



Predictive models performance in financial services for identifying at-risk customers

Ashok Ghimire

Westcliff University, USA

ashok.ghimire1991@gmail.com



Corresponding Author

Ashok Ghimire

ashok.ghimire1991@gmail.com

Article History:

Submitted: 23-12-2024

Accepted: 24-12-2024

Published: 24-12-2024

Keywords: Independently and jointly used in assessment, financing, and credit, risky customers, artificial intelligence, pre-processing data, assessment parameters, post-model quantification.

Brilliance: Research of

Artificial Intelligence is licensed under a Creative Commons Attribution-Noncommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

It is therefore essential in financial service where the concept of pre-emptive modeling is used to select customers that are most likely to engage in negative behaviors like loan defaulting, fraud or churn. This paper will seek to analyze different modeling techniques like Logistic regression, Decision trees, Random forests, Gradient boosting, and Neural network among others. The paper pays considerable attention to the data preprocessing step; problems such as imbalance ratio and missing values as well as important parameters such as precision, recall, and AUC-ROC to measure the efficiency of the models. Some of these problems that have been highlighted in the review include; data privacy, compliance to laws and regulations among others and model interpretability which plays an important role in the financial industry. This means that new achievements like XAI, Real-time analytics, Federated learning have contributed to the improvement of model interpretability, model size and data security. Such include technological applications to increase accuracy of Risk prediction while cutting costs, and help the financial institutions retain the trust of the customer given the ever changing regulation. By integrating these novelties and ethical issues, some more financial institutions should enhance the centrality of the presented predictions to improve the identification of the risks and customers' appeal.

INTRODUCTION

As financial services change over time, the risk characteristics of buyers have become far more 'safe' to transact with, therefore the accuracy of the probabilistic and controllability of negative buyer attributes should be boosted. An at risk customer can be said to be any customer that is likely to be in debt and also include any customer that is involved in fraud and other circumstances that make him or her a risk to institutions. In this way, by defining these customers the financial institutions can apply some preventive measures including the following ones; changing the tariffs of the loans, checking the credits or working on the existing non-recourse options [1]. Therefore, predictive modeling has established itself as an important and enabled technique in this environment boosted by improvements in data analysis, machine learning, and artificial intelligence. Its technical name is predictive models; models that based their results on experience of previous performances in similar events. In financial services these models apply in credit 'grant' and risk assessment, customer 'retention', credit card 'fraud' and in the 'recovery' of del suspended debts [2].

Since the models can identify customer behavior and activity, it assists institutions in categorizing customers, identifying and segmenting profitable customers and decision making. For instance, based on the historical rating data, predictive model may classify the customer with low credit scores, and payment history showing late payment as high risk; and suggest the appropriate action if the organization does not wish to default on the credit sales. Using the predictive models is therefore useful in churning the following efficiencies which include efficiency, scalability as well as accuracy. More importantly, set-and-measure techniques or fixed-rule evaluation models may not be adequate sharply of more frequent use and innovation of data in financial contexts partly because of the richer features [3]. Whereas, the predictive models in real time environment handle a huge amount of data and identify many other aspects lots of which are not easily visible to the human analyzing the data set. Doing so, because the financial institutions are also going through the process of digitization, the implementation of predictive analysis has become compelled. There are a number of challenges related to the application of forecast models however this is when they are applied as an approach [4].

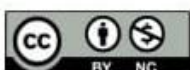




Figure: 1 showing ways to improve predictive model

This is so because of the dynamic character of financial markets and the legal and ethical element that are important cornerstones of such models. From where I stand, it cannot be the case that there are no means, methods on channels that make the bias, customer privacy and or compliance a non-concern when using or producing any form of predictive algorithms. Second, some of these models originating from these methods are of a more understandable kind while the use of algorithms of which the operation is not easily comprehensible will raise doubt from customers. The model's work for detecting customers for credit risk in financial services should be a focus of this paper [5]. It will categories kinds of models used, some measures that have been used in evaluating the models, challenges faced and achievements made in the field. Lastly, in the subsequent section of the review, the literature and new practices establishing the elaboration of the efficient use of the above and other predictive modeling approaches in the case of the financial risk management in the last few years and the probable areas for future research will be enumerated. In such a manner, it aim at creating awareness on the potential of PA as a tool in the act of constructing sustainable resilience in the financial institutions [6].

