Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata

Matheus Kempa Severino
Yaohao Peng

University of Brasilia
Laboratory of Machine Learning in Finance and Organizations

Outline

1 Motivation

2 Data and methods

3 Results

4 Conclusions and remarks

Motivation

Fraud detection is a relevant issue for the insurance market

- Brazil in 2017: USD 10.8 billion USD paid in insurance policies
  - Total value of all occurred claims: approximately USD 10.0 billion
  - Total value of proven frauds: USD 221.2 million

Motivation

- Machine learning methods can aid risk analysts in the fraud detection task
  - However, many models from this class are considered to be "black boxes", in the sense of providing results that are difficult to be interpreted
- Machine learning methods can aid risk analysts in the fraud detection task (Ngai et al., 2011; Awoyemi et al., 2017; Jurgovsky et al, 2018; Raghavan and El Gayar, 2019)
  - The literature is concentrated on credit card and telecommunication frauds, and healthcare/automobile applications for insurance frauds – studies about frauds for property insurances, especially for residential policies, are relatively scarce (Sinayobye et al., 2018)

Main contributions

1. Empirically evaluate the predictive performance of machine learning models using real-world microdata
2. Rank the features (independent variables) for each tested model with respect to their **global relevance**
3. Identify which features had a stronger **local impact** for the predictions on prominent false positive and false negative observations

Outline

1 Motivation

2 Data and methods

3 Results

4 Conclusions and remarks

Data

- Microdata between 2009 and 2018 from a major Brazilian insurance company, labeled and balanced (851 observations)
- Variables: Product Type, Coverage type, Contract channel, Automatic renewal, Past renewal, Legal person, Number of payment installments of the insurance value, Time of approval of the insurance policy, Differences of Days between contract term start/end and claim date, Insured amount, Insurance premium, Age, Gender, Income range, Marital status, and Number of previous claims of the customer in the company

Models and empirical experiments

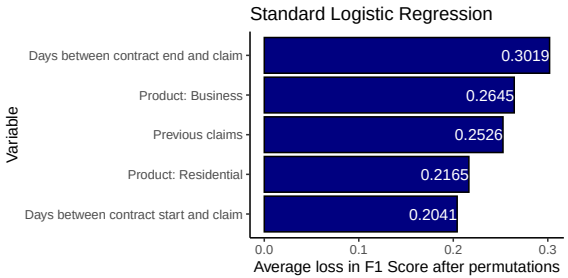
- Machine learning models: Standard logistic regression, Logistic regression with elastic-net regularization, Naive Bayes, K-Nearest Neighbors, Support Vector Machine (Polynomial/Gaussian Kernel, Deep Neural Network, Random Forest, and Gradient Boosting Machine
- 10-fold cross-validation (200 observations for the training set, 651 for testing set), repeated over 1000 samples
- Evaluation metrics: Accuracy, Precision, Recall, F1 Score, Kappa, and MCC
- Robustness check: Model Confidence Set procedure

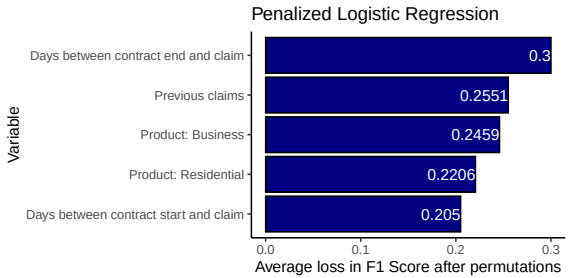
eXplainable Artificial Intelligence (XAI)

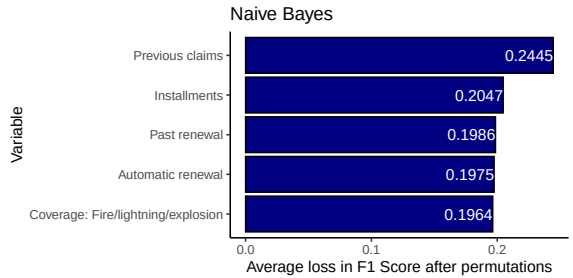
XAI methods aim to provide interpretable outputs to machine learning models while maintaining their flexibility and generalization power

