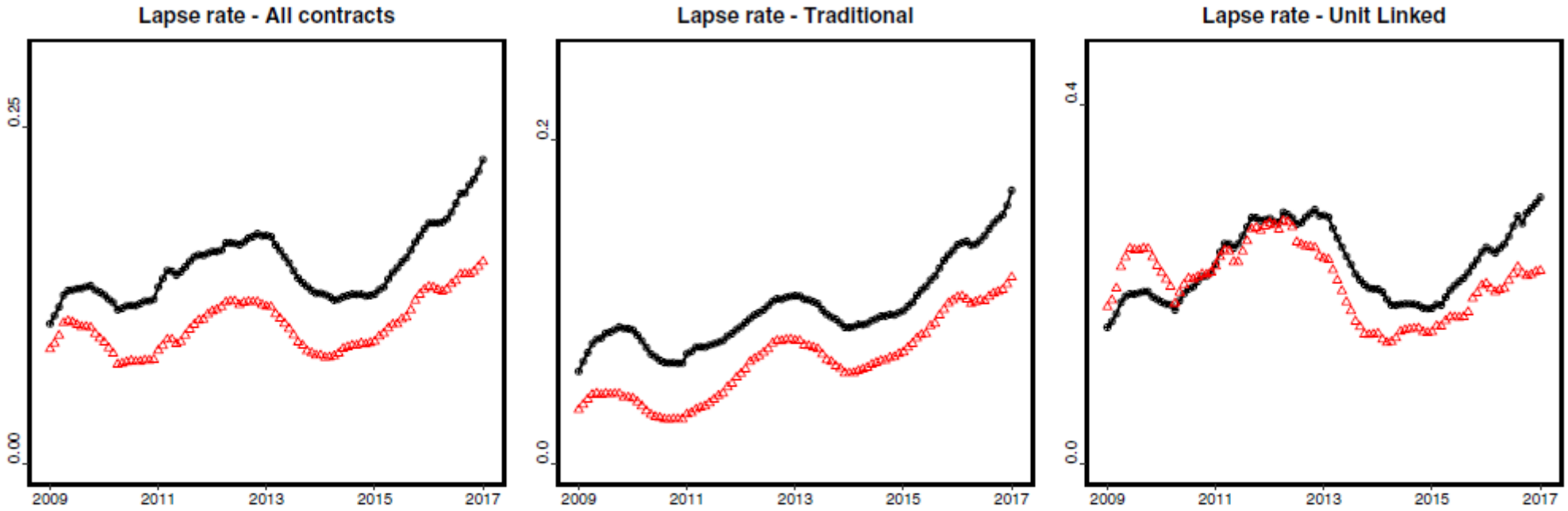
  Traditional plans do not have investment options.

  - **Unit Linked**

  Provides you with investment options based on your risk profile.

- The contracts refer to three different companies belonging to the same insurance group.
  - As the companies refer to different distributors, either bank assurance relationship or financial advisors, the lapse phenomenon is company-specific because the lapse decision may be driven by commercial policies of distributors

Early lapse



Lapse rates of all contracts, traditional contracts and unit-linked contracts at monthly frequency. At each month t, we report the lapse rate computed as the cumulated number of lapses in $[t-11, t]$ divided by the total number of outstanding contracts at $t-12$ (black solid line) and the lapse rate computed as the cumulated lapsed capital in $[t-11, t]$ divided by the total capital of the contracts outstanding at $t-12$ (red dashed ($\triangle$) line)

Dataset

- Gender: policyholder's gender (54% male and 46% female in the dataset).
- Age: age of the policyholder. We compute it as the policyholder's age at the beginning of the contract plus the time passed from the beginning of the contract.
- Region: one of the 4 macro-regions where the policyholder lives, North-West (Region=1), North-East (2), Center (3), and South & Islands (4).
- Company: the insurance company issuing the policy. The parent company develops its business through three companies. We identify them as company A (COMP=1), B (2), and C (3).
- Time from start: the time passed from the incipit of the contract.
- Time to expiry: the time left (in years) to the expiry date of the contract. We also have perpetual contracts with no expiration date (in this case we set the value of the variable to -1 in RF analysis while we consider a dummy variable in case of a logistic model to capture a perpetual contract).
- Contract size: the insured capital.

- Premium: policy premium. This variable is given by the contract size if there is only one payment and by the contract size divided by the number of payments if there are multiple payments.

- Product type: the dataset contains several different types of policies, we categorized them into two classes: traditional contracts (PT=0) and unit-linked (1).

We also consider macroeconomic variables that may drive the lapse decision.

- Disposable income: the yearly growth rate of Italian disposable income.
- Inflation: the yearly Italian inflation rate.
- Eurostoxx: the yearly growth rate of the Eurostoxx index.
- RiskFree: the Italian risk free rate.

