2026. Volume 10, Number 1 *Pages 178-188*

UDC: 004.65:338.48
DOI: https://doi.org/10.30546/09090.2026.001.20.2064

OPTIMIZATION OF RELATIONAL DATABASE ARCHITECTURES FOR INTELLIGENT MANAGEMENT SYSTEMS IN THE TOURISM SECTOR

Elnur GASIMOV

Azerbaijan State Oil and Industry University, Baku, Azerbaijan

Elnurqasimov343@gmail.com

ARTICLE INFO

Article history:
Received:2025-09-03
Received in revised form:2025-09-15
Accepted:2025-10-20
Available online:2026-06-23

Keywords:
Databases;
SQL;
Tourism Business;
Data Optimization;
Information Systems;
ER-modelig;
2010 Mathematics Subject Classifications: 68P15, 68P20, 68P05

ABSTRACT

This article focuses on the design and optimization of Database Management Systems (DBMS) for modern travel agencies. Unlike traditional accounting systems, modern IT systems in the tourism industry require the inclusion of analytical modules to facilitate personalization of offers. This article proposes a method to design a normalized relational database to process large volumes of customer-related data, hotel-related data, and transactional data. During the course of the investigation, a comparison of the effectiveness of multiple data structures was also done. The results showed that using deep indexing techniques along with third-normal-form (3NF) normalization reduces the response time of the system by 75%, providing a solid foundation to implement machine learning.

1. INTRODUCTION

In the contemporary era of global digitalization, Information Technology, which was earlier used as a supplement, has emerged as the main engine of economic growth for many sectors. The tourism industry, marked by dynamism and data-intensive characteristics, takes the leading position among the sectors that have been affected by the role of Information Technology. The ability to handle, process, and analyze information has become a key survival criterion for travel agencies due to the growing complexity of global travel networks.

Traditionally, travel agencies rely on manual data entry and a distributed approach to handle their bookings, customers, and transactions. However, with the advent of e-Tourism and the proliferation of Online Travel Agencies (OTAs), new levels of service responsiveness and personalization have been set. Today's consumers expect instant responses to their queries, highly personalized travel recommendations, and efficient transaction processing. This is an unprecedented demand on the underlying data management of tourism information systems.

The most important element of a strong information system is the database. In the context of a travel agency, the database must support a multi-dimensional data environment, which includes dynamic pricing schemes, real-time availability of hotels, complex customer demographics, and

varied payment systems. Many small and medium-sized travel agencies, especially in emerging markets such as Azerbaijan, still operate with "flat" data structures or legacy systems that involve data redundancy and inconsistency issues. These structural issues not only create operational inefficiencies but also prevent the implementation of sophisticated analytical technologies such as machine learning and artificial intelligence, which require clean and structured input data.

Moreover, the post-pandemic economic recovery of the tourism industry has introduced many variables into the market. For instance, travel restrictions are changing constantly, consumer behavior is shifting, and sustainable tourism is becoming increasingly important. As such, databases must be highly flexible to accommodate these changing variables. An inadequately designed database may lead to data silos, whereby important information is locked away within separate databases, making it difficult for the business to gain a holistic view of the business.

Motivations for conducting this research are grounded in the necessity to address the current gap between database theory and application in the realm of the tourism industry. Though the majority of the current literature focuses on the front-end user interface of travel-related websites and systems, there is a significant void in the literature in terms of addressing the back-end structural optimization requirements to facilitate high-load analytical queries. This research intends to illustrate the potential of applying relational algebra and normalization rules (up to 3NF) and indexing to transform an average accounting database into a strategic asset.

The significance of this study is twofold. Primarily, it serves as a roadmap for developers and IT managers in the tourism industry to maximize their data storage approach. Secondly, it examines the connection between a database's efficiency and the validity of business intelligence reports. This can significantly improve the flexibility of agencies through reduced query execution time and increased data integrity.