DEVELOPMENT OF VARIOUS PROGNOSIS MODELS APPROPRIATE FOR USAGE OF THE FINANCIAL SECTOR

Almost all business problems are normally solved under these predictive models which have virtually turned out to be an industrial pillar in financial services. It also allows the valuation or of the financial credits using past with present data making models to come up with the right decision of the financial status of institutions [7]. Other results like loan defaults, credit score deterioration, customer churn and other fraud related activities are estimated by the acquired patterns and, based on these, appropriate predictive models are built. The existing types of predictive models are;

Logistic Regression: In this case logistic regression which is helpful in solving binary classification issues is used in explaining ability to pay or likelihood that a customer will be a defaulter based on past records [8].

Decision Trees and Random Forests: To be a bit more precise, such models are easy to use and also rather effective when working with big sets of data; and indicate, the probable cause of customer risk.

Neural Networks: Due to their capability to scan intensely through the many complex, cross over and presumably non-HT relations in a dataset, neural networks are most appropriate for the more valuable augment operations such as credit rating or fraud analysis [9].

Predictive models in financial services are applied across multiple domains, such as:

Credit Risk Assessment: A customer behavior credit control system Fresh mean includes issuance of the likely default on the loan along with the credit limits.

Fraud Detection: Besides, the detecting false patterns in the transactions to give hints concerning the fraud.

Predictive modeling



Figure: 2 showing predictive modeling

Customer Retention: Thus, precisely identifying the customers who might churn in the near future, and guaranteeing that different techniques are developed for the specific goals of customer retention. The inputs of these models vary from basic transactional data, to payment history data, demographic data, and right up to behavioral data. Besides, there has been progress in artificial intelligence and machine learning, which have improved the ability of firms to create a dependable and replicable instrumental prediction used in real-time supervision and decision-making among the firms in the finance business [10]. On the financial aspect, risk management, business processes, and the confidence in the financial companies have become the concentration since the parameter has increased its financial in formativeness. However, innovation cannot be an auxiliary of, but it may become an adversary to, regulators and ethics [11].

CRITICAL STATISTICAL MEASURES USED TO ASSESS THE EFFICIENCY OF A MODEL FOR A FORECAST

As can be seen, the efficiency of the large number of models requires estimating as their results set outcomes for the customer and institutions in the financial sector. Factors involved in models help in determining how well or badly the model is identifying risky customers at the same time helping in evaluating the efficiency, time and fairness of the concerned model [12].

Accuracy: Accuracy reports total efficiency of results that give percentage to correct option of total post it options whether correctly or inapt. However, what was very easy to compute, even this accuracy parameter may not look very accurate in case of a model dealing with an imbalanced data set, where the density of the minority class, let's say the risky client for us, is much less than the total customers—a false negative issue [13].

Precision: Accuracy – refers to the number of true positive — number of risky customers that are indeed risky and found to be such by the model to the total number of risky customers that the model has flagged. That is understandable because the percentage of true positives is high when precision is high; consequently, interventions are designed to target the right populace [14].

Recall: Sensitivity or Recall reveals the number of the actual endangered clients who were included in the constructed profiting model. A high recall score ensures that the model is capable of identifying most of the at-risk customers thus eliminating on cases of Miss them – false negatives [15].

F1 Score: As you may remember, F1 score comes from the precision and the recall and hence can be appropriately used where the accuracies of the two are proportional. It is useful most of all when the cost of false positive and false negative result is equal and the volume of samples is different [16].

OTHER RELEVANT METRICS

Specificity: The number ratio of accurate customer's non-risk classification and potential gainful users who can be rather inconveniently contacted.

Confusion Matrix: A larger toolbox that also shows the actual positive, actual negative, false positive and False Negative



results for high level evaluation.

Cost-Sensitive Metrics: Measures like this, especially when those effectively deem a product junk (false positives), or exclude a product (false negatives) – all something that filters particularly when needed to an important degree in the financial services arena [17].

Based on the business environment of an organization and its objectives there is choosing of performance measure. For instance, in fraud detection the cost of wrongly classifying negative observations as positive have high life time cost while the cost of missing on positives is low they hint at the fact that both precision and recall should be moderately high [18]. Therefore, their strategic consideration enable financial institutions easily, accurately, fairly and efficiently to apply the related models for identifying customer at risk.