For each contract, we have the following information: starting date, expiry date, and the eventual lapse date.

For each contract, we have an observation for each year in which the contract is active.

Our dataset contains 9 309 755 observations.

Research Question

- Linear models (like logistic regression) are good enough to explain the agents financial decisions (lapses)?

Machine Learning Approach: RF

Given the binary nature of the target variable, our RF classifier is based on a combination of classification trees and of tree bagging methods

- Classification trees are recursive algorithms that split the dataset into smaller sets (*nodes*) in a step by step fashion thanks to binary rules defined on the features of the observations.
- The complete dataset is the root node of the tree. To split the node, a feature is selected and a binary rule, e.g., if the value of the feature is larger or smaller than a given threshold, is defined on it in order to obtain two disjoint datasets. These two datasets become nodes of the tree.
- This procedure is repeated for each new node until a stopping criterion is met, e.g., the specified maximum depth of the tree is reached. So a classification tree is a growing tree with nodes refining the information about the exogenous variables to classify an item in the database (lapse or no lapse).

To calibrate the parameters of the RF we have to set a priori some hyper-parameters.

The hyper-parameters are chosen evaluating the performance of the calibrated models on the validation set.

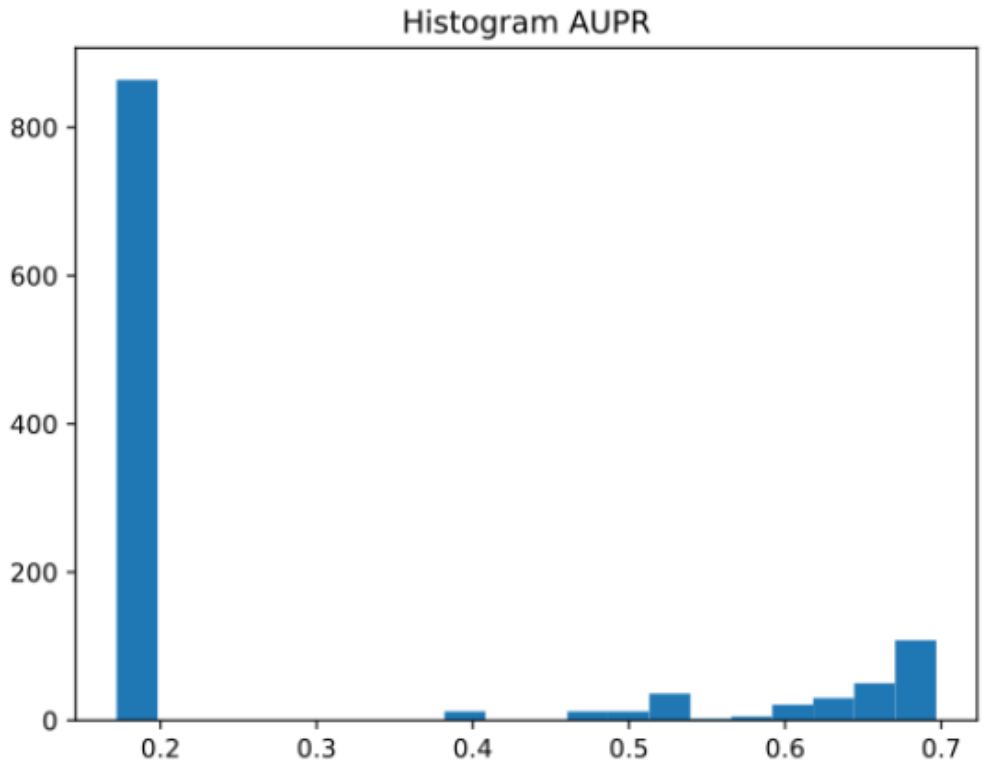
In our analysis we consider the following hyper-parameters:

- **maximum tree depth**: the maximum depth for a tree;
- **minimum leaf size**: the minimum number of observations contained in a leaf. A split will only be considered if the number of observations belonging to each child node will be higher than the threshold;
- **minimum decrease of impurity after a split**: a node is split if it generates an impurity decrease greater than or equal to the threshold;
- **minimum split size**: the minimum number of observations belonging to a node required to split it;
- **number of trees**: the number of trees in the forest.

We split our dataset as follows: training set (70%), validation set (15%), test set (15%).

In the case of the logistic model, as it does not present hyper-parameters, the validation and test set are combined to evaluate the out of sample performance of the model.