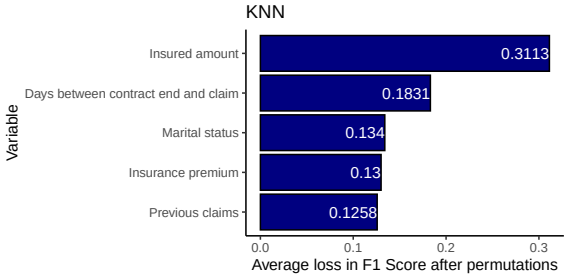
- Global importance: Model-agnostic permutation-based approach (impact on F1 Score)
- Local importance: Shapley Additive Explanation

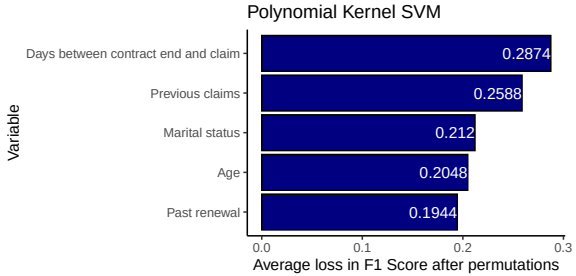
Global variable importance

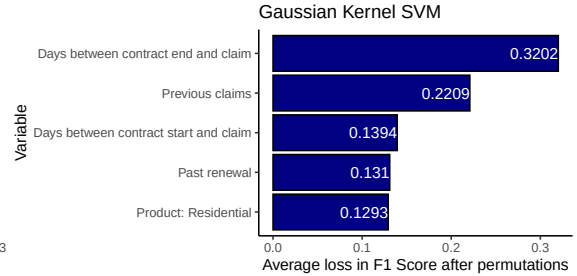


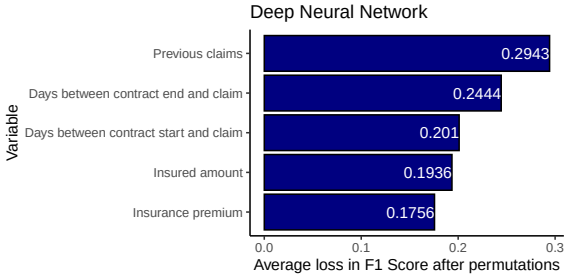


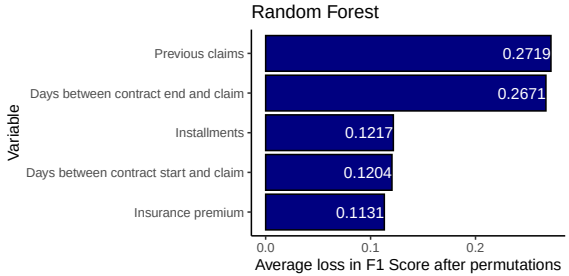


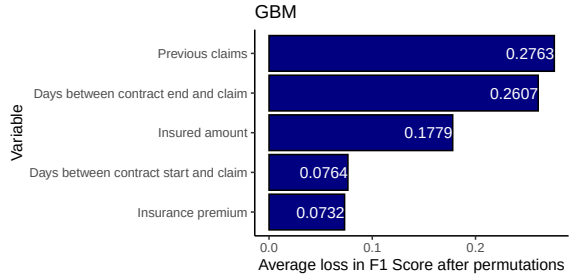












Shapley Additive Explanation (SHAP)

For each observation of interest $x_{\#} \in \mathbb{R}^p$, the SHAP value for the i-th feature is:

$$\varphi(x_{\#}, i) = \frac{1}{p!} \sum_{\Pi} \Delta^{i|\pi(\Pi,i)}(x_{\#})$$

$$\begin{aligned}\Delta^{i|\pi(\Pi,i)}(x_{\#}) =& \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})|x_{\pi(\Pi,i)_1} = x_{\pi(\Pi,i)_{1_{\#}}}, \ldots, x_{\pi(\Pi,i)_{|\pi(\Pi,i)|}} = x_{\pi(\Pi,i)_{|\pi(\Pi,i)|_{\#}}}, x_i = x_{i_{\#}}\} - \\ & \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})|x_{\pi(\Pi,i)_1} = x_{\pi(\Pi,i)_{1_{\#}}}, \ldots, x_{\pi(\Pi,i)_{|\pi(\Pi,i)|}} = x_{\pi(\Pi,i)_{|\pi(\Pi,i)|_{\#}}}\}\end{aligned}$$

