MACHINE LEARNING EXPLAINABILITY IN LOAN TRANSPARENCY DETERMINATION

Emma Quyen Do, Kyle Spurlock, Olfa Nasraoui
Knowledge Discovery & Web Mining Lab, Department of Computer Science & Engineering, http://webmining.spd.louisville.edu/
NSF CSR REU, Summer 2024
University of Louisville, U.S.A., University of California-Berkeley, U.S.A

Introduction

Machine Learning models:
- White box: transparent algorithms
- Black box: high accuracy, less transparency on interpretability

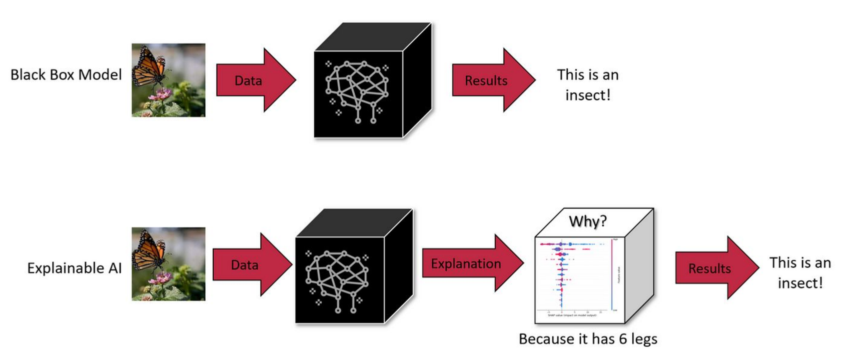

Figure 1: Explainability in AI framework. Source: https://encord.com/blog/model-robustness-machine-learning-strategies/

Explainable AI: produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)
- *Important* for transparency, trust, fairness, accountability, legal and ethical compliance.
- Model-agnostic methods: Applicable to any model, such as LIME (Local Interpretable Model-agnostic Explanations).

LIME:

- Model-agnostic.
- Approximating the complex model locally with a simpler, interpretable model.
- Understanding why a particular prediction was made by analyzing the local behavior around the instance of interest .

Companies oftentimes rely on highly accurate black box models to make a decision that requires highly complex computation.

However, much of it is not transparent as to how they came up with the final verdict.

Research Goals

Primary goal:
- Discover how to enhance the explainability of black box machine learning models.
- Conduct experiments to look inside a financial institution loan default dataset and try to gain an understanding of the features that the model thinks matter most when it comes up with a prediction.
- Create model-agnostic methods that:
  - Find local importance of feature changes for reaching the desired goal.
  - Predict if a borrower would have the ability to pay back the loan or not using a black box model called Random Forest.
  - Analyze the weight of each feature to the prediction to gain an understanding of how black box model produce a highly accurate prediction.
  - Evaluate if the most important feature weights is ethical and fair.