Maximum tree depth	5, 10, **50**, 100
Minimum leaf size	5, **10**, 50
Minimum decrease of impurity after a split	**0**, 0.1, 0.2, 0.3
Minimum split size	5, **10**, 50, 100
Number of trees	1, 2, 5, 10, **50**, 100



Model Selection

- Accuracy (threshold 0.5 & 0.75)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- ROC Curve and AUC

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}$$

The *Receiver Operating Characteristic* (ROC) curve is built plotting the false positive rate (*FPR*) against the true positive rate (*TPR*), moving the threshold from 0 to 1.
The ROC curve illustrates the predictive ability of a binary classifier, visualizing the trade-off between TPR and FPR, and thus suggesting an optimal threshold that minimizes the misclassification error.

- PR Curve and AUPR

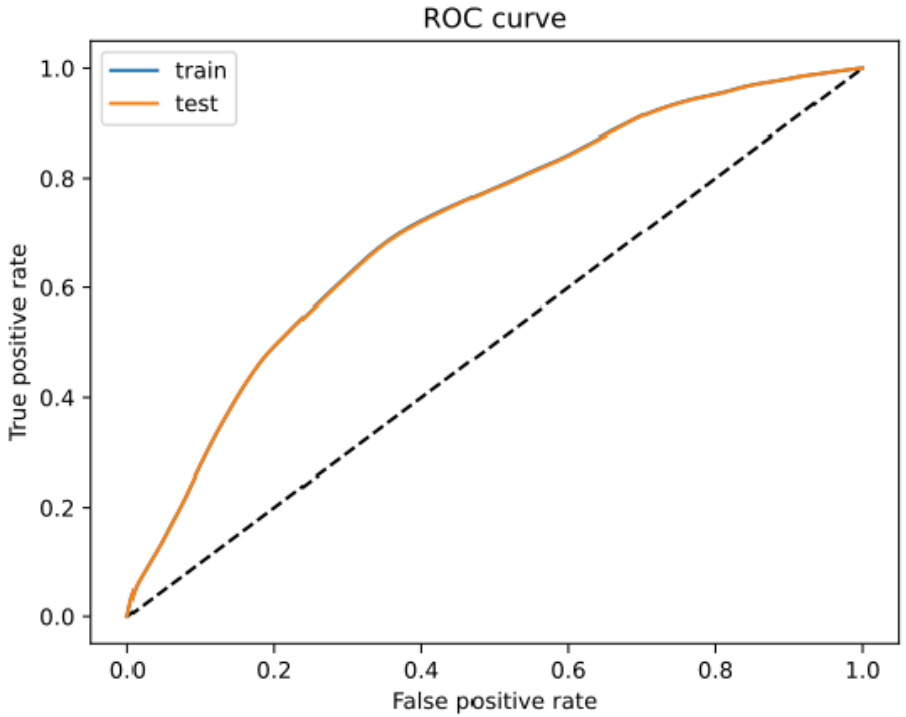
The *Precision Recall* (PR) curve is obtained by plotting the *Precision* (the percentage of observations classified as positive that are indeed positive) against the *Recall* (the percentage of positive observations classified as positive) as the threshold defining the classifier output varies from *0* to *1*.
The *Area Under PR curve* (AUPR) is the integral below the PR curve.

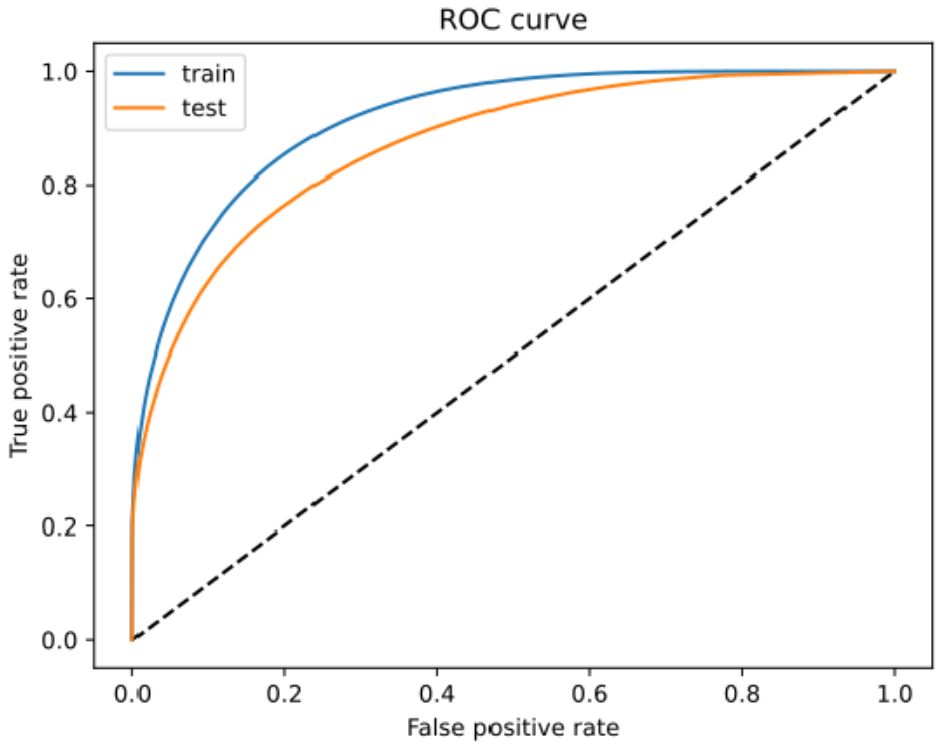
	Training	Test
Accuracy (50%)	82.8381 %	82.8111 %
Accuracy (75%)	82.7985 %	82.7721 %
AUC	70.4532 %	70.4110 %
AUPR	31.3710 %	31.3622 %