- $\pi(\Pi, i)$ is the set of the indexes associated with features that comes before the i-th variable in permutation Π
  - For permutation $\Pi = \{3, 1, 4, 2\}$, $\pi(\Pi, 4) = \{3, 1\}$ and $\pi(\Pi, 3) = \emptyset$

Shapley Additive Explanation (SHAP)

For each observation of interest $x_{\#} \in \mathbb{R}^p$, the SHAP value for the i-th feature is:

$$\varphi(x_{\#}, i) = \frac{1}{p!} \sum_{\Pi} \Delta^{i|\pi(\Pi,i)}(x_{\#})$$

$$\begin{aligned}\Delta^{i|\pi(\Pi,i)}(x_{\#}) =& \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})|x_{\pi(\Pi,i)_1} = x_{\pi(\Pi,i)_{1_{\#}}}, \ldots, x_{\pi(\Pi,i)_{|\pi(\Pi,i)|}} = x_{\pi(\Pi,i)_{|\pi(\Pi,i)|_{\#}}}, x_i = x_{i_{\#}}\} - \\ & \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})|x_{\pi(\Pi,i)_1} = x_{\pi(\Pi,i)_{1_{\#}}}, \ldots, x_{\pi(\Pi,i)_{|\pi(\Pi,i)|}} = x_{\pi(\Pi,i)_{|\pi(\Pi,i)|_{\#}}}\}\end{aligned}$$

- $p!$ is the number of possible permutations (orderings) between the p features

Shapley Additive Explanation (SHAP)

For each observation of interest $x_{\#} \in \mathbb{R}^p$, the SHAP value for the i-th feature is:

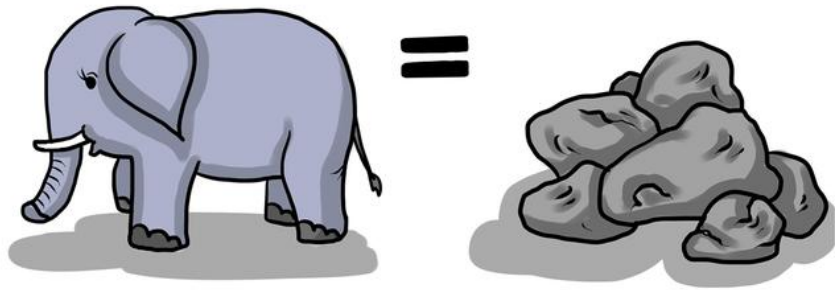
$$\varphi(x_{\#}, i) = \frac{1}{p!} \sum_{\Pi} \Delta^{i|\pi(\Pi,i)}(x_{\#})$$

$$\begin{aligned}\Delta^{i|\pi(\Pi,i)}(x_{\#}) =& \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})|x_{\pi(\Pi,i)_1} = x_{\pi(\Pi,i)_{1_{\#}}}, \ldots, x_{\pi(\Pi,i)_{|\pi(\Pi,i)|}} = x_{\pi(\Pi,i)_{|\pi(\Pi,i)|_{\#}}}, x_i = x_{i_{\#}}\} - \\ & \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})|x_{\pi(\Pi,i)_1} = x_{\pi(\Pi,i)_{1_{\#}}}, \ldots, x_{\pi(\Pi,i)_{|\pi(\Pi,i)|}} = x_{\pi(\Pi,i)_{|\pi(\Pi,i)|_{\#}}}\}\end{aligned}$$

- $\Delta^{i|\pi(\Pi,i)}(x_{\#})$ is the difference between the expected predictions using the features with indexes $i \cup \pi(\Pi, i)$ and $\pi(\Pi, i)$

Shapley Additive Explanation (SHAP)

The intuition of SHAP values is rather similar to Cao Chong (196–208)'s method to weigh an elephant



Shapley Additive Explanation (SHAP)

The intuition of SHAP values is rather similar to Cao Chong (196–208)'s method to weigh an elephant

- SHAP's equivalent of the buoyancy principle is the **local accuracy** property:

$$f(x_{\#}) = \sum_{i=1}^{p} \varphi(x_{\#}, i) + \mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})\}$$

