

Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending



Miller Ariza

Javier Arroyo

Antonio Caparrini

Maria J. Segovia

Universidad Complutense de Madrid

- In finance, interpretability is essential for regulatory entities and end-users (Carvalho, 2019; Babel et al., 2019), as it was stated in Basel III and by the Financial Stability Board in 2017
- In the FinTech project we are in contact with industry and regulators that have highlighted that need
 - We included a use case on machine learning explainability (Bussmann, Giudici, Marinelli, & Papenbrock, 2020)

- After a suptech seminar on credit assessment for p2p lending, Bank of Spain was concerned on how to make interpretable machine learning models for credit scoring
 - We decided to explore the topic
- The work that we will present here is the result of that exploration
 - While the work is on p2p lending most of the conclusions can be interpreted on credit risk in general

- P2P lending eliminates the intermediation of traditional institutions, and consequently, **information asymmetry** is much more marked than in traditional banking.
- P2P platforms offer **scoring** models to mitigate risk, protect investors and maintain financial stability (Ahelegbey, Giudici and Hadji-Misheva, FinTech-ho2020, 2019).
- Reliable, interpretable and explainable credit risk is even more challenging in P2P lending than in traditional products.

- **Machine learning** is a powerful prediction tool, but banks, regulators and end-users are proceeding with caution in its adoption for credit risk modeling due to the lack of interpretability elements.
- Thus, they typically rely on statistical and interpretable models such as logistic regression

- Interpretability is a construct that includes several aspects: **explainability**, causality, transparency, transferability, ...
 - Ontological and epistemological perspective (Carvalho et al., 2019) .
- New tools are being developed to add interpretability to machine learning methods
 - Shapley values, LIME, DeepLift, **SHAP values**...

- Logistic Regression (LR) is typically used in credit risk.
 - Regulatory entities and end users require interpretable models and hence typically rely on models such as LR.
 - LR provides a high interpretability but, possibly, a limited predictive accuracy (Bussmann, Giudici, Marinelli, & Papenbrock, 2020).
 - LR models sometimes include new variables from text, network data, etc... and interpretability is used) to include new variables and assess their predictability and their interpretability.

- It is rare to find studies that use ML methods and include some type of explainability or interpretability analysis.
- Most studies focus on accuracy or on classification performance.
- The most frequent interpretability element is feature importance for methods based on decision trees.
- This lack is noted in surveys about the use of machine learning in finance (Lessmann et al., 2015; Andriosopoulos, 2019; Leo et al., 2019).

- Private companies are turning their attention to interpretability
 - E.g. The Explainable Machine Learning Challenge proposed by the Fair Isaac Corporation (FICO).
- Interpretability is one of highlighted aspects to be discussed in the financial digitalization project in the EU (Expert Group on Regulatory Obstacles to Financial Innovation (ROFIEG, 2019)

“...top recommendations in terms of regulatory reform, highlighting the need to address as a matter of priority:

The explainability and interpretability of technology, especially AI, as measures to protect consumers and businesses and facilitate supervision, or to meet supervisory expectations ...”

- Here we fill the gap in the literature with an **in-depth comparison** of the **predictability and interpretability** of **machine learning models and logistic regression** for granting scoring in p2p lending.
- We use the **SHAP values** to interpret machine learning which can reflect dispersion, nonlinearity and structural breaks in the relationships between each feature and the target variable.
- The comparison reveals that the machine learning alternative is superior in terms of not only classification performance but also explainability.