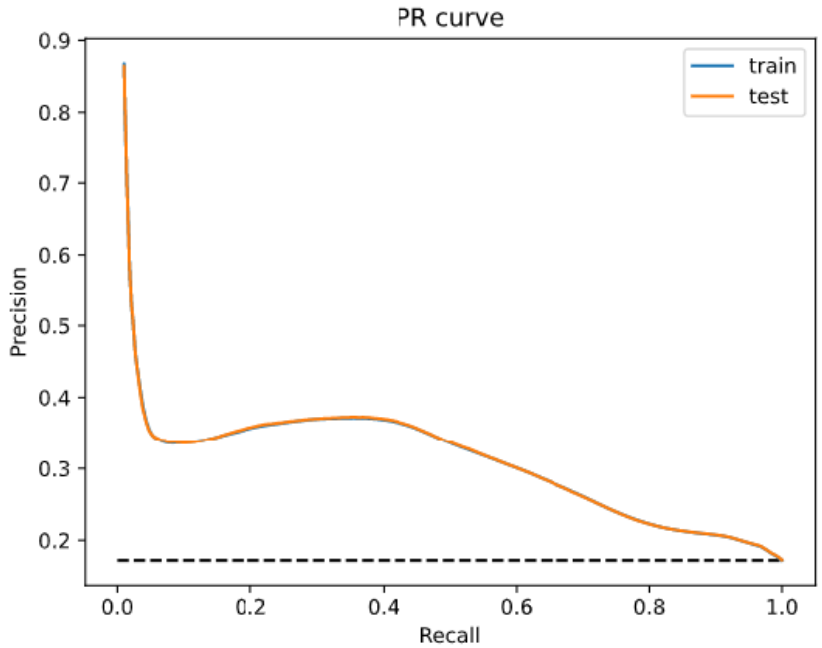
Table : Performance metrics for the logistic regression on the training and test dataset.

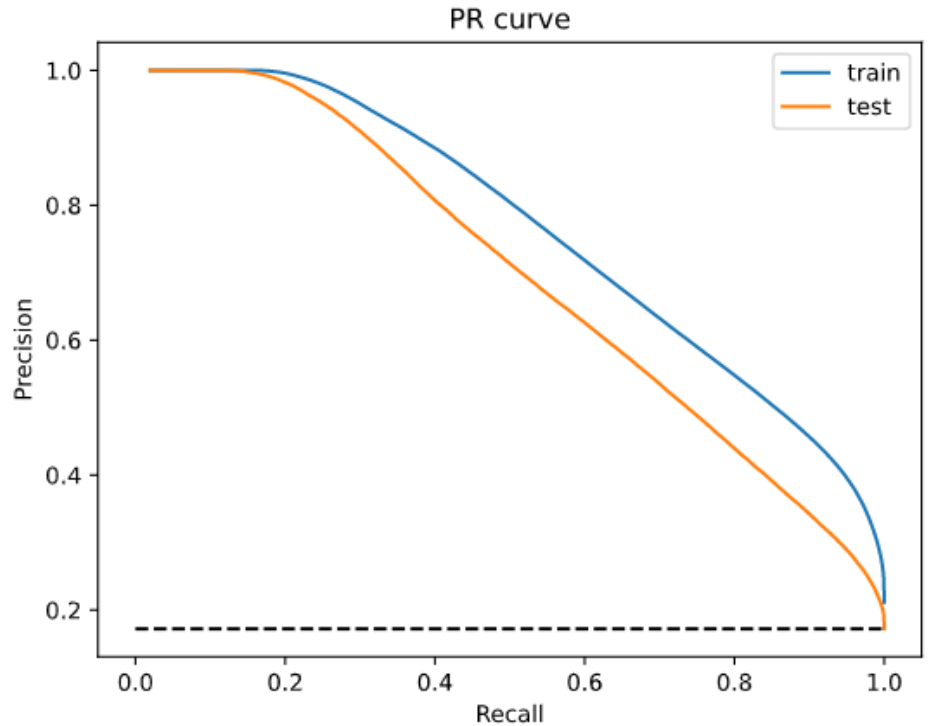
	Training	Validation	Test
Accuracy (50%)	89.1470 %	88.0488 %	88.0271 %
Accuracy (75%)	87.3582 %	87.0261 %	86.9586 %
AUC	92.5806 %	88.2619 %	88.3149 %
AUPR	76.5305 %	69.5697 %	69.7618 %

Table : Performance metrics for the selected RF models on the training, validation and test sets.


