Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

Khalid and Javed

An Hybrid Algorithmic Framework for Credit Risk Assessment Based on Ensemble Learning Techniques

[1]Dr. Usman Khalid

[2] Hira Javed

How to cite this article:

Khalid, U., & Javed, H. (2020). An Hybrid Algorithmic Framework for Credit Risk Assessment Based on Ensemble Learning Techniques. *Journal of Technological Information, Management & Engineering Sciences, 1(2), 20-35.*

Received	Accepted	Published
20 July 2020	24 August 2020	28 November 2020

Abstract

Credit risk assessment plays a crucial role in the functioning of lending institutions, as it impacts loan approval decisions, pricing, capital structure, portfolio management, and financial stability. While traditional score cards and single predictive models can deliver clear and effective decisions that are also operationally efficient, they can be difficult to build to capture the non-linear relationships, the complex interactions, evolving borrower patterns and the data heterogeneity. It suggests a hybrid algorithmic framework which integrates data preparation, interpretable baseline modeling, ensemble learning, probability calibration, optimization of decision-threshold and post-deployment monitoring. The framework adopts logistic regression as a transparent benchmark and ensembling the same through validation-based ensemble method, by combining tree-based learners such as random forests and gradient-boosted decision trees. It also includes methods for handling class imbalance, temporal validation, explainability, fairness checks and governance controls. The paper does not make any claims of empirical performance without data, but rather provides a research design and evaluation protocol to be used to test the framework with borrower-level repayment outcomes. Some key metrics for core evaluation are: area under the receiver operating characteristic curve (ROC-AUC), precision-recall AUC, Brier score, calibration curves, expected credit loss, and subgroup performance. The paper proposes a new evaluation method of the usefulness of a hybrid system beyond accuracy ranking that takes into account calibration, stability, interpretability, fairness and decision utility. The proposed approach provides a well defined structure for building credit-risk models that are technically sound and appropriate to responsible use in the lending business and subject to local regulatory and institution-specific model validation.

Keywords: *credit risk, ensemble learning, logistic regression, random forest, gradient boosting, model calibration, explainable artificial intelligence, responsible lending.*

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)

This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

Khalid and Javed

Introduction

Credit risk refers to the likelihood of a borrower's failure to pay back the promised financial obligations. For the lenders, this risk estimation helps them underwrite, price loans, assign credit limits, provision, collect, and manage their portfolio. A wrong estimate may result in two expensive mistakes: risky applicants whose failure to pay would be too high, or the failure to make a loan to a creditworthy applicant, and the loss of a profitable business. When conditions are changing rapidly, customers are changing, lending products are changing and data sources are changing, reliability of risk estimates becomes even more important.

Traditional credit rating models have been based on statistical models like logistic regression and scorecards. Such approaches have appeal because their score contributions and coefficients can be communicated to analysts and applicants and because they can be accomplished with relatively small computers. However, relationships between borrower characteristics and default may be nonlinear. These relationships can be complicated by the interaction between debt burden, level of income stability, utilization, the repayment history and macroeconomic factors and can be hard to capture in a simple additive model without careful design.

There are flexible alternatives to machine-learning methods. Random forests combine several decision trees and are able to capture nonlinear effects and interactions. Gradient-boosted decision trees are sequential learners that improve previous mistakes, and can be very effective on structured tabular data. However, it can be hard to interpret a complex model, it can be poorly calibrated, and it can generate historical inequities in the data it models. High discrimination does not necessarily mean high responsibility or usefulness for a credit decision system.

The authors suggest a hybrid algorithmic framework which merges the advantages of transparent statistical modeling and ensemble learning. In a general sense, the framework is hybrid since it incorporates several predictive learners, combination of validators, calibration, cost-sensitive threshold selection, explainability, fairness evaluation, and monitoring. The goal is not to make the claim that a particular ensemble will always perform better than a simpler

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

ensemble, but to create a repeatable method to determine when and how it is helpful.

1.1 Research Objectives

The study aims to address five objectives: (1) develop a framework for borrower default risk estimation that is modular in structure; (2) provide a defensible comparison between logistic regression and ensemble learners; (3) demonstrate how to incorporate interpretability, calibration and fairness into model selection; and (4) propose operational governance for deployment and monitoring.