1 Data preparation

- Data Cleaning
- Partition:
 - Training set: From June 2007 to July 2015
 - Test set: from August 2015 to December 2018
- Definition of credit default event
- Definition of input variables



2 Model adjustment

- LR fitting with standard set
- LR fitting with balanced set
- Machine Learning hyperparameter optimization
 - Genetic algorithm
 - Maximization of BAC
 - K-fold CV (k=4)
- Machine Learning fitting (DT, RF, XGBoost)



3 Model evaluation

- Model selection using BAC
- Global performance assessment
 - Statistical tests on
 - ACC
 - AUC
 - KS
- Performance assesment for each class
 - Precision, recall and F1 score



4 Model explanation

- Coefficient analysys and inference (only LR)
 - Shap values based assessment
- Feature importance, monotonicity and SHAP values
 - Feature importance
 - Monotonicity
 - Dependence plots



- Data from Lending Club available in Kaggle and extensively analyzed in the literature
- The dataset has 1,347,681 obligations, we split it in a chronological way
 - Training from June 2007 to July 2015: 657,602 obligations
 - Test from August 2015 to December 2018: 690,079 obligations
- Definition of the credit default event

$$Y = \begin{cases} 1 & \text{charged off } (\cong 20\%) \\ 0 & \text{fully paid} \end{cases}$$

- As input variables we only considered those available at the time of the application (granting model)
 - Categorical
 - Employment length
 - Previous experience with Lending Club
 - Purpose of the loan
 - Home ownership
 - State in US
 - Quantitative
 - Revenue
 - Debt ratio (dti)
 - Loan amount
 - FICO (credit bureau score)

- Logistic Regression
 - Standard model (**LR**)
 - Data set adjusted for class imbalance (**LR.BS**)
 - Adjusted using a hybrid of under and oversampling
- Machine learning models
 - Decision Trees (**DT**), Random Forest (**RF**), eXtreme Gradient Boosting (**XGB**)
 - Adjusted using hyperparameter optimization in a k-fold cross validation setting (k=4)
 - Genetic Algorithm with 20 epochs of 50 individuals
 - Maximizing Balanced Accuracy (BAC)

- Model selection using Balanced Accuracy (BAC)
- Global performance assessment
 - Accuracy (ACC), AUC and Kolmogorov-Smirnoff (KS) validate with statistical tests
- Performance assessment for each class to better understand the classifier
 - Precision, recall and F1 score

Model	Training				Test			
	BAC	ACC	AUC	KS	BAC	ACC	AUC	KS
LR	50.1	81.9	65.4	22.1	50.2	78.1	66.6	23.9
LR.BS	61.0	61.0	65.4	22.1	61.9	60.5	66.6	23.8
DT	61.5	58.5	66.1	22.2	60.5	60.8	64.7	18.0
RF	62.7	62.7	67.8	26.1	61.4	64.4	66.3	24.0
XGB	62.8	62.6	68.0	27.2	62.4	63.6	67.4	26.4



Statistical significance of the results in the test set


Differences	ACC ^a	AUC ^b	KS ^c
XGB-LR	-14.54***	0.81***	2.51***
XGB-LR.BS	3.11***	0.84***	2.58***
XGB-DT	2.82***	2.71***	8.49***
XGB-RF	-0.78***	1.06***	2.36***

***significant at the 0.01 level.

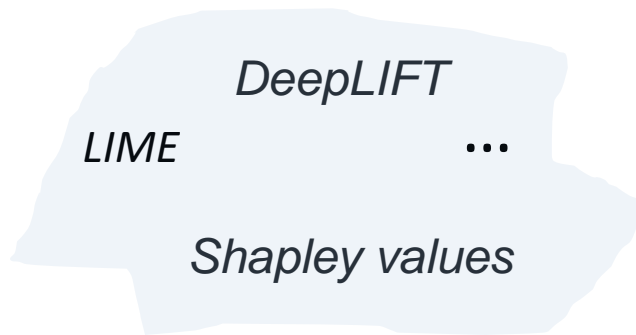
^a McNemar test, ^b DeLong test and ^c Krzanowski-Hand test.

Results for each class in the test set

Model	Default			Non default		
	Precision	Recall	F1	Precision	Recall	F1
LR	44.4	0.7	1.3	78.3	99.8	87.7
LR.BS	30.7	64.4	41.6	85.7	59.4	70.2
DT	30.1	60.1	40.1	84.6	61.0	70.9
RF	32.0	56.0	40.7	84.5	66.7	74.5
XGB	32.1	60.2	41.9	85.3	64.5	73.5

- 
- XGB globally performs better than the other methods.
 - However, the results illustrate the complexity of the classification problem and the subtlety of the different behaviors that the classifiers exhibit.