ROC curve
train
test
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve
train
test
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


PR curve
train
test
Precision
Recall
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0

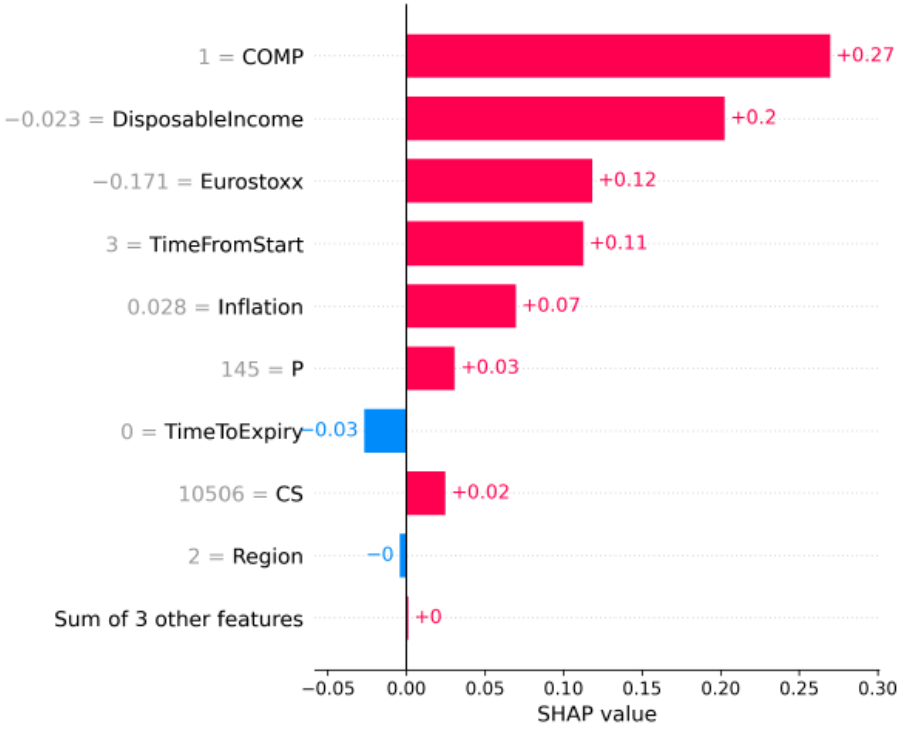

PR curve
train
test
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0

Interpretability

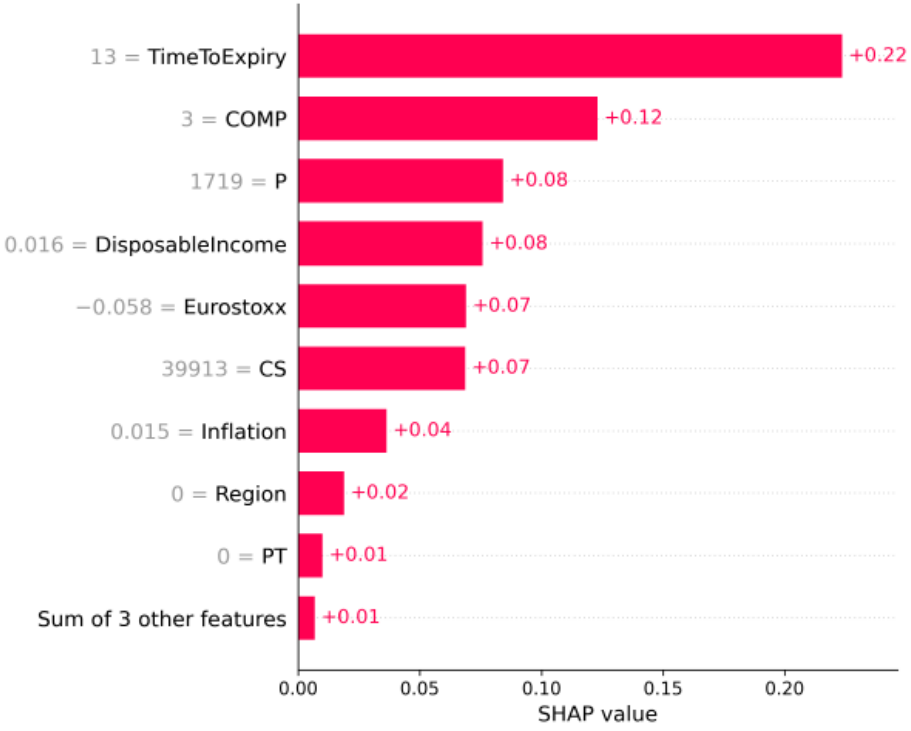
Local Interpretability

- The SHAP (SHapley Additive exPlanations) method allows us to capture the impact of the different variables/features on the ML classifier output. The method borrows from cooperative game theory and consists in the calculation of SHAP value, which represents a measure of the importance of a feature.
- More precisely, the SHAP value of a feature measures how much it contributes, either positively or negatively, to the classifier prediction.
- The goal of the SHAP method is to explain a prediction computing the contribution of each feature to the prediction itself. The method shows the contribution of each feature to push the model output from the base value (the average model output over the training dataset) to the model output associated with the observation.