Shapley Additive Explanation (SHAP)

For instance, to calculate the SHAP value of a specific observation $x_{\#}$ for x_2 with $p = 3$:

Permutation	**A** = With x_2	**B** = Without x_2	**A - B**
$\Pi_1 = \{1, 2, 3\}$	$i \cup \pi(\Pi_1, i) = \{1, 2\}$	$\pi(\Pi_1, 2) = \{1\}$	$\Delta^{i\|\pi(\Pi_1,2)}(x_{\#})$
$\Pi_2 = \{1, 3, 2\}$	$i \cup \pi(\Pi_2, i) = \{1, 3, 2\}$	$\pi(\Pi_2, 2) = \{1, 3\}$	$\Delta^{i\|\pi(\Pi_2,2)}(x_{\#})$
$\Pi_3 = \{2, 1, 3\}$	$i \cup \pi(\Pi_3, i) = \{2\}$	$\pi(\Pi_3, 2) = \emptyset$	$\Delta^{i\|\pi(\Pi_3,2)}(x_{\#})$
$\Pi_4 = \{2, 3, 1\}$	$i \cup \pi(\Pi_4, i) = \{2\}$	$\pi(\Pi_4, 2) = \emptyset$	$\Delta^{i\|\pi(\Pi_4,2)}(x_{\#})$
$\Pi_5 = \{3, 1, 2\}$	$i \cup \pi(\Pi_5, i) = \{3, 1, 2\}$	$\pi(\Pi_5, 2) = \{3, 1\}$	$\Delta^{i\|\pi(\Pi_5,2)}(x_{\#})$
$\Pi_6 = \{3, 2, 1\}$	$i \cup \pi(\Pi_6, i) = \{3, 2\}$	$\pi(\Pi_6, 2) = \{3\}$	$\Delta^{i\|\pi(\Pi_6,2)}(x_{\#})$

The average of (**A** - **B**) over all feature permutations measures the relative impact of x_2 on the model's prediction for observation $x_{\#}$

