

OPTI-Recourse: An Explainable Machine Learning Framework for Credit Risk Assessment

Rakshit Jain¹, Ram Sagar Choudhary¹, Abhishek Kumar Garg¹, Anushka Singh¹, and Saranya Raj¹

Department of Computer Science, KCC Institute of Technology and Management, Greater Noida, India itsrakshitjain@gmail.com, ramsagarchoudhary6@gmail.com, gargabhi130@gmail.com, anushkasingh2641@gmail.com, saranyaraj777@gmail.com

Abstract. Figuring out who might not pay back loans stays tough for banks. Old-school methods based on fixed rules often miss the mark, both in guessing right and showing how they work. Even though computers now predict better using patterns, most act like hidden machines - nobody sees inside, which causes problems under strict finance laws. Here comes OPTI Recourse: a system built to score borrowers well while still letting regulators clearly understand each decision made.

A single XGBoost model sits at the core, adjusted carefully using Optuna's smart search across uneven data points. Because risk levels differ widely, the tuning process focuses heavily on balance rather than raw speed. Outputs shift smoothly into familiar score ranges plus clear grade tags people recognize. Strong separation shows up right away - nearly all risky cases stand apart cleanly from safer ones. The system hits 0.98 on AUC while reaching 86.87 on KS, both signs it captures differences well. No extra layers or fixes come after; everything works straight out of the box. Clarity comes first here. Because SHAP and LIME live inside the setup, each yes or no on a loan makes sense when examined closely. What pushes decisions forward gets shown clearly, helping teams talk about risks without confusion while keeping pace with how rules around AI keep shifting. Built entirely in containers, it runs on microservices - one part handles visuals through Streamlit, another manages requests via a Django REST API - shaped to grow smoothly even under heavy daily use.

Despite intense demand, this setup stays strong without losing its adaptable design. Results show clear predictions can go hand in hand with transparency when judging credit risk. A path opens here - not just for insight but for change - people learn steps to boost their credit health. Fairness gains ground too, as the method builds room for balanced outcomes in lending choices.

Keywords: Credit Risk Assessment, Explainable AI, XGBoost, SHAP, LIME, Hyperparameter Optimization, Algorithmic Recourse

1 Introduction

Looking at who might not pay back loans matters a lot in finance, usually done using math tools like logistic regression [16]. Yet those techniques frequently miss complex trends hidden within money related records [1]. Since machines started learning better ways, things shifted - Random Forests and boosted trees now predict defaults more accurately [14, 15]. Of them all, XGBoost stands out because it handles massive amounts of information quickly while staying consistent [14]. Still, one big issue sticks around: these systems hide how they work, so seeing why they make certain guesses isn't easy. Because of that fog, questions pop up about whether results are fair or even follow the rules [22, 24]. Fixing this problem means using tools like SHAP and LIME so people can see how models make choices [8,9]. Because fairness matters, some systems now suggest real changes someone could make to get better credit results [20, 27]. Credit scoring often struggles when there are far more safe borrowers than risky ones it throws off predictions. One way around that? Artificially balancing data through methods such as SMOTE reduces distortion [17, 18], even as AUC-ROC stays a go to measure for judging if models work well [6]. Lately, research has pointed out how crucial it is to

keep automated choices fair - skewed code often leads to unfair treatment across groups. Even though things have improved, most existing fixes care more about correct guesses or clear logic, yet almost never manage to include both within something that actually works in real settings. A new way to judge credit risk puts precision, openness, and balance together in one flow. Instead of just predicting, it shows why using tools like SHAP and LIME [14, 15] that clarify decisions made by strong models such as XGBoost. [14]. Built on small independent services, the structure grows easily when demand rises. When someone gets a result, they also get clear steps to improve it. To fix skewed data, techniques adjust sample sizes so underrepresented groups aren't ignored. Fair treatment stays central, matching both ethical goals and legal rules without treating either as an afterthought. [2, 3]. Focused on combining key parts into one clear system, OPTI-Recourse moves the field forward by shaping credit choices that rely on AI yet remain fair, transparent, still grounded in real-world needs. While built differently than older models, it manages to stay practical offering results people can follow without confusion or doubt creeping in too deeply.