Local Interpretability

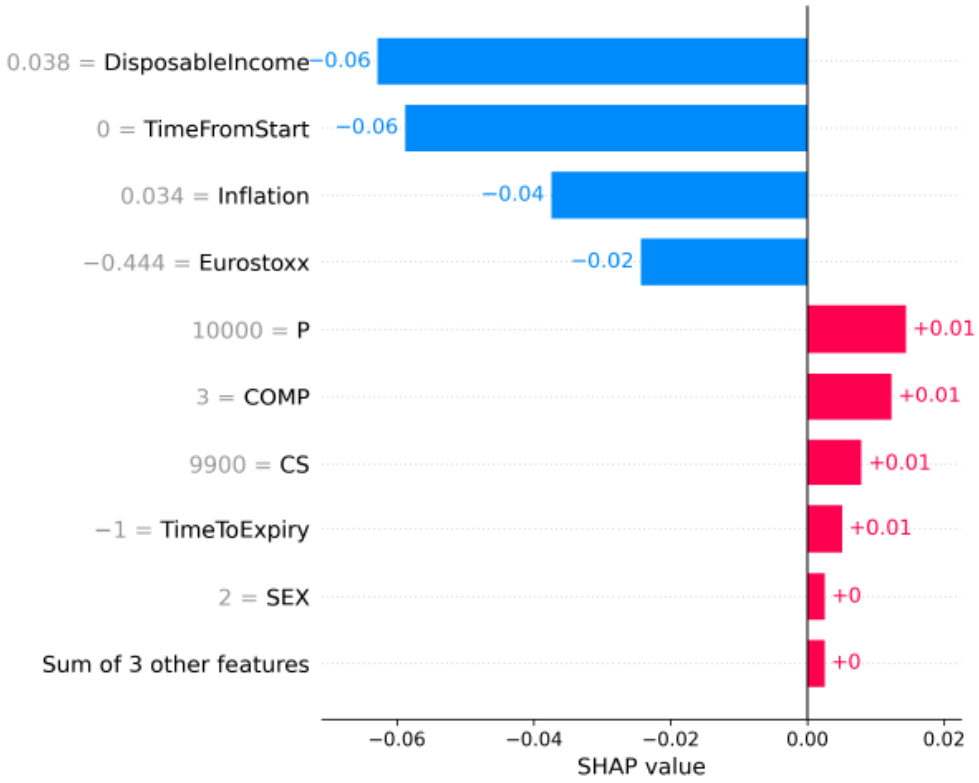


(a) true positive

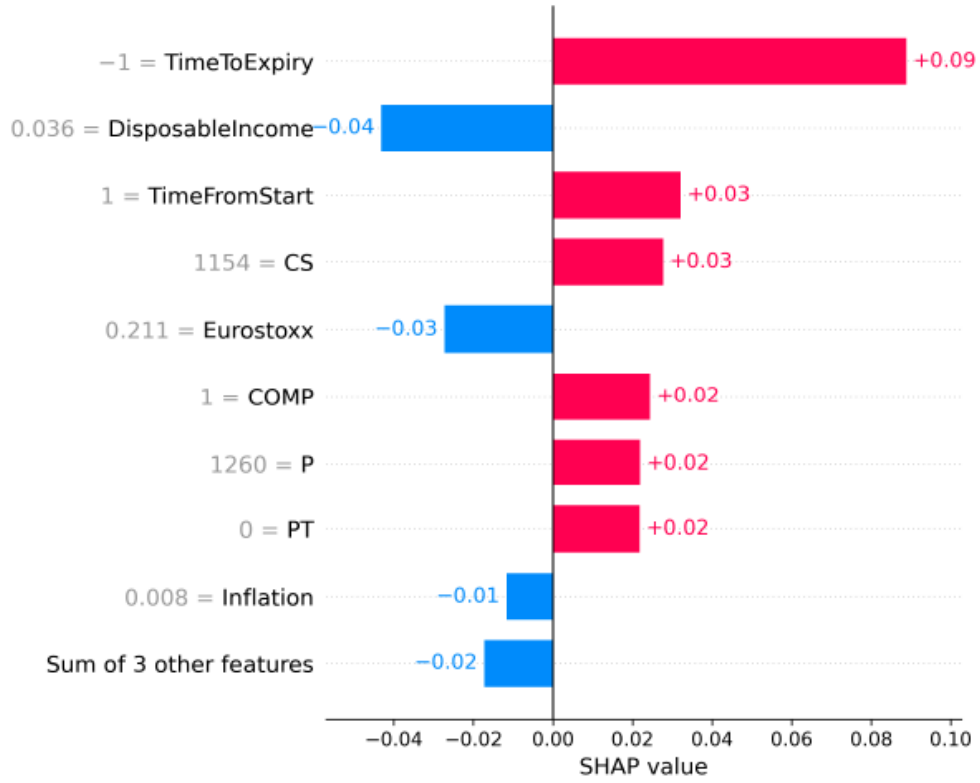


(b) false positive

SHAP values of the four different observations (each feature is accompanied by the corresponding value assumed in the observation). The base rate for the lapsing class is 17.2%. We recall that P is the premium, CS is the contract size and PT is the product type.