The types of data that were used in the studies and preliminary data processes that have been conducted on the collected data are also presented here. In financial services predictions, which are also especially important while working with an 'at-risk' customer, data makes up the largest part of a model [19]. Through the amount and guns of the data and through the reach of the data all configurations of the predictions used in different kinds of prediction models are determined. However, raw data is not always clear and free from noise and bias; raw data is very noisy and unreliable and the process of preparing data becomes very sensitive to derive real value [20].

TYPES SOURCES OF DATA IN FINANCIAL SERVICES

Predictive models rely on diverse data sources to identify at-risk customers, including:

Transactional Data: This include customer transaction information for instance; transaction frequency, transaction volume and transaction types that help in developing behavioral patterns of formulation of financial formatter or risk [21].

Credit History: Information can be accessed from the credit bureaus in form of credit scores, history of payment, loan accounts and credit utilization.

Demographic Data: His age, income, his working state or her living state together with other customer information assist in putting the customer in his financial risk position [22].

Behavioral Data: Data such as log in details about a customer, payments and or use of services paints a picture of the increasing or reducing economic value of a human being to the society.

External Data: External circumstances, the state and the analysis of information gathered from the social networks will improve the level of differentiation due to the fact that it provides universal data [23].

Preprocessing Techniques

Preprocessing is the technique that make it possible for a dataset to be clean and rid of all things that could have had a negative impact on the finding when done through data in predictive modeling. Key steps include:

Feature Engineering: This improvement includes the addition of new parameters, which enrich models and depiction of a certain problem at least, calculation of the DTI or customer attrition rate. However, categorical variables are converted to number form by other options such as one-hot encoding or label encoding most of the time [24].

Outlier Detection: A technique for finding out other potential bifurcation points in the same data set for which the performance of other yet different models may be wrong.

Data Normalization and Scaling: As features in the datasets all those that have variation must be standardized since, for instance, while some independently influential variables can overwhelm the other variables particularly when using the nearest neighbor distance based methods [25].

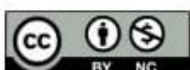
Addressing Imbalanced Data: The problem of skewed target classes is as follows: In financial datasets a minority of customers are in risky, hence target classes are skewed. The methods which have been recommended to be used to deal with this type of a problem are oversampling (SMOTE), under sampling, or use of cost-sensitive algorithm [26].

Dimensionality Reduction: Some of the features were removed in a way to get rid of problem of number of features which are often irrelevant; feature extraction were also used including principal component analysis (PCA). Real-time data preprocessing in the framework of big data analytics similarly becomes one of the most complex processes among those discussed in this article [27].

Data Privacy and Compliance: Due to the data offered by customers appropriate legal requirements including GDPR or CCPA need to be compliance.

Dynamic Nature of Data: Currency financial behaviors result into updated of such datasets and the overall preprocessing workflows thus the frequent updating [28].

Heterogeneous Data Sources: Quantitative data in terms of format, quality and structure origin can be combined with other data not without issue. This means that only the models to be used for analysis will access the best, important and balanced quality data that is very vital in identifying the vulnerable customers. It is a point that institutions interested in asserting the right and efficient data preprocessing systems for realizing the best result in the application of the predictive analytics in managing the financial risk may consider [29]. In the area of operational financial data analysis the most fundamental problem of customer risk evaluation poses several significant questions related to data heterogeneity and relative randomness, issues of model interpretability, legal and ethical concerns. Such factors may dilute the characteristics of the predictive models to enable adjustments on endorsement of new strategies by financial institutions [30].



DATA-RELATED CHALLENGES

Incomplete or Inaccurate Data: Real data is constituted by the actual data supplied by customers added to missing data, wrong data or obsolete data in the customer data base and in the transaction data.

Data Imbalance: In a company, a population at risk is usually the small number of the total number of patients, and this changes data. This state can make the given predictive models pay more attention to the majority class; hence failure in identifying the potentially ill people [31].

Data Silos: They observed that it has been very difficult as a result of which especially financial institutions have not been able to compile data in a more integrated fashion which gives a much cleaner data for analysis [32].

Model Interpretability: This is why the first-order models obtained upon training with neural network are incredibly difficult to understand and we get the name 'black box'. Of these, the propositions for customers and financial firms require explain ability of financial services to increase confidence in such models [33].