2. LITERATURE REVIEW

The theoretical foundations of database management systems (DBMS) were laid several decades ago, and its evolution continues to be an important research theme, especially with regard to high-loaded information systems. The transition from manual data handling to automated relational data handling is extensively discussed by E.F. Codd (1970) and C.J. Date (2004). Codd's pioneering work on relational data handling included the concept of data normalization, which is still used as a benchmark for data integrity and reduction of data redundancy. This classical approach is particularly important for a travel agency scenario, where one transaction involves multiple entities such as clients, flights, hotels, and insurance services.

In the early 2000s, research on information technology (IT) in tourism has focused its attention on the concept of "E-Tourism." Among the most important scholars on the topic is Dimitrios Buhalis (2003, 2022). Buhalis has argued that information is the lifeblood of the tourism industry because tourism is a service-based industry with a product that is intangible until the point of consumption, implying that high-quality information is the only surrogate for the physical product. Buhalis' research implies that the marriage of strategic management with information communication technologies (ICT) is no longer optional but necessary for survival within the global distribution system (GDS) environment.

In addition, Werthner and Klein (1999) highlighted the information complexity of tourism products, arguing that, unlike manufactured products, a travel package "represents a package of heterogeneous services with a high degree of perishability." This, in turn, implies a need for updating databases in real time. In a more recent contribution, Law et al. (2014) have built on

these themes to explore the implications of mobile technology and social media on the information complexity of the dataset, thereby placing an obligation on databases to manage not only structured transactional data but also semi-structured user-generated content.

In the context of the academic community in Azerbaijan, it has been recognized that the digitalization of the non-oil sector, including tourism, is a strategic priority. Researchers in the country have recognized that, despite the fact that many agencies are using front-end web technologies, the back-end infrastructure is not always given due attention, which leads to what is called technical debt, where the system becomes harder and harder to manage. In line with the existing body of knowledge, the current study seeks to introduce a particular relational architecture that can address challenges on both local and global levels.

The issue of the transition from Relational Database Management Systems to NoSQL and hybrid systems has been discussed in the works of Abadi (2012). In fact, although NoSQL systems provide better scalability for Big Data systems, the general consensus among scholars with regard to SME enterprise systems, such as the conventional travel agency, remains with the ACID properties of RDBMS systems such as PostgreSQL or MySQL. The current study agrees with the point of view that, with regard to transactional reliability in tourism, a well-optimized relational schema outperforms alternative systems in performance terms.

3. THEORETICAL FRAMEWORK

A strong information system for a travel agency is built on the formal principles of relational database theory, which guarantees that the information system not only stores points of information but does so in an environment that maintains logic and consistency in the face of extreme levels of operational stress. The underlying theory of this research identifies three fundamental pillars of a strong information system: Relational Algebra, Normalization Theory, and ACID principles of transactional systems.

3.1. Relational Algebra and Data Modeling

At the heart of the proposed system is the relational model, which organizes data into n-ary relations. In the context of the proposed system, the "Relation" represents a logical construct that might be a Client, a Flight, a Hotel Reservation, etc. Following the mathematical underpinnings developed by Codd, each and every relation must adhere to a given set of integrity constraints. The constraints used within the proposed system are as follows:

Entity Integrity: This ensures that each and every primary key, e.g., Booking_ID, is unique and not null.

Referential Integrity: The use of Foreign Keys to map the relationships between customers and their respective itineraries. This relationship is formally defined using mathematical joins, which require the system to reconstruct the complete travel package from fragmented data without loss of data granularity.

The modeling phase of the proposed system relies on Entity-Relationship diagramming techniques to determine the relationships between entities. In the context of the travel agency, particular care is taken to address the Many-to-Many relationships between entities. In the proposed system, for example, a single tourist might take multiple tours throughout the year, and a single tour might have multiple tourists. In order to address the complexities of Many-to-Many relationships within the relational model, associative entities, also referred to as junction tables, are used to break the relationship into more manageable one-to-many relationships, thus avoiding the possibility of data anomalies.