1.2 Research Questions

The paper addresses the following questions: How can ensemble learning be combined with interpretable credit-risk modeling? What other measures of evaluation are more informative than accuracy? How to choose a threshold when there are different costs of false approval and false rejection? Lastly, what are the controls that must exist to ensure that the model's behavior is reliable and fair following deployment?

2. Literature Review

The statistical classification of credit has been around for a long time. Logistic regression is still popular, since it estimates a binary probability and yields understandable coefficients, which can be analysed for direction and size. A typical application would be to determine whether a loan would default on a given performance window. The model predicts the log-odds of default based on variables observed from the borrower and loan. A linear predictor in log odds space is limited in flexibility, but transformations and bins, as well as interactions, can add flexibility while maintaining a fairly interpretable structure.

The idea of ensemble learning is to combine several models to make the system robust or predictive power better. Bagging trains models on the resampled data, and averages their predictions, while Random Forests introduce random feature selection to decorrelate trees. Boosting trains a sequence of weak learners, allocating more weights to those observations that are still hard to predict. Gradient boosting (Friedman, 2001; Breiman, 2001) and modern implementations like XGBoost (Chen & Guestrin, 2016) and LightGBM and CatBoost (Chen et al.,

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

2019) have become popular choices for tabular prediction due to their ability to work with limited manual feature engineering (Breiman, 2001; Friedman, 2001; Chen & Guestrin, 2016).

The quality of the data and validation design influence ensemble performance. Random Train Test Split could result in an inflated performance if the future is different from the past or if records from the same customer are located across partitions. Out-of-time evaluation is a benefit for credit-risk studies since a model is built using past observations and then tested on future outcomes. The outcome horizon should also be carefully determined: If default is considered over the next 12 months, then the feature snapshot should reflect only information available at the decision date, and observations whose outcomes have not been reached should be ignored as non-defaults.

The concept of predictive ranking is not the same as probability quality. The ROC-AUC measure reflects the percentage of times that a randomly selected defaulting borrower has a higher risk score than a non-defaulting borrower. In cases of low default rates, precision-recall AUC may be more informative. Calibration checks the consistency between the probabilities predicted and the frequencies of the events that actually occur. A model can be very accurate in rank order borrowers but systematically overestimate or underestimate their default probabilities and thus can mislead pricing and expected loss calculations. The Brier score and log loss are correct scoring rules which reward inaccurate probabilities (Brier, 1950). When applied to an appropriate validation set, calibration techniques like Platt scaling or isotonic regression can enhance probability estimates.

Explainability and governance become key in a credit decision as it impacts financial services access. Global feature importance can be used to summarize the model behavior while local explanation methods can explain the factors that lead to a specific prediction. For instance, in the cooperative-game-theoretic approach of SHAP, feature contributions are computed based on Shapley values (Lundberg & Lee, 2017). Explanations should not be given as causal evidence and should not be presented as proof that changing a feature will change the risk of a borrower.

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

They must be evaluated for stability, reasonableness, and adherence to adverse-action and consumer protection requirements.

Careful definition is needed to achieve fairness in credit modeling. The differences in the approval rates or error rates between groups may be due to historical access, label quality, economic factors and feature distributions. In itself, the exclusion of protected attributes in the model is not sufficient to ensure fairness, since there might be correlated variables serving as proxies. Meanwhile, the fairness measures might clash, especially if the group base rates are different. Institutions should therefore record their selected fairness goals, legal restrictions, data limitations and trade-offs instead of a single measure. The NIST AI Risk Management Framework (NIST, 2023) is a model governance framework which prioritizes principles of validity, transparency, accountability, and continuous risk management.

3. Proposed Hybrid Framework

The proposed framework is designed as a series of linked modules. The input, output and control of each module are defined, allowing the quality of the model to be monitored from the raw data to the lending decision. Separation of prediction and policy is intentional: Prediction is done by the model, and action is based on business cost, regulatory rule, and risk appetite.