Methodology

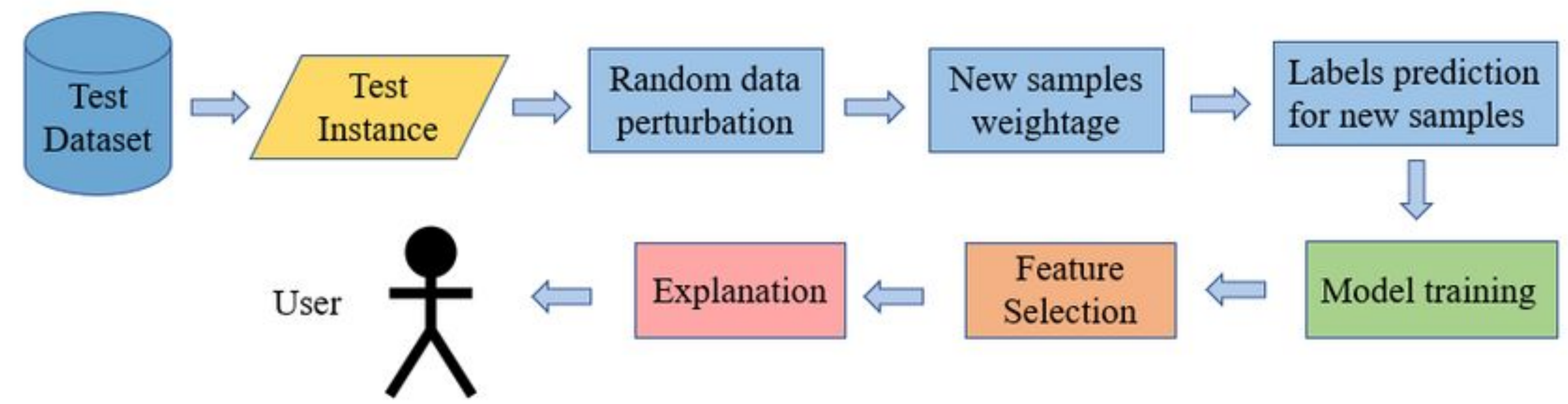

Figure 2: LIME framework within a black box model. Source: https://www.researchgate.net/figure/The-workflow-of-LIME-method_fig2_365101230

$$\xi(x) = \arg\min_{g \in G} \left(\mathcal{L}(f, g, \pi_x) + \Omega(g)\right)$$

- Assigns weights to the perturbed instances based on their similarity to the original instance.
- Determined using Euclidean distance, ensuring that data points closer to the original instance have a greater influence on the explanation.

- Generates synthetic data points by perturbations around the reference point of interest (to be explained) and uses those neighboring data points to fit a local linear regression model, such that the linear model weight of each feature can serve as explanation score for the reference point.
- LIME uses the original model to predict outcomes for these perturbed instances.
- This new training dataset is used to fit a *local* interpretable model, such as a linear model or decision tree.

Experiments & Evaluation

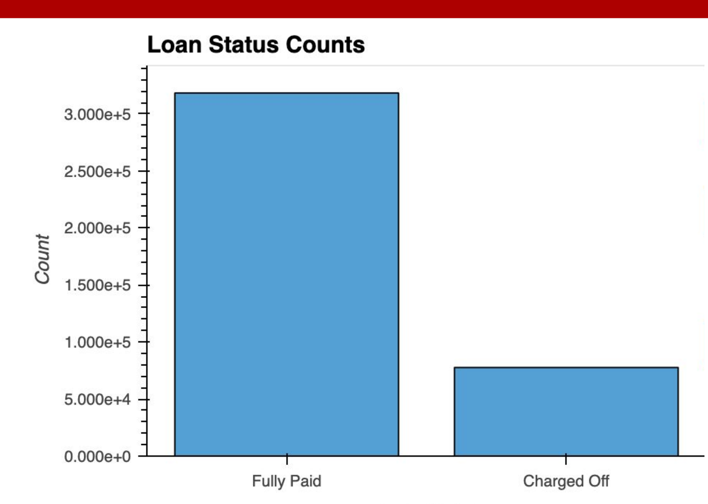

Figure 2: The count of fully paid and charged off borrowers.

- **Figure 2** shows the count of fully paid vs charged off loan status, showing that our dataset is imbalanced.
- The test set comprised 33% while the remaining 67% will be allocated to the training set with the Random Forest model.
- ROC shows the accurate and trade off between true positive rate with respect to each false positive rate.