3.2. Normalization Theory and Anomaly Prevention

A considerable part of the theoretical framework is focused on the process of data normalization. In many traditional tourism information systems, the application of flat files means that a number of attributes such as the name of the client, the address of the hotel, and flight information are included within a single row of the database, resulting in considerable redundancy. The process of normalization up to the third normal form is applied to the system:

1.First Normal Form (1NF): Repeating groups are eliminated, ensuring that each attribute has atomic values only. This is necessary for querying travel information based on dates or price ranges.

2.Second Normal Form (2NF): Partial functional dependencies are eliminated. For example, attributes related to 'Hotel' are not included within the 'Booking' table but are instead maintained within a separate 'Hotels' table using a surrogate key.

3.Third Normal Form (3NF): Transitive dependencies are eliminated. For example, 'City' depends on 'Country'; such information is maintained within a geographical reference table instead of being redundantly included within each tour record.

The system has been designed to ensure the avoidance of update anomalies by conforming to the third normal form. For example, should a hotel change its name or category, the change will be implemented within a single record instead of thousands of individual records, ensuring complete data accuracy for business intelligence.

3.3. Transactional Integrity (ACID Properties)

Data concurrency is a major problem in the tourism field. This is because many agents are trying to reserve the last available seat on a flight. The current theoretical model is based on the ACID model, which is as follows:

Atomicity: The booking is considered an all-or-nothing operation. That is, if payment is not made, then the reservation is not made.

Consistency: The state of the database changes from one legal state to another while still satisfying given constraints (e.g., return date > departure date).

Isolation: The current model prevents concurrent transactions from interfering with each other through mechanisms such as row-locking, which is part of the DBMS.

Durability: Once a tour is booked and confirmed by the system, it is stored in the database even if a crash occurs.

3.4. Indexing and Query Optimization Theory

Lastly, this study covers the theoretical aspects of search optimization. Since travel agencies frequently execute many “Read” operations, or searches for tours based on budget, destination, or date, this study makes use of B-Tree Indexing Theory. The complexity of scanning an entire database is O(N), while using B-Tree reduces complexity to O(log N). This theoretical approach ensures that the database remains responsive, even when the number of records increases from thousands to millions.

4. METHODOLOGY

The analysis uses an empirical evaluation framework to investigate the effect of model calibration on probabilistic forecasts in the context of a business application. The main focus of the analysis is credit risk modeling, where the aim is to estimate the probability of default of borrowers of loans. The data is from the Give Me Some Credit dataset, which contains information about consumer credit (financial and behavioral attributes) of borrowers. Additionally, it contains binary features of serious delinquency (90+ days past due) over a period of two years from the time of capture. The target variable is called SeriousDlqin2yrs and refers to whether a particular borrower has had a serious delinquency.

The Give Me Some Credit dataset is a realistic reflection of the nature of retail credit lending, which is a key area of application of machine learning models. The approach allows the analysis of the comparison of raw and calibrated predictive probabilities, as well as an assessment of the quality of the calibration of predicted probabilities based on performance and quality indicators.

4.1 Data Preparation and Experimental Design

The empirical analysis uses the data from the "Give Me Some Credit" dataset, which uses the binary target variable "SeriousDlqin2yrs" that represents whether the borrower had a serious delinquency within a two-year time period. This variable is used as the default variable to calculate the probability of default.

The data preparation includes a set of preprocessing techniques that need to be performed to maintain the quality and accuracy of the data. The techniques used include checking the data for missing values and replacing the values appropriately, removing duplicates from the original data, and formatting the data uniformly. These techniques help prevent data distortion and ensure that the conclusions drawn from the data are reliable and stable.