(c) true negative



(d) false negative

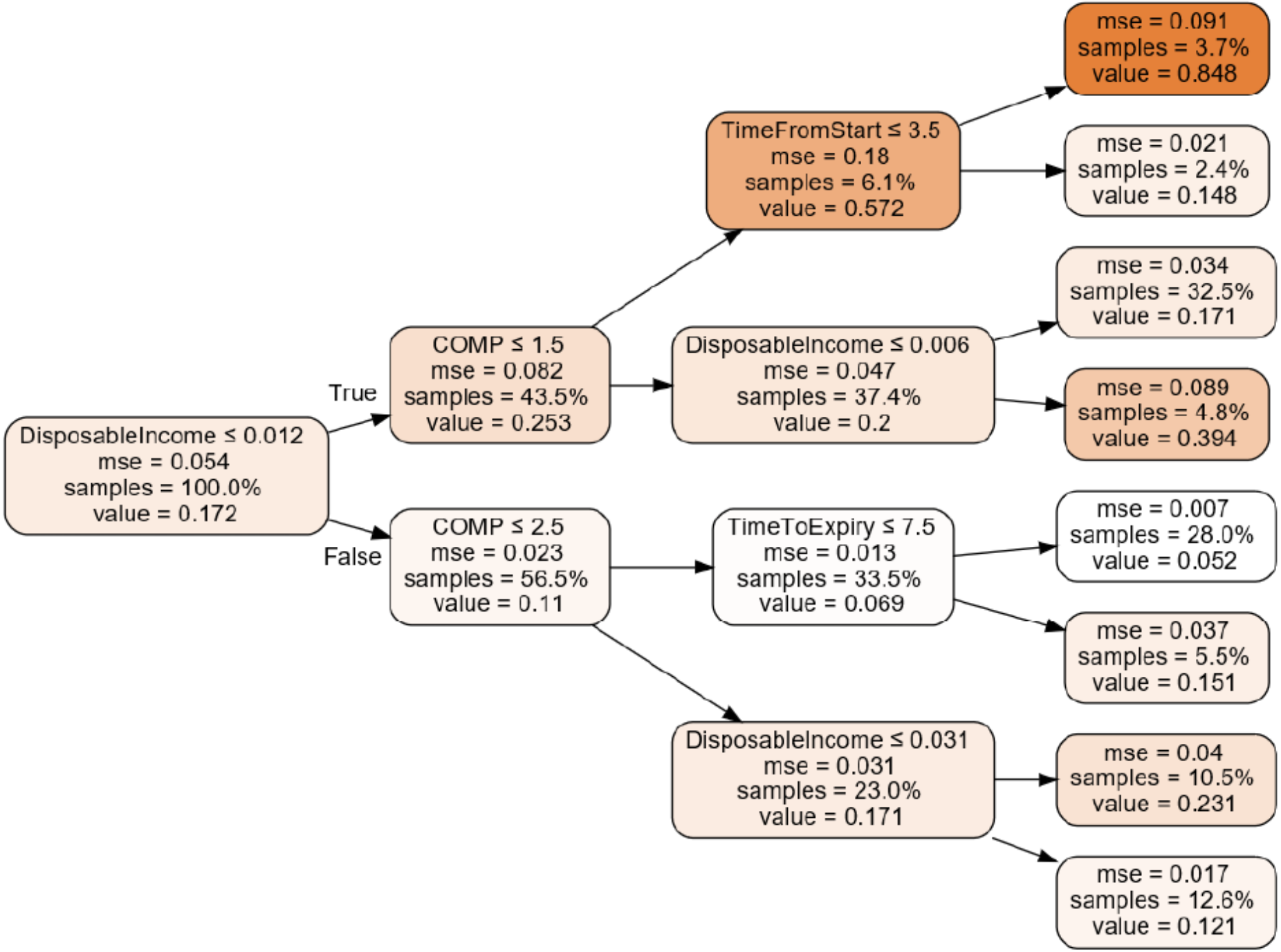
Global Interpretability

- *Local* methods are useful to investigate the results connected to a specific observation. In some applications, it is interesting to have a global view of the model's behavior.
- To accomplish this task, we fit a single regression tree on the predicted probabilities of our RF.
- Regression trees have basically the same structure as classification trees, but the dependent variable is a continuous variable (e.g, the probability to lapse), and the measure of impurity is the MSE of the observations in the node. The quantity predicted by each leaf is the average of the values of the target variable of the observations in the leaf.
- The idea is to estimate a regression tree, with low depth, in such a way that its output is as close as possible to our ``black box'' output (the outcome of the RF).
- The scope is to produce a human-readable tree which is able to represent the main drivers of the ``black box'' classification.

Global Interpretability

depth	2	3	4	5	6	10	25
Accuracy	25.57%	45.00%	55.69%	63.93%	73.10%	88.89%	99.39%
# Features	2	4	6	9	10	12	12

Global interpretability decision tree varying its depth: accuracy in reproducing the "black box" output, and number of features used to fit the random forest with a single decision tree.


DisposableIncome ≤ 0.012
mse = 0.054
samples = 100.0%
value = 0.172
True
False
COMP ≤ 1.5
mse = 0.082
samples = 43.5%
value = 0.253
COMP ≤ 2.5
mse = 0.023
samples = 56.5%
value = 0.11
TimeFromStart ≤ 3.5
mse = 0.18
samples = 6.1%
value = 0.572
DisposableIncome ≤ 0.006
mse = 0.047
samples = 37.4%
value = 0.2
TimeToExpiry ≤ 7.5
mse = 0.013
samples = 33.5%
value = 0.069
DisposableIncome ≤ 0.031
mse = 0.031
samples = 23.0%
value = 0.171
mse = 0.091
samples = 3.7%
value = 0.848
mse = 0.021
samples = 2.4%
value = 0.148
mse = 0.034
samples = 32.5%
value = 0.171
mse = 0.089
samples = 4.8%
value = 0.394
mse = 0.007
samples = 28.0%
value = 0.052
mse = 0.037
samples = 5.5%
value = 0.151
mse = 0.04
samples = 10.5%
value = 0.231
mse = 0.017
samples = 12.6%
value = 0.121

- We notice that observations with disposable income growth rate smaller than 0.012 are more likely to lapse. We also notice the relevance of the Company (COMP) which acts at the second level of the tree. For example, if the disposable income growth rate is low and the policy is stipulated with Company A (COMP=1), then the probability to lapse increases considerably, reaching a 57% probability.