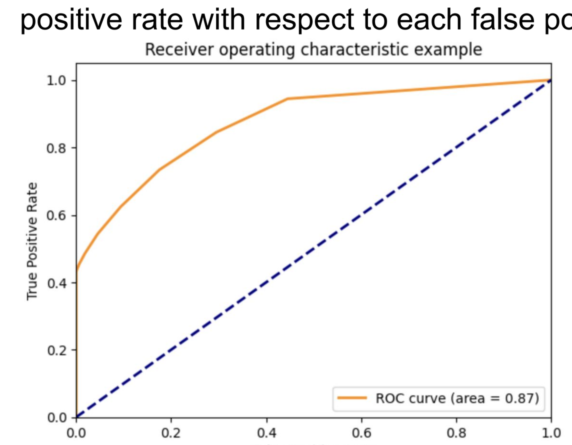

Figure 3: The ROC that shows performance of the model

- The area under the curve (AUC) in **Figure 3** shows that our model performs better than random chance and reaches higher accuracy when the false positive rate is relatively small compared to the true positive rate.
- **Figure 4** shows the confusion matrix which summarizes the predictions made by the model.

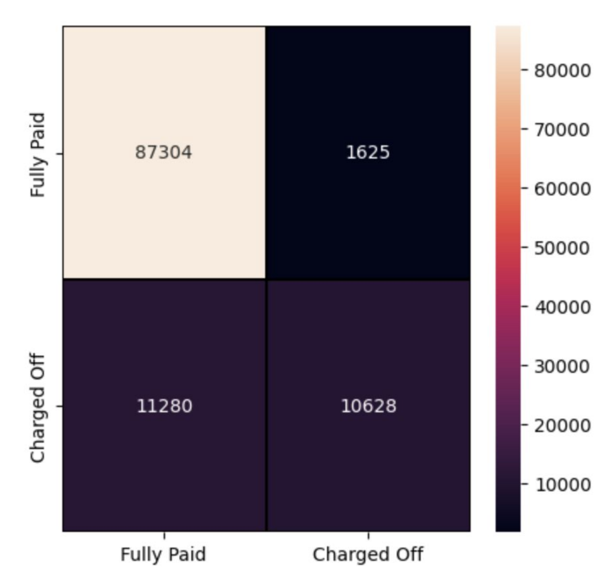

Figure 4: The confusion matrix of the model

- After we trained the model, select a point that has been classified by the model as class 1 (Charged Off) and apply LIME to get the feature weights of the decision shown in **Figure 5**.