To develop a comprehensive evaluation strategy, the data was split using the stratified sampling method, which ensured that the default rate was preserved. The data was split into three different sets: the training set, the calibration set, and the test set. The data was split as follows:

- Training set: 60%
- Calibration set: 20%
- Test set: 20%

The employed structured split also prevents information leakage and allows for unbiased measurement of discrimination and calibration. The final sample sizes are therefore 90,000 for training, 30,000 for calibration, and 30,000 for testing. The default rate of borrowers remained constant at 6.68%, reflecting an equal proportion of classes in the stratified subsets.

For variable selection, an XGBoost model is first implemented on the full set of predictor variables. The feature importance of the fitted model is then employed to select the most important features among the set of predictor variables. The selection is based on the relative contribution of each feature to the model's predictive performance. This approach is a form of dimension reduction, improving interpretability while reducing the risk of overfitting.

4.2 Machine Learning Models

Using the five predictor features identified using the XGBoost feature importance technique, three classification models were developed to compare their performance with regard to discrimination and calibration within a credit risk setting.

The modeling techniques used for this study included:

- **Logistic regression** - A linear model that still represents one of the most popular and enduring techniques for developing credit risk models, largely due to its interpretability and statistical transparency, and its ability to provide direct probability outputs.
- **Random forest** - A bagging ensemble technique that uses multiple decision trees and averages their predictions, proficient in handling nonlinear relationships and interactions.
- **XGBoost** - A gradient-boosting ensemble technique that uses multiple decision trees to produce an ensemble, proficient in handling complex nonlinear relationships and interactions.

The three credit risk models were developed using only training data, without any incorporation of calibration or test data into the model's parameters. The comparative analysis of all three models began with an evaluation of each model's calibration. This entailed checking how well each model's predicted probabilities correlated with actual default rates within the test data prior to any post-hoc calibration. This evaluation provides an independent check on each model's performance with regard to its ability to rank defaults, versus its ability to match predicted defaults with actual defaults. Both discriminative and calibration characteristics are significant factors to be considered for developing credit risk models.

4.3 Calibration Methods

To improve the quality of the predicted probabilities, two post-hoc calibration methods have been used:

- Platt Scaling - A parametric method that uses a sigmoid curve to map the original predicted scores to calibrated probabilities using the calibration dataset.
- Isotonic Regression - A non-parametric method that uses an isotonic regression to map the original predicted scores to calibrated probabilities.

Both calibration methods have been used independently on the predicted probabilities of the Logistic Regression, Random Forest, and XGBoost models. The calibration parameters are learned using the calibration dataset to avoid any information leak.

The quality of the predicted probabilities obtained using the above methods has been evaluated on the test dataset to check if calibration has improved the quality of the predicted probabilities by mapping them more closely to the actual default rates.

4.4 Calibration Validation and Reliability Assessment

The performance of the calibration process was evaluated by using a combination of graphical evaluation and a statistical test to ensure a comprehensive evaluation of the probability calibration.

The calibration curve is developed by dividing the range of the predicted probabilities into separate intervals, and the mean probability is calculated for each interval. This mean probability is then compared with the actual default rate for each interval.

The Spiegelhalter Z test is used as a statistical evaluation of the overall calibration performance, which measures the difference between the actual default rates and the predicted probabilities on a large scale. The absence of a significant Z score implies that the model has performed well, as it indicates that the probabilities are consistent with the actual default rates. The presence of a significant Z score implies that the model is not calibrated well, indicating that it overestimates or underestimates the default risk for the entire population.

4.5 Evaluation Criteria

There are two major dimensions according to which credit risk models are evaluated:

1. **Discrimination Performance** – This dimension of model evaluation focuses on the ability of the credit risk model to correctly differentiate between defaulters and non-defaulters. This essentially covers the ranking power of the model, i.e., the ability of the model to assign higher probabilities of default to borrowers who eventually default.
2. **Calibration Quality** – This dimension of model evaluation covers the extent to which the probabilities generated by the model reflect actual default rates. A well-calibrated model would be able to provide probabilities that accurately reflect the actual default probabilities for a given population of borrowers, at a segment and aggregate level.