3.1 Data Acquisition and Target Definition

Sources of data might include application data, verified income/employment data, current debt load, repayment history, credit utilization, loan terms, and acceptable macroeconomic data. Alternative data must be used only when lawful, relevant, sufficiently reliable and assessed for disparate impact. Document data provenance, consent, or other legal basis; access permissions; retention periods; and update frequency.

Before models are developed the target variable has to be defined. A binary target, for instance, can be set to 1 if a borrower becomes delinquent or defaults within 12 months after loan origination and 0 if the borrower does not become delinquent or default during the period. The definition must reflect the lender's products, policies and regulations. Restructured loans, early settlements, missing

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

outcomes, and loans that are sold/transferred must be treated explicitly to prevent any ambiguity in the labels.

Understanding data quality and feature engineering.3.2 Data Quality and Feature Engineering.

The data pipeline should validate for duplicates, impossible values, incompatible units, missing values, outlier detection and field definition inconsistencies. Values that are missing can be imputed with training-set statistics or model-native methods, but the imputation must be learned, using the statistics or model, solely from the training set. One-hot encoding, well-thought-out ordinal encodings, and target-leakage free encodings are all possible methods for encoding categorical features. Numerical variables can be converted to create meaningful bands of risk and/or to decrease skew. All transformations shall be repeatable, and shall be version controlled.

Decisions time boundary should be respected in feature engineering. For instance, a repayment-history feature might provide a summary of repayment history prior to the application, but may not contain payments made after the date the decision was made. Underwriting models should not include identifiers that are indirectly used to identify the target, or automatic collection actions, or variables that are only obtained when a loan goes bad. If more than one record is being partitioned for the same borrower, ensure there isn't any cross split contamination.

Data Analysis: Logistic Regression

The default probability is estimated by logistic regression in the following way:

$$P(Y = 1 \mid X) = 1 / (1 + \exp(-(\beta_0 + \beta^T X)))$$

In this case, Y is the indicator, X is the feature vector, and β is the model coefficients, β_0 is the intercept. Overfitting can be alleviated by regularization when there are lots of correlated predictors. The baseline serves as a clear reference point, and as a helpful diagnostic—an ensemble is justified to a degree if it offers measurable improvements in validation performance, stability, or decision usefulness.

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

3.4 Ensemble Learners

The framework assesses at least two of the tree-based learners. A random forest is an average of many decorrelated trees, which have less variance than an individual tree. The gradient-boosted trees construct an additive model that involves each new tree serving to enhance the existing objective. Hyperparameters such as tree depth, learning rate, the number of trees, minimum leaf size and row subsampling, regularization should be optimized within cross validation that is performed, only, during the training period.

The predicted probabilities from baseline and tree-based models can be incorporated into a soft-voting ensemble. Suppose that $p_1(x)$, $p_2(x)$, …, $p_m(x)$ denote the component probabilities. The probability of both of them occurring is:

$p_hybrid(x) = \Sigma_i w_i p_i(x)$, where $w_i \geq 0$ and $\Sigma_i w_i = 1$.

Selection of weights can be based on a validation measure such as log loss or brier score and constraints can be used to prevent overfitting. An alternative approach is a stacked model: The predictions of the base learners are fed to a simple meta-learner. The stacking predictions should be predictions that are made outside the test set, or the base predictions would be in-sample predictions, which would lead to leakage and biased estimates.

3.5 Calibration and Decision Layer

Once the model has been selected, probability calibration is assessed using data not used in the base model. Platt scaling is a logistic mapping of model scores, whilst isotonic regression learns the mapping without any parametric assumption. Overall and, if possible by sample size, relevant sections of the portfolio should be tested for calibration. If the calibration is not good, the model should not be used as the probability of the output.

A decision threshold converts a probability to an action. If predicted risk falls below a threshold, the application is approved or referred for approval; however, one threshold may not be appropriate for the various factors associated with the loan, such as the interest margin, size of the loan, loss given default, exposure at default and affordability requirements. Threshold selection should be done in light of expected costs and benefits, taking into account law, policy and risk

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

appetite. The model should not apply thresholds based on protected characteristics.