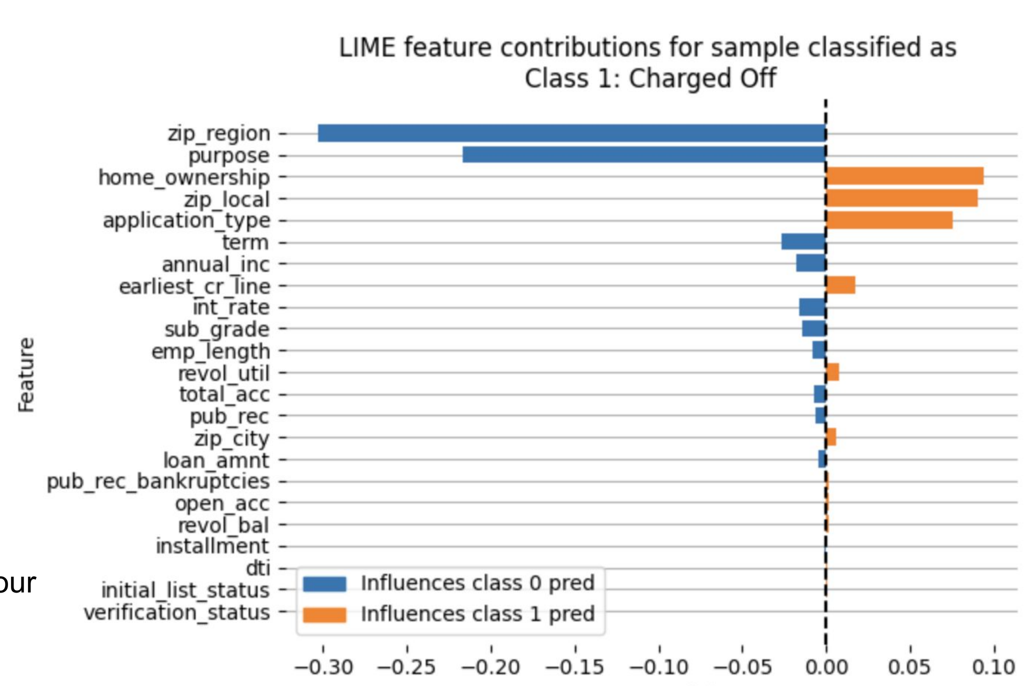

Figure 5: The weight of each features to the highly accurate prediction

- Even though the model was said to be highly accurate, after a look at the features weights, we can see that zip_region (state), purpose, home ownership and zip_local (city) are the most influential to this prediction.
- Raises a significant concern because the fact that certain applicants' living address (as inferred from their zip code) plays a crucial factor in rejecting (or accepting) their loan is a form of financial redlining (or privilege in the opposite/acceptance case).
- **Figure 6** shows values of each features in a charged off point that was predicted by the model ranking from most influential to least influential.

Sample: 368351, Actual: 1, Pred: 1

zip_region	9
zip_local	00
home_ownership	RENT
term	36
int_rate	9.76
sub_grade	7
annual_inc	75000.0
zip_city	37
revol_util	85.3
loan_amnt	8000.0
verification_status	Source Verified
earliest_cr_line	2006
emp_length	9.0
pub_rec	0.0
purpose	debt_consolidation
dti	19.42
installment	257.24
open_acc	10.0
revol_bal	20802.0
pub_rec_bankruptcies	0.0
total_acc	21.0
application_type	INDIVIDUAL
initial_list_status	w

Figure 6: The values of each features from most weight to least weight

Conclusions and Future Work

Conclusion:
- We applied a model-agnostic method onto a Random Forest Classifier (black box model)
- Although the black box model gives highly accurate results, once we applied LIME, we discovered hidden biases.
- Ensuring that models do not perpetuate or exacerbate existing biases is crucial for maintaining fairness and trust in automated decision-making systems
- The results highlight the necessity for transparency and accountability in AI models, particularly in finance, where decisions have substantial real-world impacts.

Limitation and Future Work:

Limitation:
- LIME is a post-hoc method
  - It learns a proxy explanation model to approximate a previously trained black box model, using data that were not part of the original analysis,
    - This means that explanations might not fully capture the underlying relationship in the original model.
- Might risk overinterpretation.

Future Work:
- Improve the model performance: train the model on XGBoost or some other black box models.
- Fairness Metrics:
- Calculate fairness metrics such as disparate impact, equal opportunity difference, and demographic parity to assess the fairness of the model's predictions.

Acknowledgement

This research was supported by the U.S. National Science Foundation (NSF) under grant CNS-2349076.

References

1. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D. and Zhu, J., 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8* (pp. 563-574). Springer International Publishing.
2. Assegie, T.A., 2023. Evaluation of local interpretable model-agnostic explanation and shapley additive explanation for chronic heart disease detection. *Proc Eng Technol Innov*, *23*, pp.48-59.
3. Kardys, I., Hoeks, S., van Domburg, R., Lenzen, M. and Boersma, E., 2013. Tools and techniques–statistics: analysis of continuous data using the t-test and ANOVA. *EuroIntervention J Eur Collab Work Group Interv Cardiol Eur Soc Cardiol*, *9*(6), pp.765-767.
4. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. and Kagal, L., 2018, October. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.
5. Cunningham, P. and Delany, S.J., 2021. K-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR)*, *54*(6), pp.1-25.
6. Lakshmi, J.V.N., 2016. Stochastic gradient descent using linear regression with python. *International Journal on Advanced Engineering Research and Applications*, *2*(7), pp.519-524.
7. Weng, L., 2017. How to Explain the Prediction of a Machine Learning Model?. *Blog Post, https://lilianweng. github. io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model. html*, pp.1-7.
8. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
9. Tamagnini, P., Krause, J., Dasgupta, A. and Bertini, E., 2017, May. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd workshop on human-in-the-loop data analytics* (pp. 1-6).