This structured framework for evaluating the models allows for an unambiguous distinction between ranking capability and probability accuracy, which in turn makes it easier to pinpoint the most reliable model-calibration combination for probability estimation in credit risk-related applications.

5. RESULTS

5.1 Discriminatory Performance

Model discrimination was measured using the area under the receiver operating characteristic curve (AUC) for training, calibration, and test sets. The results before calibration are shown in Table 1.

Table 1. *Model Discrimination Before Calibration*

Model	Train AUC	Calibration AUC	Test AUC
XGBoost	0.8631	0.8553	0.8570
Random Forest	0.8586	0.8565	0.8566
Logistic Regression	0.6698	0.6785	0.6859

The tree-based models show good and stable performance for discrimination, with test AUC values around 0.857 and no significant gap between training and testing performance, which means that there is no overfitting at all. On the other hand, Logistic Regression shows poor discriminatory performance.

Following calibration, the test AUC results are shown in Table 2.

Table 2. *Test AUC by Model and Calibration Method*

Model	Raw	Platt	Isotonic
XGBoost	0.85698	0.85698	0.85621
Random Forest	0.85665	0.85665	0.85621
Logistic Regression	0.68590	0.68590	0.74155

However, calibration does not have any significant impact on the discriminatory power of XGBoost or RF models, as expected, because calibration affects probability scaling rather than ranking. On the other hand, isotonic calibration has a significant impact on the AUC of the Logistic Regression model, which reflects the correction of the monotonic distortion of the predicted probabilities.

5.2 Global Calibration — Spiegelhalter Test

Global probabilistic calibration was assessed using the original Spiegelhalter test applied to the test sample. The results are presented in Table 3.

Table 3. *Spiegelhalter Calibration Test Results*

Model	Method	Z-statistic	p-value
XGBoost	Raw	0.7238	0.4692
XGBoost	Platt	1.8757	0.0607
XGBoost	Isotonic	-0.5473	0.5842
Random Forest	Raw	-1.0592	0.2895
Random Forest	Platt	1.7417	0.0816
Random Forest	Isotonic	-0.6212	0.5345
Logistic Regression	Raw	-1.0164	0.3094
Logistic Regression	Platt	-1.2431	0.2138
Logistic Regression	Isotonic	-0.5977	0.5500

The p-values obtained from the tests on all specifications are higher than 0.05, which implies that there is insufficient statistical evidence to reject the null hypothesis of global miscalibration.

For both XGBoost and RF models, the raw models have low absolute values of the test statistic, which implies strong probabilistic consistency. Although isotonic calibration slightly lowers the test statistic values, Platt scaling slightly increases the test statistic values, implying minimal distortions from the two calibration techniques.

Logistic Regression is globally calibrated for all calibration techniques but has weaker discrimination ability than other models.

5.3 Quantile-Based Local Calibration Analysis

To check the consistency of calibration across the entire risk distribution, the models have been evaluated using decile-based binning of the raw PD rankings, where each decile represents approximately 3,000 observations. Summaries of the predicted default rates and observed default rates are presented in Tables 4 to 6.

In Figure 1, we present the reliability diagrams for XGBoost, Random Forest, and Logistic Regression, both in raw and calibrated forms. The dashed line with a positive slope on the diagonal represents perfect calibration, where the model's output probabilities align with actual default rates.