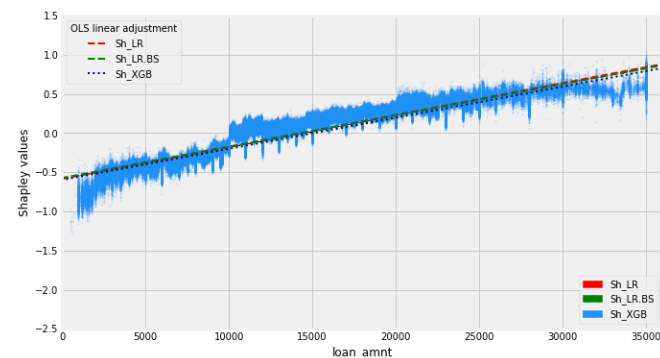
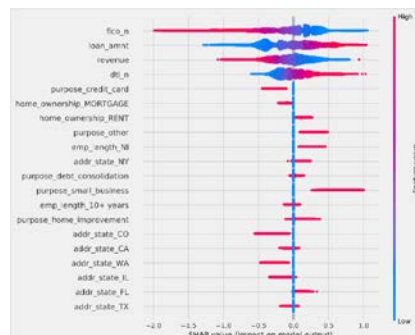
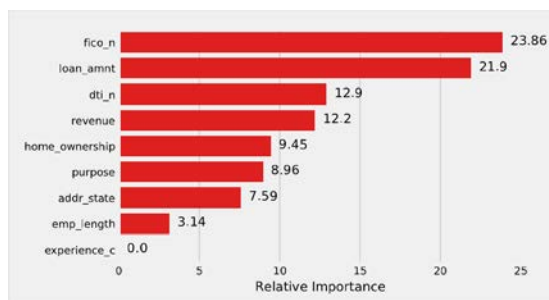
- Logistic Regression
 - Coefficients, marginal effects and odds ratios
 - Shapley values
- XGBoost
 - SHAP values: Post-hoc surrogate models of ML models



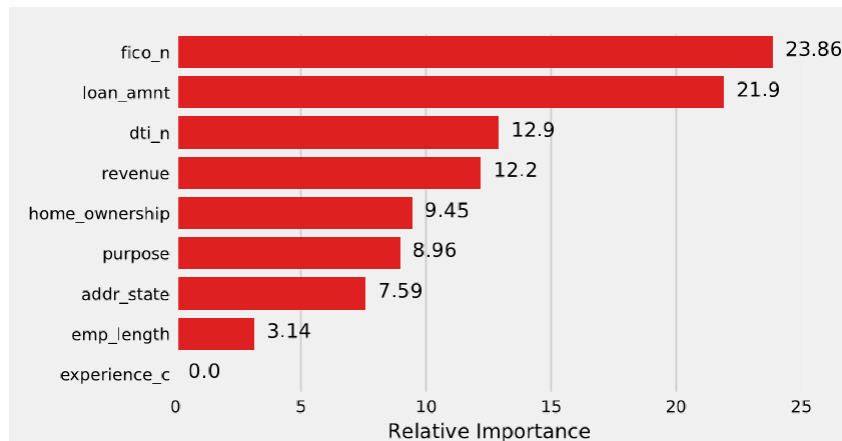
SHAP values
(Lundberg et al., 2017,
2018, 2019 and 2020)

- SHAP values estimate the contribution of each feature to the prediction of an instance.
 - These values produce an approximate explanation of the prediction.
- The individual predictions can be aggregated to measure the variable contribution.
- The contribution of categorical variables are estimated for each category
 - We adjust the SHAP values for the categories of categorical variables to better account for dependence among categories.