- According to our analysis, the leaf with the highest lapse probability (84.8%) is obtained with a disposable income growth rate lower (or equal) than 0.012, a policy stipulated with Company A and with time from start smaller (or equal) than 3.5 years.
- On the other hand, the leaf with the smallest lapse probability (5.2%) is obtained with a disposable income growth rate higher than 0.012, a policy stipulated with Company A or B, and with time to expiry smaller (or equal) than 7.5 years.
- Notice that the role of disposable income growth rate is controversial: while in the first node a low disposable income growth rate results in an increase of the probability to lapse (from 17.2\% to 25.3\%), the opposite happens in the second node of the third level.
- We can conclude that the contribution of the growth rate of the disposable income to the lapse decision is nonlinear and can not be fully captured through the coefficient of a linear model, as in the logistic regression. Notice that also Company A is crucial both in determining a lapse and a non-lapse event. We can conclude that the interaction among variables is truly nonlinear.
- The role of the Company, e.g. launching of new products or the frequency with which products are proposed to the customers could have an important impact on the lapse decision, plays an important role

Conclusions

- The reason to analyze the lapse decision through ML methodologies is that the different hypotheses proposed in the literature are not able to capture adequately the phenomenon.
- Behavioral and commercial reasons seem to play a significant role.

To investigate them in an agnostic way, it is useful to use ML tools that allow data to speak for themselves.

We confirm that a RF performs better than the classical logistic model to predict the lapse decision.
The ML methodology allowed us to discern the relevance of a wide set of exogenous variables:

- The main result of our analysis is that the important drivers of the lapse decision are the time passed from the incipit of the contract and the time to expiry, as well as the insurance company, the contract size, and premium.
- Other features of the policyholder (gender, age of the policyholder, region) or of the contracts (product type) as well as macroeconomic variables (with the exception of the disposable income growth rate with a nonlinear effect) play a limited role.