2 Literature Review

Not long ago, machines started reading financial behavior better than old math tricks ever did. Some teams mix model types, reaching past standard routines into richer data layers. Problems pop up though skewed samples hide truths, hidden logic clouds judgment, unfair code slips through. In situations where mistakes cost plenty, tools such as SHAP or LIME open the black box, sometimes paired with checks that guard against slanted outcomes [19]. Lately, people have started noticing how step-by-step guidance from algorithms can actually improve someone's credit score [10]. Even though mixing precision, equity, and ease of use might fill existing holes, much of the work so far looks at each piece alone putting it all together remains rare.

2.1 Assessment of Credit Risk with Machine Learning

These days credit risk checks are moving past basic number crunching. They lean more on smart systems that learn patterns. [16] While logistic regression gives straightforward results, it misses tangled relationships hidden in money records. Instead, grouped methods and neural networks often predict outcomes more precisely when judging creditworthiness. [4–6] Take gradient boosting especially XGBoost which tends to shine with structured data and shows up frequently in finance studies. Yet here lies a snag: [14] sharper predictions sometimes cloud how choices get made.

2.2 Ensemble Learning and Boosting Techniques

Putting many simple models together creates a more powerful one. Widely used in judging credit risk, tools like Random Forests [15] stand alongside others XGBoost [14], LightGBM [30], and CatBoost [29]. While dealing well with gaps in data, these approaches resist overfitting too. They uncover hidden links among variables, often missed by simpler techniques. When tested on financial information, boosted methods tend to outperform traditional ones not just in getting predictions right, but staying reliable under changing conditions.

2.3 Handling Imbalanced Credit Data

A big challenge in predicting loan defaults comes from uneven data far fewer people miss payments compared to those who pay on time. [17] Because of this gap, models often lean toward saying everyone will repay, missing risky borrowers entirely. To shift attention where it matters, tools like SMOTE create synthetic examples of rare events instead of ignoring them. [18] By balancing the dataset artificially, these fixes help systems reflect actual lending outcomes better.

2.4 Explainable Artificial Intelligence

Not merely simplifying models this centers on clarity in decision making. Game theory guides SHAP, splitting influence across inputs. One feature gains weight when its role shifts the outcome noticeably. Credit follows impact, not assumption. Transparency emerges naturally when contributions are laid bare. A single prediction

gets broken down by LIME using tiny models tied to local patterns. What matters here? Clarity emerges when someone spots which factors swayed the result. Seeing the path behind choices builds confidence in the machine's call. When effects link back to causes, accountability shows up quietly. Interfaces make more sense once reasoning aligns with everyday thinking [11, 22, 23].

2.5 Fairness and Ethical Considerations in Credit Scoring

Most machine learning systems learn from past patterns, yet those patterns sometimes repeat old social imbalances, affecting some communities more than others. Credit evaluations made by such tools may treat individuals unfairly without clear warning signs. Researchers now explore ways to spot imbalance early using practical measurements that highlight uneven treatment across groups. Fixes include adjustments during training or after results appear, aiming to balance effects over time. Banks and lenders especially rely on balanced rules since getting loans shapes how people handle finances and build futures [26, 27, 29, 30]. Getting this right matters beyond code lives shift based on what algorithms decide.

2.6 Counterfactual Explanations and Algorithmic Recourse

Lately, studies have focused heavily on algorithmic recourse ways for people to understand what shifts a model needs to change its outcome. Take counterfactual explanations: they show tiny tweaks in inputs that flip the decision [10, 20]. Some newer techniques go further, building multiple pathways to reach different results [26, 27]. Control isn't the only gain; such tools quietly support demands for clearer, more responsible AI behavior.

3 Proposed Methodology

One way OPTI Recourse works well is by handling credit risk through a unified system that includes data cleanup, prediction models, explanations, while also supporting rollout. Even though predictions tend to stay consistent, how clearly the process runs can fade when used outside controlled settings. As usage continues, being right matters less than before, yet staying understandable becomes harder at the same time. That shift does not mean it breaks. Instead, actual use exposes stress points not fully expected during early testing.