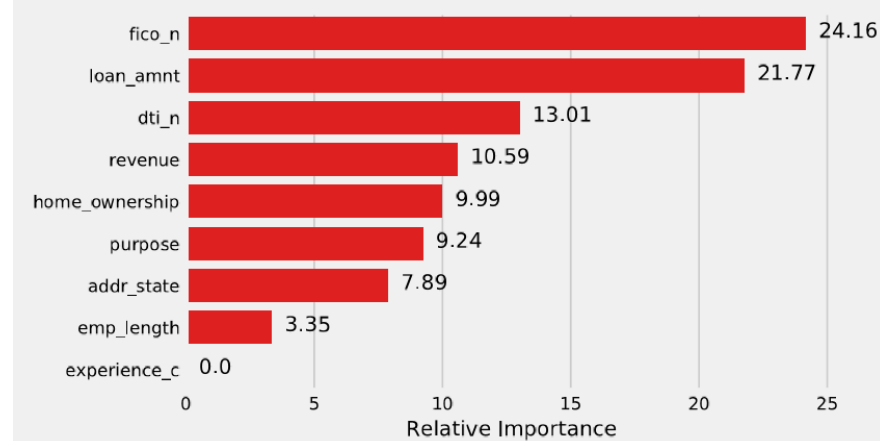
- SHAP values provides graphical tools to represent them:
 - Feature importance
 - Monotoncities
 - Nonlinearities
 - E.g. structural breaks and heteroscedasticities
 - Outlier detection