Outline

1 Motivation

2 Data and methods

3 Results

4 Conclusions and remarks

Macro-profile of the frauds

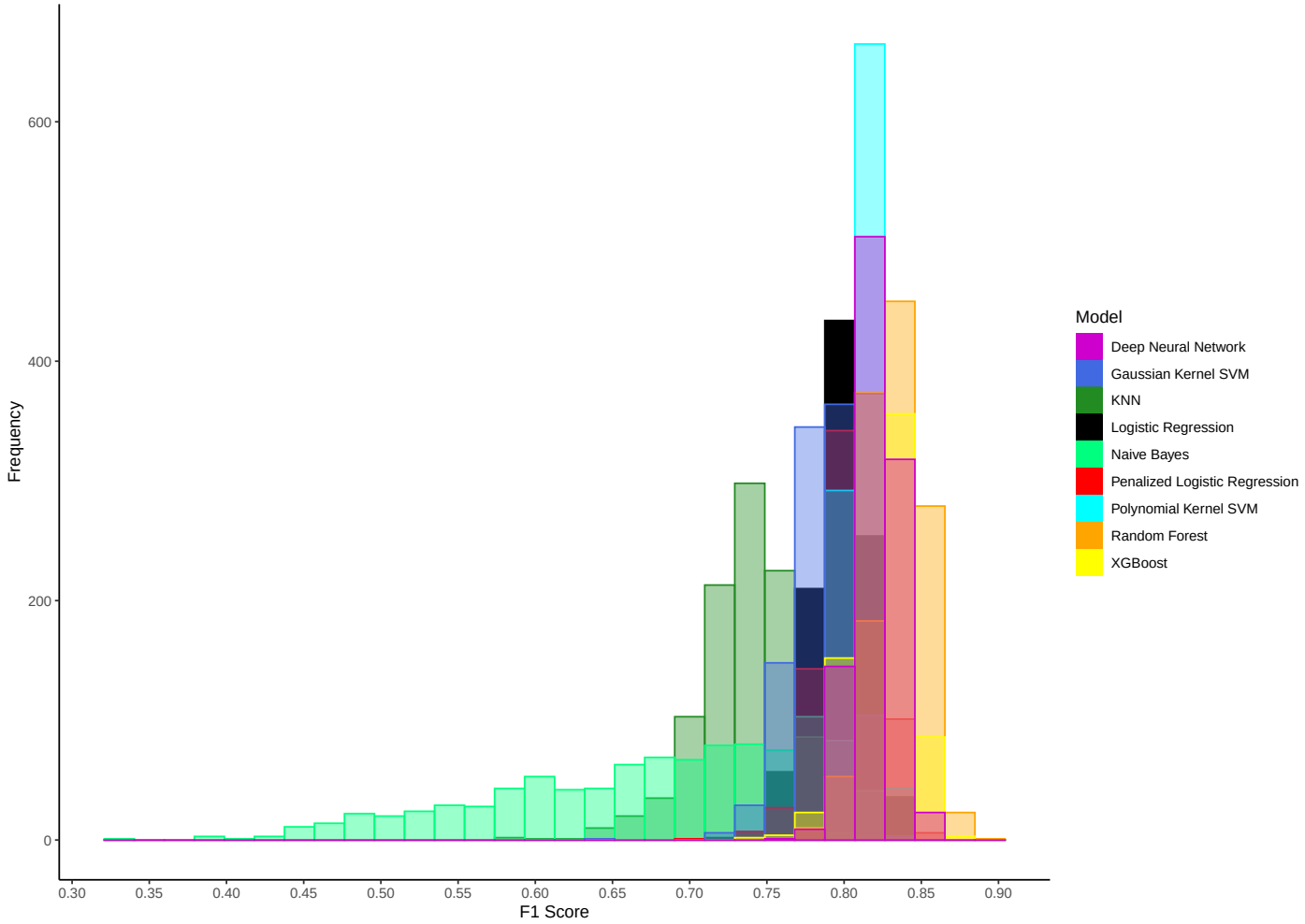
1. 60.14% of the fraudsters were male;
2. 48.16% of the frauds were premature claims;
3. 52.81% of the fraudsters were non-married;
4. 79.95% of the total coverage amount was for electrical damage or theft claims;
5. The average age of fraudsters was 41 years;
6. 72.61% of the frauds were new insurance policies;
7. Fire/lightning/explosion coverage had the highest average payment amount for detected but unproven frauds.