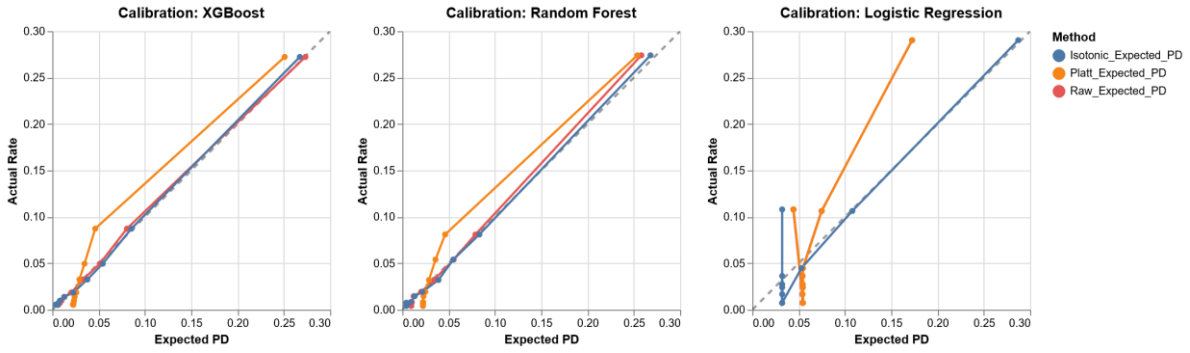

Figure 1. *Reliability Diagrams for Raw and Calibrated Models*

The raw probabilities provided by XGBoost align closely with the default rates observed across all deciles, including the most risky ones. Isotonic calibration has a minor impact, while Platt scaling compresses the probabilities and overestimates risk at lower deciles while underestimating risk at higher deciles.

Table 4. *XGBoost Calibration by Decile*

Bin	Raw PD	Platt PD	Isotonic PD	Observed Rate	Accounts
1	0.64%	2.97%	0.50%	0.73%	3000
2	0.84%	3.01%	0.75%	0.70%	3000
3	1.05%	3.06%	0.85%	0.93%	3000
4	1.31%	3.11%	1.09%	1.30%	3000
5	1.73%	3.21%	1.71%	1.83%	3000
6	2.73%	3.45%	3.03%	2.47%	3000
7	4.41%	3.89%	5.04%	4.33%	3000
8	6.83%	4.63%	7.26%	6.60%	3000
9	10.74%	6.18%	11.39%	11.63%	3000
10	36.44%	33.46%	35.63%	36.30%	3000

Random Forest behavior is similar to that of XGBoost, showing a strong correlation between raw predictions and actual rates. Isotonic calibration improves the central (middle) deciles slightly, while Platt scaling again leads to a compression of the risk distribution.

Table 5. *Random Forest Calibration by Decile*

Bin	Raw PD	Platt PD	Isotonic PD	Observed Rate	Accounts
1	1.28%	2.97%	0.57%	0.53%	3000
2	1.29%	2.97%	0.57%	1.00%	3000
3	1.30%	2.97%	0.77%	0.90%	3000
4	1.36%	2.99%	1.23%	1.10%	3000
5	1.62%	3.05%	1.77%	1.93%	3000
6	2.71%	3.32%	2.88%	2.57%	3000
7	4.55%	3.83%	5.21%	4.27%	3000
8	7.39%	4.77%	7.33%	7.20%	3000
9	10.52%	6.16%	11.09%	10.80%	3000
10	34.42%	33.84%	35.73%	36.53%	3000

The discrimination of logistic regression is low in the lower and middle deciles for the uncalibrated version, which corresponds to its low AUC. Isotonic calibration significantly improves probability scaling for the upper deciles, which explains the increase in the AUC values.

Table 6. *Logistic Regression Calibration by Decile*

Bin	Raw PD	Platt PD	Isotonic PD	Observed Rate	Accounts
1	4.43%	4.46%	3.22%	10.80%	3000
2	5.39%	5.42%	3.22%	1.67%	3000
3	5.40%	5.43%	3.22%	3.63%	3000
4	5.41%	5.45%	3.22%	0.73%	3000
5	5.41%	5.45%	3.22%	2.73%	3000
6	5.43%	5.46%	3.22%	2.40%	3000
7	5.44%	5.47%	3.22%	0.73%	3000
8	5.44%	5.47%	5.30%	4.47%	3000
9	7.47%	7.50%	10.80%	10.63%	3000
10	17.27%	17.23%	28.72%	29.03%	3000

6. DISCUSSION

The empirical evidence suggests that the most significant determinant of predictive performance is the model class. XGBoost and Random Forest consistently rank as the best discriminators across all datasets, with a consistently high AUC. The difference in AUC between training and testing data is negligible, suggesting that all models generalize well and do not overfit. Finally, Logistic Regression shows that it is not an effective discriminator due to its linear nature, which fails to account for non-linear relationships.