3.6 Explanability, Fairness and Governance

The framework generates global explanations for review of the portfolio and local explanations of individual cases. The reason code mapping, partial dependence, and feature importance can aid in model review, but they need to be validated by domain experts. Reason codes need to reflect the model and be comprehensible for the target group. Private information about other borrowers should not be disclosed via explanation tools and sensitive variables and proxies must be legally reviewed.

Fairness testing should be considered in conjunction with compliance, legal, risk and impacted stakeholders. Some potential checks might be different approval rates, false-positive and false-negative rates, group-wise calibration, or stability in relevant groups. Confidence intervals and minimum sample rules should be reported when a small sample size causes the estimates of subgroups to be unreliable. When lawful information regarding protected groups is not available, the institution should announce such a disclaimer; it is not fair to claim fairness has been established.

3.7 Monitoring and Feedback

Population stability, drift, predicted-risk distributions, default rates, approval results, complaints, overrides, subgroup indicators should be monitored after deploy.Monitoring after deploy should include population stability, drift, predicted-risk distributions, default rates, approval results, complaints, overrides, and subgroup indicators. Investigation should be triggered by the alert, NOT automatic retraining. Retraining should follow a documented change control process, be reproducible, be approved by authorized model-risk personnel, and should include a rollback procedure. There should be an ability to review cases at the human level when applicable institutional policy and law deems that there is a need, as is the case for other cases, for such review.

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

4. Research Methodology

In this paper, a quantitative and comparative evaluation design is proposed. Since no data for the borrowers were provided, nor were any experimental results provided, the framework in this paper is not an empirical study but is instead a research proposal. There are no performance claims. The design below may be used with an institutionally approved data set or an appropriate public benchmark (depending on the license, documentation and restrictions of the data source).

4.1 Dataset and Sampling Strategy

The unit of analysis is either a credit application or an originated loan, depending on the target definition. The data set should include a decision time, the specified performance time window, outcome status and features available at the decision moment. Samples should be representative of the population to be deployed. The researcher should report on selection effects, applicants with unobserved outcomes, changes in underwriting policy and macroeconomic periods covered in the dataset if it is a historical portfolio.

4.2 Train, Validation, and Test Design

If timestamps exist, it is recommended to use a temporal split. The oldest time interval is used for training the model, a middle time interval for tuning and calibration, and the most recent stable time interval for final testing. For instance, in the analysis, a multi-year training period could be used, followed by a multi-year validation period, and then the final 12-months set with complete data — how these periods are calculated should depend on the amount of data and the business context. Alternatively, stratified cross validation may be applied if it is not possible to separate time periods, but it is recognized that it may not be suitable for future period generalization.

Hyperparameter tuning should be done within the training/validation process, not on final test set. All preprocessing, resampling, feature selection and calibration should be done without the knowledge of the final test results. Oversampling methods (e.g., SMOTE) might be explored only in the training folds, while class weighting is an often easier first choice. Evaluation should maintain the natural

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

distribution of the default in validation and test sets to ensure the precision, calibration, and losses reported are meaningful.

4.3 Evaluation Metrics

There is no single measure that is enough. Ranking Discrimination (ROC-AUC) is used for the default class when events are common, Precision-Recall AUC is used for the default class when events are rare. The Brier score and log loss are measures of the quality of probabilities. Calibration plots are used to compare event rates observed with those predicted. When operating at a threshold, the confusion matrix provides sensitivity, specificity, precision, and balanced accuracy. Evaluating the business should include a worst-case scenario analysis of the costs of failing to approve a transaction and the costs of mistaking a transaction for a false rejection, while estimating the expected credit loss.

A standard approximation of ECL is ECL = PD × LGD × EAD, with PD representing the probability of default, LGD the loss given default and EAD the exposure at default. This expression is a conceptualized version that has been simplified; in accounting and/or regulatory calculations, it may be necessary to discount, accommodate scenarios, use term structures and apply specific rules for various jurisdictions. Economic value of a model should be measured using realistic assumptions of recovery, funding cost, capital, and cost of lost business.