Classification results

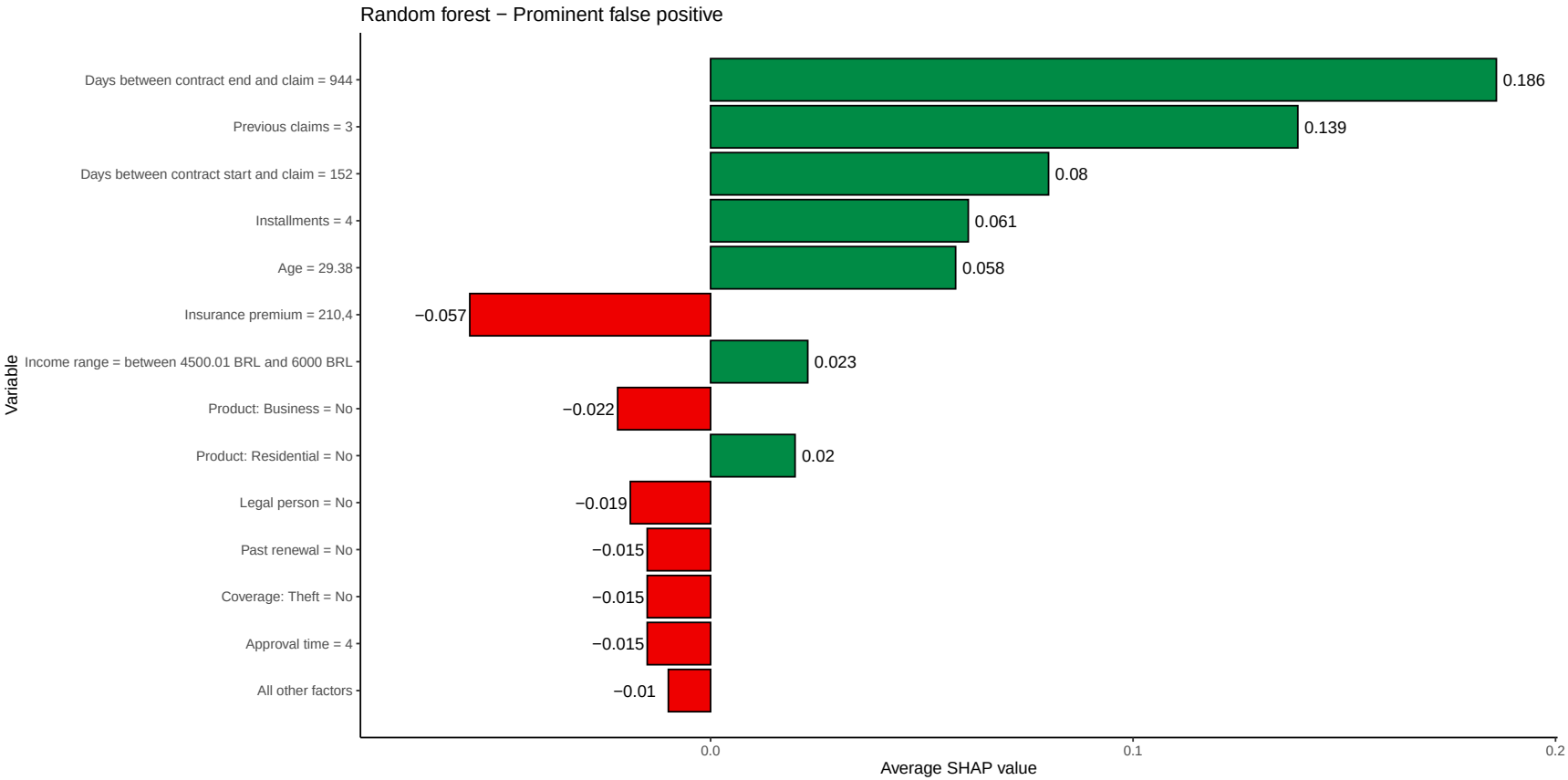
Model	Accuracy	Precision	Recall	F1 Score	Kappa	MCC
Logistic Regression	80.67% (1.60%)	80.56% (2.94%)	78.99% (3.79%)	79.67% (1.81%)	61.26% (3.21%)	61.41% (3.18%)
Penalized Logistic Regression	81.40% (1.72%)	81.27% (3.36%)	79.95% (3.99%)	80.48% (1.88%)	61.34% (3.19%)	62.93% (3.36%)
Naive Bayes	71.16% (5.66%)	73.18% (8.64%)	73.02% (5.53%)	72.39% (3.72%)	47.66% (10.28%)	49.51% (8.78%)
KNN	75.74% (2.51%)	77.74% (3.76%)	69.77% (4.67%)	73.39% (2.88%)	51.25% (5.01%)	51.66% (4.98%)
Polynomial Kernel SVM	81.34% (0.75%)	79.22% (1.15%)	82.98% (1.06%)	80.93% (0.84%)	62.92% (1.48%)	63.00% (1.48%)
Gaussian Kernel SVM	79.56% (1.71%)	79.07% (3.08%)	78.41% (4.17%)	78.53% (1.98%)	58.81% (3.36%)	59.00% (3.31%)
Deep Neural Network	81.88% (1.58%)	78.41% (3.11%)	**86.28% (3.06%)**	82.06% (1.32%)	63.84% (3.08%)	64.32% (2.80%)
Random Forest	**84.56% (1.43%)**	**84.72% (2.60%)**	82.77% (3.72%)	**83.61% (1.65%)**	**69.05% (2.97%)**	**69.24% (2.88%)**
GBM	83.21% (1.61%)	83.55% (2.97%)	81.73% (3.96%)	82.44% (1.82%)	66.20% (3.08%)	66.39% (3.02%)

Standard deviations are in parenthesis; MCS superior models at the 95% level are in bold