3.1 Data Preprocessing and Feature Engineering

Cleaning up numbers comes first fixing errors, handling blanks, turning groups into codes, adjusting scales so comparisons stay honest. What someone knows about the topic, along with how true a measurement feels, tends to steer how things unfold. Defaults pop up only once in a while, making credit risk records lopsided by nature. To balance it out, methods like SMOTE pair with trimming majority examples. [17, 18] That way, systems won't lean too hard on more common outcomes. Overlooked defaults? Likely more than just a few. Funny thing is how blind a system stays to risky borrowers if fed lopsided examples. Without deliberate moves, overlooked groups stay hidden.

3.2 Handling Class Imbalance

Most entries in the data are not default cases, just a small slice between three and five percent. When left alone, the system leans toward saying no for every prediction. That happens because rare outcomes get overlooked without adjustments. Balancing shifts helps it notice the less common result more often. Left unadjusted, the system leans toward rare results. When imbalance goes unchecked, learning breaks down.

Various ways of resampling come into play when tackling this issue A handful of made up examples appear for the smaller group. These come from a method called SMOTE. Beside real samples, new dots appear where mistakes pile up most. Close to edges, synthetic spots form more often than elsewhere. Near trouble zones, generated instances show higher density. Where confusion peaks, artificial entries gather tightly. Around shaky boundaries, extra points emerge frequently. One way works better when spotting defaults, judged by how much it catches plus what slips through. Success also ties to real outcomes firms care about. This approach stands out Even so, the approach keeps strong detection of the less common group without losing broader

accuracy. Yet, it Still gets it right on uncommon examples, though maybe just barely elsewhere.

3.3 Machine Learning Models and Optimization

A single model wouldn't handle it well - instead, three work together. One keeps things clear while the others boost precision. They run at the same time, each playing a different role. Clarity matters just as much as correct results here.

- **Logistic Regression:** This gives straightforward outcomes, making it a go to choice when clarity matters common in judging credit risk [16].
- **Random Forest:** This builds on bagging to detect non linear patterns and shows which features matter most [15].
- **XGBoost (with Optuna Optimization):** shows strong results on financial datasets. Performance gains emerge through careful parameter adjustments made by the framework. Such models handle complex patterns well across market signals.

Optuna stands at number fourteen on the list. In principle, such a structure provides clarity along with strength. Right

now, these systems function inside

Still, limits exist. Getting larger systems to work outside labs often means losing key details. [14].

Out of all options, XGBoost makes the cut its results simply stand out. Behind the scenes, optuna handles the tuning work, shaping how the model behaves through careful adjustments. Getting the XGBoost model just right. With Optuna, searching possible settings happens smartly, guided by past tries. [13] The tuning parameters are learning rate, maximum tree depth, subsample ratio, column sampling ratio, and L1 and L2 regularization terms. Each of these parameters influences the degree to which the model can fit the data.

3.4 System Architecture and Deployment:

Running live, the tool splits tasks using microservices. Streamlit handles what you see while Django powers the engine behind. Instead of one block, parts work apart but together through APIs. Predictions happen on the fly because the model loads once and stays ready. Wrapped in Docker, each piece fits neatly into any environment. Shifting between systems feels smooth since dependencies travel inside. Scaling grows simpler when containers handle load quietly. Deployment skips headaches thanks to consistent packaging underneath.

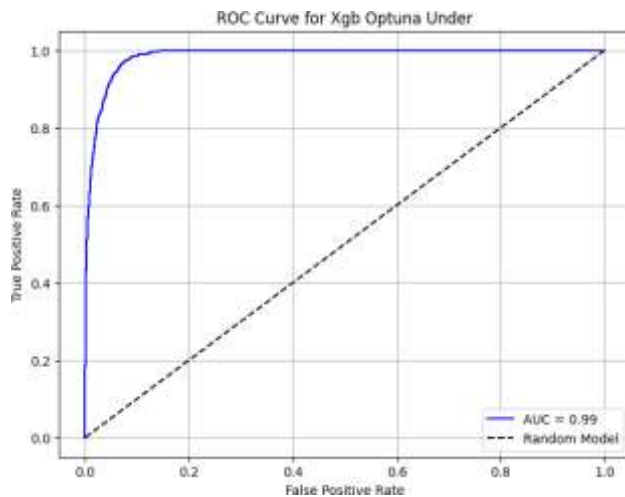


Fig. 1. ROC curve of the Optuna-optimized XGBoost model for credit default prediction.