(4.4) Statistical Uncertainty & Robustness

If the difference in performance is significant, it should be followed by a confidence interval or repeated temporal assessments, if possible. If relevant dependence is respected, uncertainty in AUC, Brier score and other metrics can be estimated by bootstrap resampling. Performance should be evaluated by product, geographic, vintage, and risk segment (if available and statistically justified). Stress tests can determine the likely impact of a range of scenarios, such as income, unemployment, interest rates, or repayment, but should not be interpreted as a forecast.

5. Proposed Algorithm

The end-to-end algorithm aims to reduce leakage and ensure that the decision policy differs from probability estimation:

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Step 1: Establish the event objective, performance time period, decision date, allowable data, and purpose.

Step 2: Generate a time-aware dataset and perform data quality, missingness, label maturity, and feature availability audits.

Step 3: Divide data into training, validation and final test sets chronologically.

Step 4: Fit preprocessing and a regularized logistic-regression baseline on training data.

Step 5: Tune random forest and gradient-boosted tree models with training-only CV.

Step 6: Make out-of-fold predictions and fuse learners using constrained soft voting or leakage-safe stacking.

Step 7: Choose the candidate (by using the validation discrimination, calibration, stability and decision-cost metrics).

Step 8: Calibrate probabilities on a separate validation segment and assess the calibration approach.

Step 9: Set operating thresholds based on expected costs, affordability rules and approved risk appetite.

Step 10: Do explainability, fairness, privacy, security and model-risk audits.

Step 11: Test once on the untreated test group and record everything that happens and what didn't work.

Step 12: Deploy with versioning, monitoring, human escalation and a tested roll back procedure.

6. Evaluation Plan and Expected Findings

The four main candidates for the proposed experiment are logistic regression, random forest, gradient-boosted trees, and a hybrid ensemble. The same definition of a target, data partitions, and evaluation population should be used for all candidates. The budgets for tuning should be approximately comparable and the results should be reported for model level metrics and for decision level results. The intent is to determine if complexity is adding value in the real world, not just improving one of the headlines.

HERON Leading strengths Main risks/Checks

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Logistic regression Probability is easy to interpret; efficient; baseline is transparent; may underfit nonlinear patterns; check functional form and calibration.

Random forest Attracts interactions; leads to variance reduction by bagging (bootstrap); may not be as calibrated as other methods; may be less stable and/or explainable

Tree-based methods: Gradient-boosted trees Flexible; powerful for tabular data Sensitive to tuning and drift; watch for overfitting and subgroup effects

Hybrid ensemble Can combine complementary error patterns Added complexity; validate weights, calibration, reproducibility and governance

It is hypothesized that tree-based learners can be more effective at discriminating if the data have nonlinear effects and interaction terms, and that a calibrated ensemble may provide a good ranking-probability quality control. This is a hypothesis to be tested, not a finding. The ensemble may not perform better than single models if learners make the same mistakes or the data size is small or the labels are noisy or the evaluation phase is significantly different from the training phase. If the performance of the simpler model is similar and the transparency or stability of the simpler model is materially superior, then a simpler model should be used.

The results should be presented in a table that includes the test ROC-AUC, precision-recall AUC, brier score, log loss, calibration error, sensitivity at a specified approval rate, the expected loss and confidence intervals. The report should contain the following: The default prevalence and the threshold selection rule. It should also present performance by vintage and meaningful segments, up to privacy and minimum sample constraints. When comparing, test observations need to be the same, and models should not be optimized based on the results of tests.

6.1 Practical Interpretation

Even if the model has a small improvement in AUC, it might only be of minor business value if there is a lack of calibration in the model, or if the improvement is only seen in a small portion of the business. In contrast, a model that performs

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

with a comparable AUC can be beneficial if it has better calibration, can alert to poorer performing portfolios earlier, or if it is more robust in times of stress. The selection criterion is then multidimensional – discrimination, calibration, cost, stability, explainability, fairness, and operational feasibility all play a role.

7. Discussion

The framework is based on the fundamental idea of applied machine learning: A model is only part of a decision system. A good algorithm might appear to be great but fail to perform in production due to data lineage, target construction, validation, calibration, decision policy, human review, and monitoring. Model selection should not be limited by accuracy alone since a high accuracy score can be obtained by predicting the majority class in low default cases.