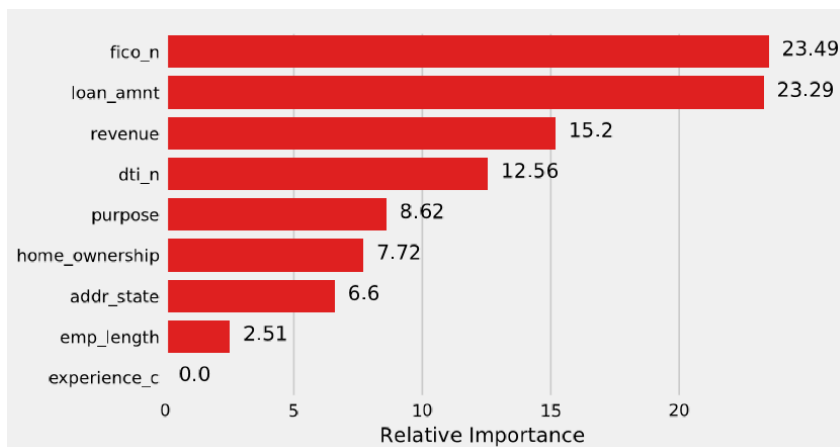
LR



LR-BS

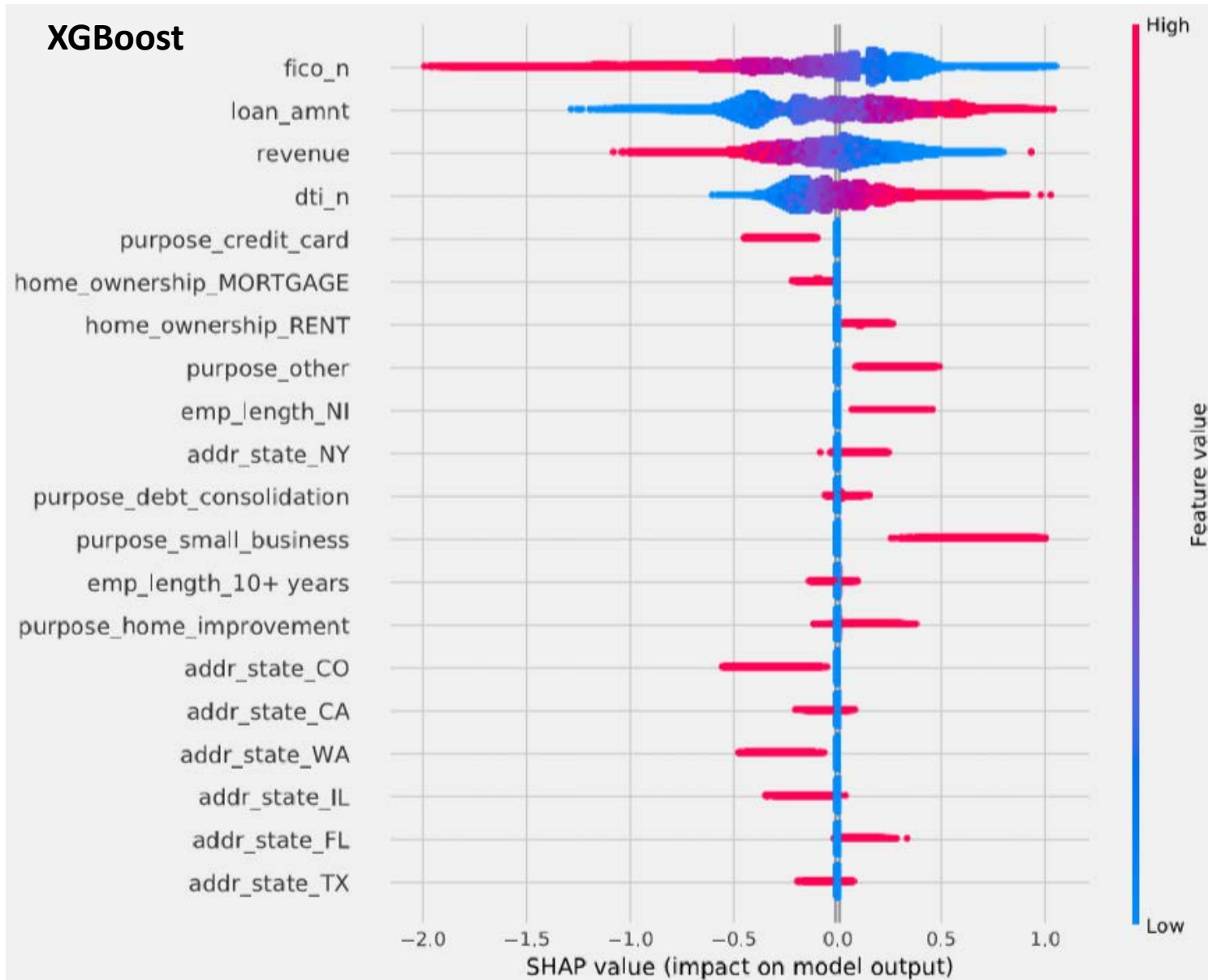


XGBoost

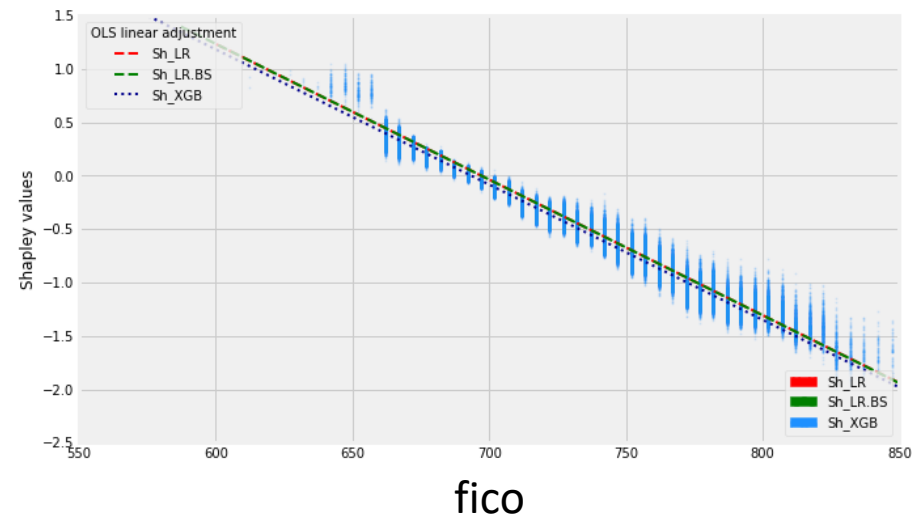