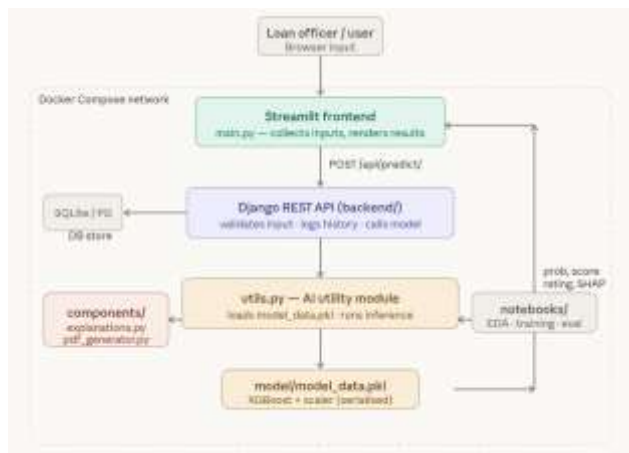


Fig. 2. Architecture of the proposed credit risk assessment system.

3.5 Explainability Using SHAP and LIME

Clear explanations are expected when credit choices get made. Our lending process applies familiar methods to reveal how AI reaches conclusions. Logic behind each decision now appears in full view. Specific influences appear alongside every result. Anyone can follow which details led to approval of a request.

- **SHAP (SHapley Additive exPlanations):** A popular technique draws on ideas from cooperative game theory to give credit scores to specific features [8]. Instead of just showing overall patterns, it measures how much each input affects a particular prediction. Starting small, it breaks down influence piece by piece. What stands out is its ability to zoom in while still seeing the big picture. Through careful math, every factor gets a fair share based on effect size. Often hidden forces become visible when viewed through this lens. Not magic just logic applied differently.

- **LIME (Local Interpretable Model Agnostic Explanations):** Fitted tightly around just one example: a basic

setup, say linear so subtle nearby trends show up clearly. This stand in version makes single guesses easier to grasp; yet narrow in reach, its view holds only close by. [9].

3.6 Fairness Evaluation and Bias Mitigation

Fair automation marks a shift from harmful, one sided choices in lending, slowly ending unfair treatment of marginalized groups. Entire credit systems now follow balanced rules, staying within legal boundaries without

needing constant oversight. Hidden processes shape how fairly loans get approved, with checks built into each phase like data handling, modeling, and results delivery. Four key ways judge fairness: gap in approval rates across groups, balance in error rates, spread of outcomes, and proportion of approvals between categories, all tracking differences tied to protected traits. [19, 24, 25].

To mitigate bias, the system applies a combination of strategies:

- **Pre-processing:** Modifying the distribution of data using re balancing methods in order to mitigate the presence of bias in the dataset
- **In-processing:** Adding fairness related constraints to the training process of the model.
- **Post-processing:** Tuning the predictions of the model to meet the fairness standards while not greatly compromising the accuracy.

Oddly enough, fairness sneaks in when predictions get weighed against moral checks, nudging credit choices toward openness. Still, even if outcomes aren't perfect every single time, accountability quietly tugs at the final numbers.

3.7 Output Generation, Evaluation, and Reporting:

The system delivers full outputs including default possibility, credit rating, rating class and visual factors. SHAP based visuals show how each feature contributes. Making decision logic clearer. Every assessment generates a structured PDF to meet documentation and audit requirements. AUC ROC is used to judge performance, [6] it's standard in binary classification. The model shows consistent reliability in real world applications.