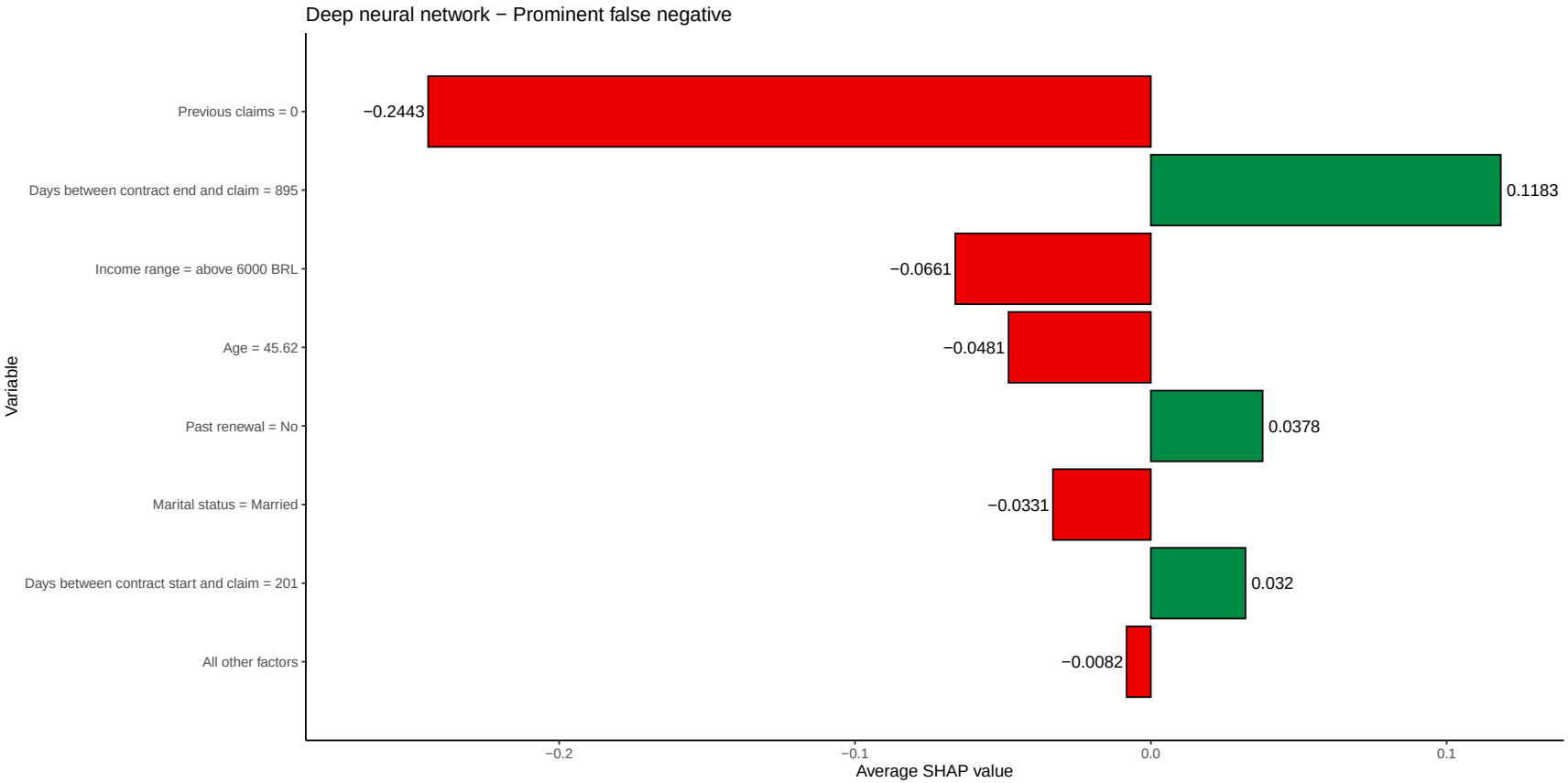
Classification results



Local feature importance: prominent false positive



Local feature importance: prominent false negative



Outline

1 Motivation

2 Data and methods

3 Results

4 Conclusions and remarks

Main conclusions

- Random forest exhibited the best overall out-of-sample results
- Deep neural network was the most consistent model in avoiding false negatives
- Number of previous claims and indicators of premature claims were relevant to detect frauds, but also had strong impacts on misclassified observations

Recommendations for future studies

- Consider additional models and hyperparameter tuning
- Take a deeper look at ensemble-based models
- Probabilistic approach for insurance premium pricing
- Incorporate the spatial dimension into the models

Thank you!

lamfo.unb.br

lamfo@unb.br

lamfo-unb.github.io