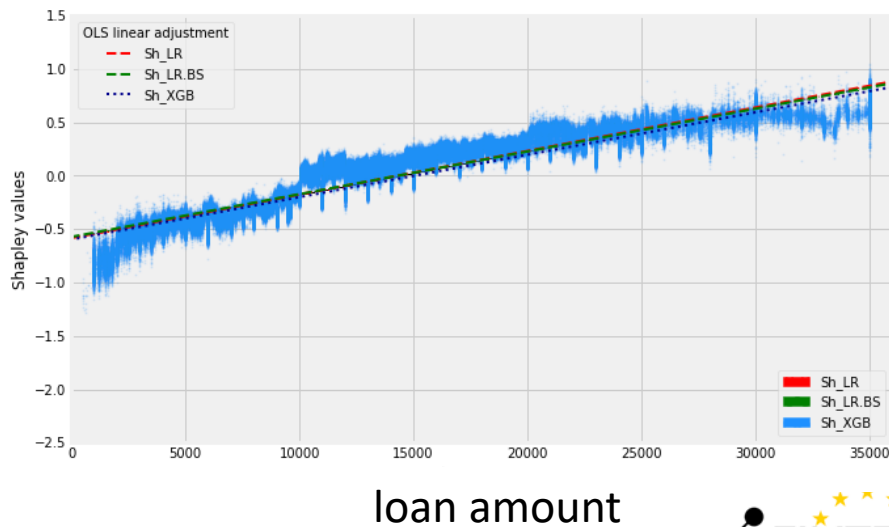
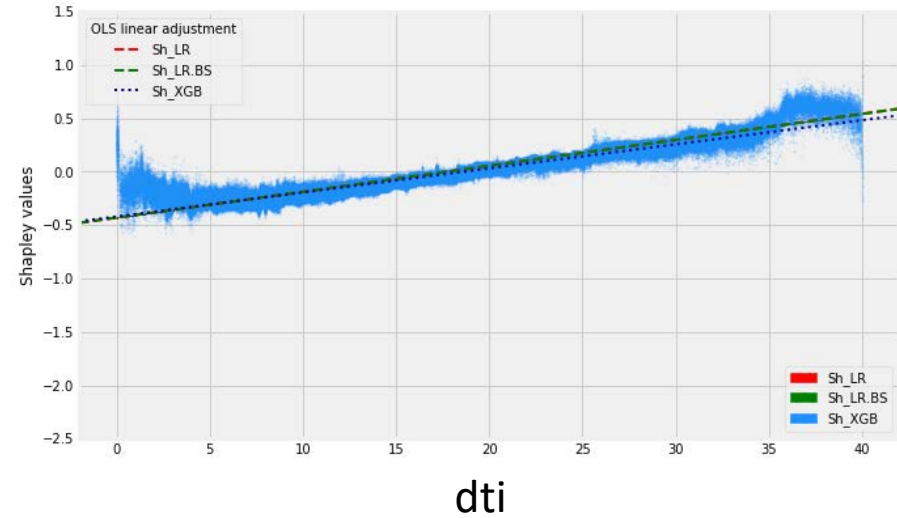
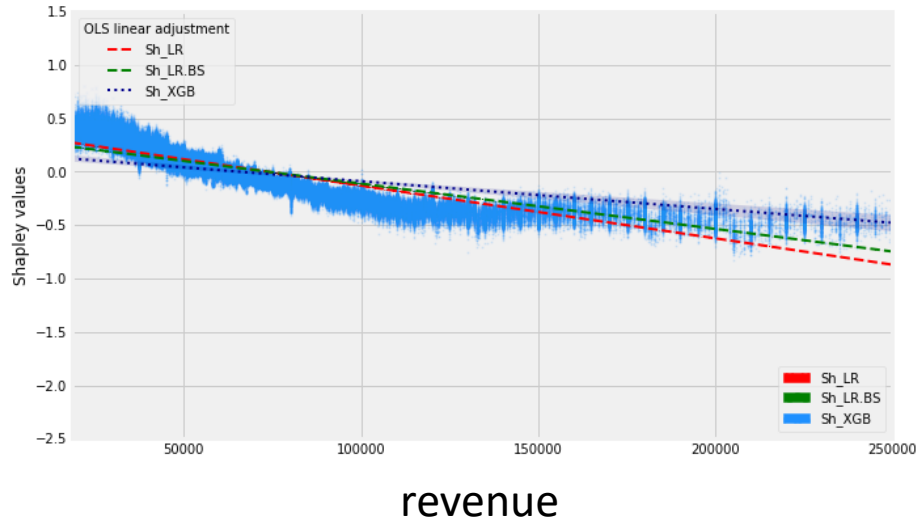