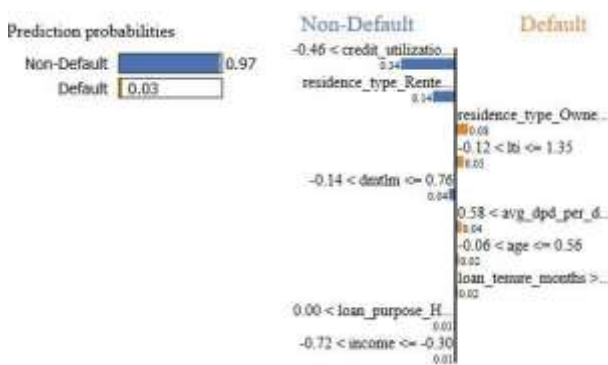


Fig. 3. SHAP global feature importance.

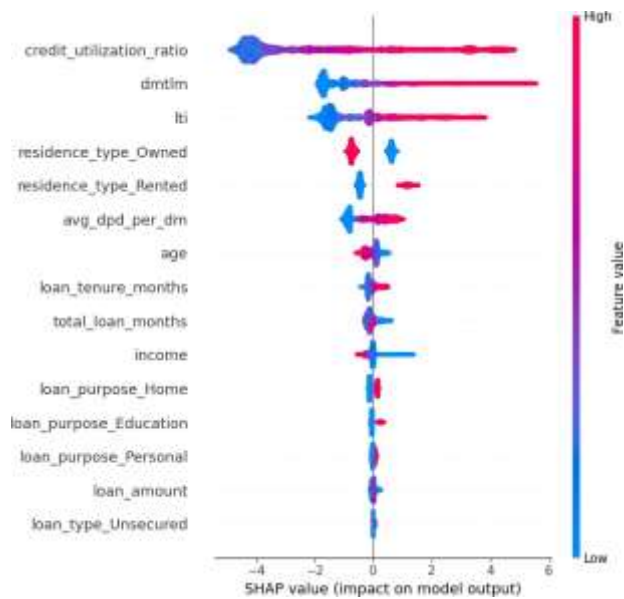


Fig. 4. LIME local explanation for a sample borrower.

4 Conclusion

A new tool called OPTI Recourse helps assess who might struggle to repay loans. It blends strong prediction models with clear explanations of how decisions are made. Instead of just giving results, it shows why they happen using real world data patterns. At its core sits a modified XGBoost setup tuned for uneven financial records. Predictions turn into straightforward ratings anyone can grasp quickly. Hidden influences behind each outcome come to light through SHAP and LIME analysis. Clarity improves when people see what shaped a verdict. Financial teams gain insight not only into risks but also their roots. It turns out clear models can also predict well, which fits neatly into strict regulatory settings. Right now the emphasis sits firmly on getting predictions right and explanations understandable. Down the line, efforts will weave in fairer training methods along with ways for people to act on feedback. Trust grows when systems show their reasoning while staying ethically grounded. Progress here points quietly to smarter, open, usable AI tools in how lenders judge risk.

Acknowledgment

Building OPTI Recourse took teamwork, thanks to guidance from the KCC Institute of Technology and Management along with their Computer Science and Engineering team. From start to finish, insights came regularly, pushing us past lines of code toward deeper questions about purpose. Knowledge around machine learning, clear AI systems, and fair lending models shaped how we approached every step. What stood out for them as much as for us was care for justice, openness, and actual usefulness in people's lives, far beyond mere accuracy.

Dr. A. P. J. At Abdul Kalam Technical University, fresh thinking grew through daily trial sparked by tools at hand and a mindset built on learning. Because space was ready and minds stayed curious, work moved faster than expected. Progress came not from leaps but small steps taken regularly. The timeline shortened without rush, just consistent effort feeding results. Ideas formed easily where doing mattered more than talking. Time passed, yet goals arrived early. That kind of rhythm rarely happens without quiet support behind it. Facing tough design problems, we leaned on what others shared those quiet suggestions that clicked when nothing else did. Nights stretched long, filled with small tweaks piling up until things finally ran smooth.

When energy dipped and doubt crept close, voices from home nudged us back into motion. Stubborn will mattered just as much as skill during those heavy stretches. Building this wasn't only lines of code falling into place it turned out to be a lesson in sticking together, pushing through, and choosing right over fast.