The hybrid method allows for flexibility. Logistic regression gives a well known benchmark and a challenging model. Extensive manual specification is not needed to capture nonlinear interactions in tree ensembles. A prediction layer using probability combination can make use of complementary predictions, and a calibration layer can make output probabilities more meaningful for calculations of expected losses. The models, however, have maintenance costs, are harder to explain, and can cause instability if the component predictions change. This ensemble should be kept only if there is meaningful incremental value provided by the ensemble, as indicated by its validation.

The strategy also establishes organizational needs. Consistent data pipelines and documentation of the models are essential for a risk analyst. There is a need for evidence of the compliance achieved by model inputs, explanations, and decisions in compliance teams. Business owners should have an honest definition of risk appetite and the costs that are applied to make decisions on thresholds. Internal audit and model-risk management must be able to see versions, test results, overrides and monitoring records. These responsibilities should be delegated prior to the point of model operational importance, not after.

7.1 Ethical and Regulatory Considerations

If credit is unequally distributed in the past, or collection takes place in different ways, then this can be a case of exclusion being exacerbated. In this regard, data

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

minimisation, access control, limitation of purpose, and careful management of sensitive variables are crucial. Legally required notices and meaningful reasons must be given to applicants in case of any adverse decision. When human review can be done, reviewers must be trained and empowered to challenge outputs, not just confirm the automated output.

Regulatory requirements are different across jurisdictions and for different products. This paper does not offer legal counsel or give any indication that it meets the standards of any specific country's law. An institution should map the model before deployment to relevant banking, consumer-protection, privacy, anti-discrimination, outsourcing and model-risk rules. Validation should be based not only on a statistical audit but also legal review of data and decision process.

7.2 Limitations

The primary drawback is that this paper is proposing a framework without reporting results of a new borrower data set. Empirically, the expected benefits need to be determined. Public data sets might not reflect the present day institution's applicants, products or macroenvironment. Labels assigned in the past may be skewed for negative reasons because defaults are only known for borrowers who were granted credit. Performance can be compromised following deployment through macro-economic changes, policy changes and strategic behaviour. Explainability approaches are only approximations, and do not prove causation. Lawful availability of the data and the selection of normative criteria are limitations of fairness assessment.

8. Conclusion

In this paper, a hybrid algorithmic framework for credit-risk assessment based on ensemble learning techniques was presented. It features a logistic-regression benchmark, along with random forest and gradient-boosted tree learners, and then has validation-based ensembling, probability calibration, threshold optimization, explainability, fairness review, and continuous monitoring. The framework has been developed to cover the predictive performance and the practical needs of good lending.

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

The main recommendation is to consider the credit-risk system as end-to-end decision processes. Selection of a model should occur only after leakage-safe temporal validation and comparisons across discrimination, calibration, loss, stability, interpretability and subgroup behavior. The use of a hybrid ensemble may be justified if the ensemble provides strong and material improvements over simpler alternatives, but not if it is more complex. Future work should follow the framework with a well-documented dataset and compare other calibration and ensembling approaches, as well as test performance over an economic cycle, and examine the impact of model explanation and human review on decision quality and applicant outcomes.

References

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brier, G. W. (1950). Confirmation of predictions in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Chen, T., & Guestrin, C. (2016). XGBoost: An Adaptive Boosting Tree System that Scales to Any Size. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Friedman, J. H. (2001). A Gradient Boosting Machine (GBM) for function approximation using greedy algorithms. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hand, D. J., & Henley, W. E. (1997). Consumer credit scoring using statistical classification techniques: A review. Journal of the Royal Statistical Society: Series A, 160(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Lundberg, S. M., & Lee, S.-I. (2017). Common interpretation of model predictions. In 30th Annual Conference on Advances in Neural Information Processing Systems.

The National Institute of Standards and Technology. (2023). NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://doi.org/10.6028/NIST.AI.100-1

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 20-35

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit Scoring and Its Applications 2nd ed. SIAM.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The weak link of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7

[1] Department of Artificial Intelligence, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.