- Based on Shapley values and aggregating the categorical variables
- The order is similar for the three methods
- Quantitative variables are much more important than the categorical ones

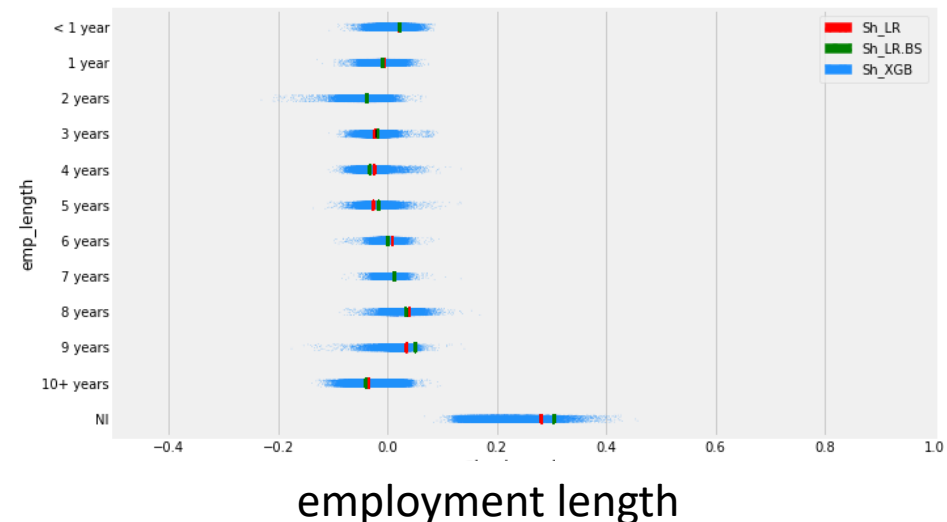
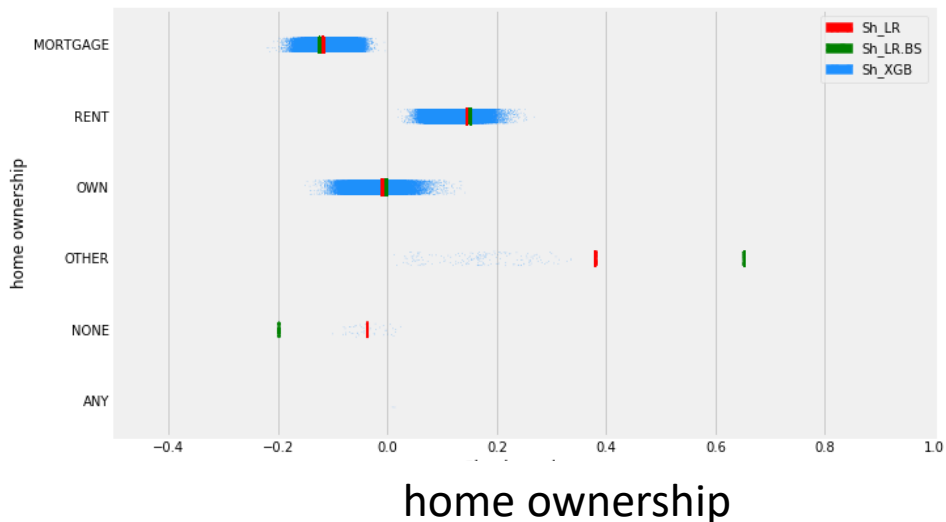
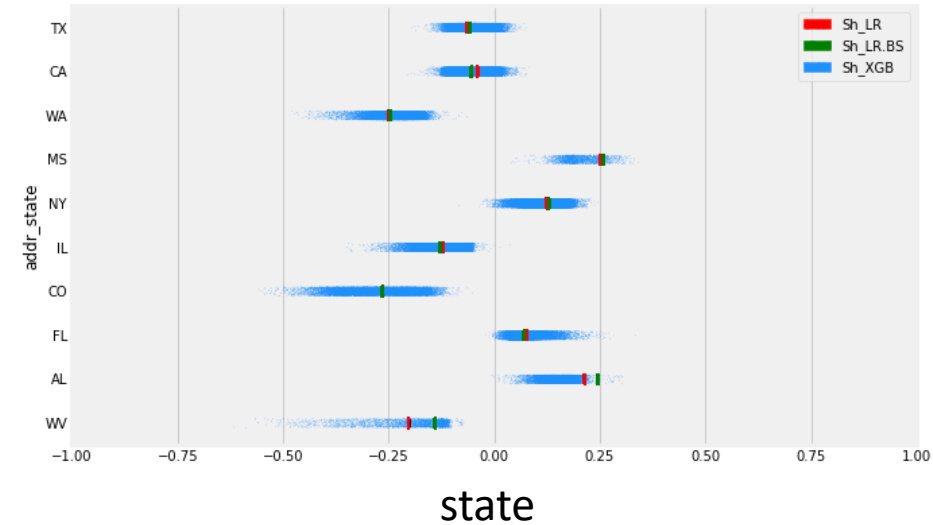
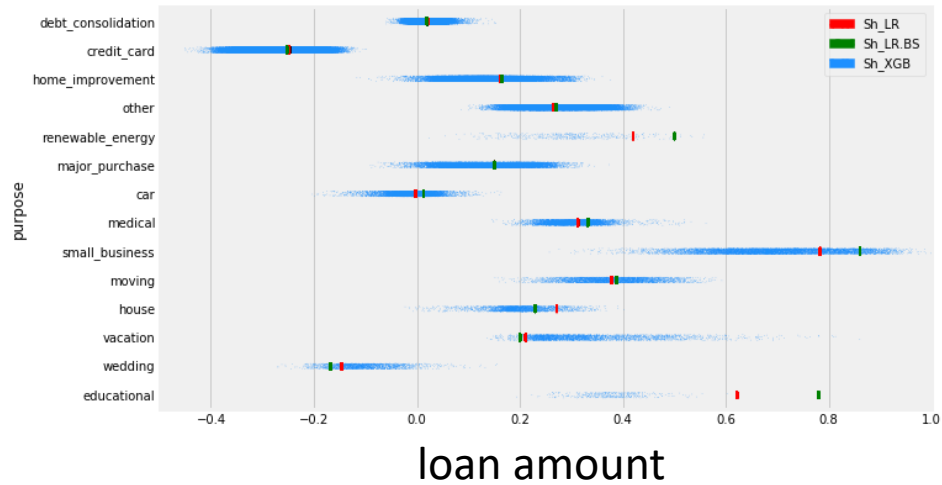
Model explanation: Feature importance



Quantitative variables



Categorical variables



- Machine learning approaches typically used in P2P lending credit risk models did not address interpretability.
- The good performance of the tree boosting classifier (XGBoost) is in line with other results in the literature
 - But our article shows that the **better results come from a better description of the relationships among the variables.**
 - Machine learning models can detect complex nonlinear relationships that cannot be reflected by logistic regression.
- We present a modeling alternative in credit risk to p2p market, reliable and explainable, as required by regulators and industry
 - Explainability generates trust in the model

Shapley values theory is still under development, for example, to better account for **dependence** and include other aspects, such **causality and inferential tools**.

This will **lead to a wider adoption** of machine learning models in credit risk modeling and other domains.

Thank you!

For those interested:

Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. (2020) *Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending*. **IEEE Access** 8, 64873–64890

<https://ieeexplore.ieee.org/document/9050779/>