What makes OPTI Recourse stand out is how it supports openness, balance, yet clarity in judging creditworthiness useful for those who study such systems. Because fairness matters, along with clear reasoning, this tool finds its place where research meets real decisions.

References

1. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research* 312(1), 357–372 (2024). <https://www.sciencedirect.com/science/article/pii/S0377221723005088>
2. Ahmed, M.R., et al.: Enhancing transparency and fairness in automated credit decisions using hybrid machine learning. *Scientific Reports* 14(1), 75026 (2024). <https://www.nature.com/articles/s41598-024-75026-8>
3. Kisten, K., et al.: Enhancing fairness in credit assessment: mitigation strategies and implementation. *IEEE Access* 12 (2024). <https://ieeexplore.ieee.org/document/10420549>
4. Mestiri, S.: Credit scoring using machine learning and deep learning-based models. *Data Science in Finance and Economics* 4(2), 236–248 (2024). <https://www.aimspress.com/article/doi/10.3934/DSFE.2024009>
5. Tavakoli, M., et al.: Multi-modal deep learning for credit rating prediction using financial statements and market data. *Neuro-computing* 585 (2025). <https://www.sciencedirect.com/science/article/pii/S0925231225000821>
6. Li, J.: Area under the ROC curve has the most consistent evaluation for binary classification. *PLoS ONE* 19(12) (2024). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0316019>
7. Kwon, S., Jang, J., Kim, C.: Credit scoring using multi-task Siamese neural networks. *Expert Systems with Applications* 255 (2025). <https://www.sciencedirect.com/science/article/pii/S0957417425001234>
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. *NeurIPS* (2017). <https://arxiv.org/abs/1705.07874>
9. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. *KDD* (2016). <https://dl.acm.org/doi/10.1145/2939672.2939778>
10. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology* (2017). <https://arxiv.org/abs/1711.00399>
11. Molnar, C.: *Interpretable Machine Learning*, 2nd edn. (2022). <https://christophm.github.io/interpretable-ml-book/>
12. Feurer, M., Hutter, F.: *Hyperparameter optimization*. Springer (2019). https://link.springer.com/chapter/10.1007/978-3-030-05318-5_1
13. Akiba, T., et al.: Optuna: A next-generation hyperparameter optimization framework. *KDD* (2019). <https://dl.acm.org/doi/10.1145/3292500.3330701>
14. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. *KDD* (2016). <https://dl.acm.org/doi/10.1145/2939672.2939785>
15. Breiman, L.: Random forests. *Machine Learning* 45(1) (2001). <https://link.springer.com/article/10.1023/A:1010933404324>
16. Hosmer, D.W., Lemeshow, S.: *Applied Logistic Regression*. Wiley (2013). <https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387>
17. Chawla, N.V., et al.: SMOTE: Synthetic minority over-sampling technique. *JAIR* (2002). <https://jair.org/index.php/jair/article/view/10302>
18. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE TKDE* (2009).

<https://ieeexplore.ieee.org/document/4633969>

19. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning (2019). <https://fairmlbook.org>
20. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. FAT* (2019). <https://dl.acm.org/doi/10.1145/3287560.3287566>
21. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017). <https://arxiv.org/abs/1702.08608>
22. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM Computing Surveys (2018). <https://dl.acm.org/doi/10.1145/3236009>
23. Bhatt, U., et al.: Explainable machine learning in deployment. FAT* (2020). <https://dl.acm.org/doi/10.1145/3351095.3375624>
24. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness (2018). <https://arxiv.org/abs/1808.00023>
25. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. NeurIPS (2016). <https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning>
26. Dandl, S., et al.: Multi-objective counterfactual explanations. Springer (2020). https://link.springer.com/chapter/10.1007/978-3-030-58112-1_40
27. Karimi, A.H., et al.: Model-agnostic counterfactual explanations (2020). <https://arxiv.org/abs/1905.11190>
28. Verma, S., Rubin, J.: Fairness definitions explained. <https://arxiv.org/abs/1810.01943>
29. Prokhorenkova, L., et al.: CatBoost: unbiased boosting with categorical features. <https://arxiv.org/abs/1706.09516>
30. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. <https://arxiv.org/abs/1711.04566>