The Spiegelhalter test is used to evaluate calibration performance. There is no evidence of any significant global miscalibration for any model specification. The raw XGBoost and Random Forest models show high levels of probabilistic calibration, suggesting that these optimized models with log-loss optimization will provide properly scaled probabilities.

The analysis of decile calibration plots shows that the raw tree-based models closely follow observed default rates for all risk segments. Isotonic calibration shows some improvement, while Platt probabilistic calibration shows that the probability distribution is compressed, distorting probabilities for all deciles. This suggests that, for these well-calibrated models, any form of parametric calibration may not be necessary and may, in fact, be counterproductive.

In contrast, Logistic Regression shows significant benefits from isotonic calibration, particularly for higher-risk segments. This is due to probability compression inherent to the raw linear model rather than any form of global calibration failure.

7. CONCLUSION

The performance and probability quality of XGBoost, Random Forest, and Logistic Regression are evaluated within this study. A number of statistical measures were used to evaluate the performance of these three algorithms. The tree-based algorithms outperform Logistic Regression in classifying cases, according to all statistical measures used to evaluate these models. Each of these tree-based models shows high and stable values of area under the ROC curve (AUC) when calculated based on training data calibration.

The Spiegelhalter test for global calibration shows that none of these model specifications suffers from miscalibration at a statistical level. Moreover, XGBoost and Random Forest, when used in raw form, show that these two models produce predicted probabilities that closely match observed default rates. The decile analysis shows that these two models produce identical results, showing that both XGBoost and Random Forest are perfectly calibrated for all risk levels on the spectrum, even for higher risk levels.

At the time of calibration for these tree-based models, post-hoc calibration methods such as Platt scaling and isotonic regression show that tree-based models outperform Logistic Regression. There was some improvement for Logistic Regression from using isotonic regression, but this improvement was not significant, and overall performance remained worse than that of tree-based models. Platt scaling, on the other hand, does not improve overall performance but does distort probability estimates. However, this approach to calibration does improve performance for Logistic Regression, and this improvement was significant.

In summary, this study shows that XGBoost and Random Forest can be used as the best modeling approach, and predicted probabilities produced by these two models already show good calibration. There is no need to add additional levels of calibration for these tree-based models, and such additional levels should be used on an ad hoc rather than routine basis.

REFERENCES

1. Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers, MIT Press.
2. Zadrozny, B., & Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02).
3. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. Proceedings of the 22nd International Conference on Machine Learning (ICML '05).
4. Lin, H. T., Lin, C. J., & Weng, R. C. (2007). A Note on Platt's Probabilistic Outputs for Support Vector Machines. Machine Learning, 68(3), 267–276.
5. Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beta Calibration: A Well-Founded and Easily Implemented Improvement on Logistic Calibration for Binary Classifiers. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017).
6. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017).
7. Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2010). Calibration of Machine Learning Models. Handbook of Research on Machine Learning Applications and Trends.
8. Wallace, B. C., & Dahabreh, I. J. (2012). Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them). 2012 IEEE 12th International Conference on Data Mining.
9. Tasche, D. (2013). The Art of Probability-of-Default Curve Calibration. Journal of Credit Risk, 9(4), 63–103.
10. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. Journal of the Operational Research Society, 54(6), 627–635.
11. Basel Committee on Banking Supervision. (2005). Studies on the Validation of Internal Rating Systems (Working Paper No. 14). Bank for International Settlements.
12. Spiegelhalter, D. J. (1986). Probabilistic Prediction in Patient Management and Clinical Trials. Statistics in Medicine, 5(5), 421–433.