Regulatory and Compliance Issues: The financial institutions are situated in rather exposed zones. Depending on the nature of a particular application of A/B testing, the predictions derived from the machine learning analysis must be cast in the legal framework stipulated by the regulations like GDPR or CCPA that the customers themselves supply data to. Indeed, models are called upon to demonstrate absence of bias and/or refrain from writing out what will be used to perpetrate bias [34].

Ethical Concerns: For instance, while doing it, there is the need to appreciate that where the information employed in developing those models has characteristics of injustice in the society then those models will similarly be imprinted with that injustice. For example, some demographic classes may be chosen wrongly as risky based on training information of the model. To say that is what institutions must do to defend models that are not destructive of ethical and non-discriminant in the process [35].

Dynamic Financial Behavior: Customers and financial markets are also mobile responding the conditions inside the economy, the change of their persona – if they are buyers for their business – or something else fundamental, including the sickness – in this case, the virus pandemic. Assets of predictive modeling that need updating includes: The use of frequently changing models when there is dynamic characteristics Dynamic characteristics are critical to the method since they indicate when more frequently changing models should be used [36]. Such problems can only be solved with the help of the better technical solutions like explainable AI, data preprocessing, model monitoring along with the use of ethically correct AI and correspondence to the current framework. These threats can be maintained and opted into provided that they let financial institutions effectively differentiate customers, who are analytically high risk, as well as remain loyal and responsible [37].

COMPARISON OF ALGORISMS: PREDICTIVE MODELING

Today, there is a very large number of approaches to predictive modeling which vary in concern with the complexity of the algorithms, their interpretability and the efficiency of their performance and before proceeding to the question what kind of clients should be identified regarding credit risk in the context of financial services, these aspects should be noted. It is based on its served concern, nature of data possessed by the business and the need of the context. In the conventional use of statistics, growth of stocks is expected to decline when using techniques such as the Dow Jones Industrial Average [38].

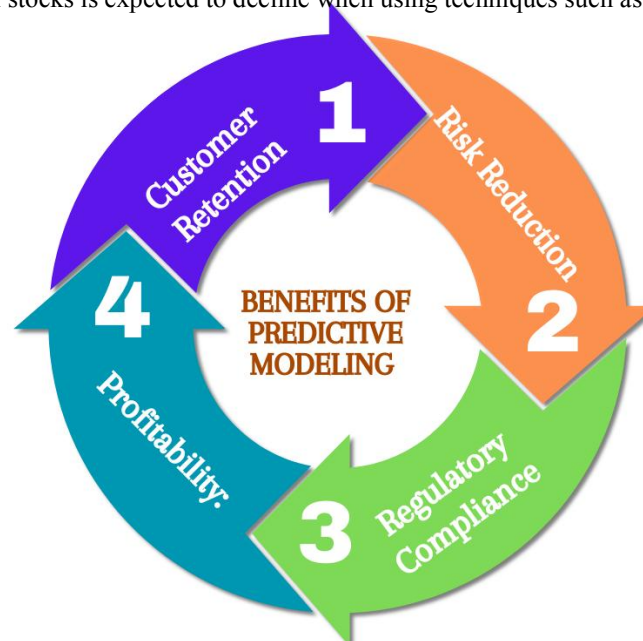


Figure:3 showing benefits of predictive modeling



Logistic Regression: meaning logistic regression is well adopted in binary classification problem which are the scenarios whereby a customer is likely to default or not requires elaboration. Thus the form is clear, easy to understand and it has also been used to test the performance on sample data during testing. But it is not suitable for models with non-linear and complex forms of data which is today's large scale financial data sets [39].

Linear Regression: In some respects, linear regression as is apply principally for use with continuous data can be used for several risk predictions. In other words, like any other persona method, it is most unsuitable for nonlinear or highly multivariate interdependence of the variables under consideration [40].

Machine Learning Models

Decision Trees: The largest purpose behind why decision trees are used is firstly, decision trees are easy to understand and secondly it is possible to acquire data that are not non-linearly related by using the decision trees. However there are decision rules which are clear, this make it easier to explain the predictions made to the clients. But they are highly claimed to be over fitted especially when a large number of data is used in the analysis [41].

Random Forests: Another exotic way to build decision trees, it poses accuracy and modality concerns and makes decision trees random forests. They are especially useful where there are distorted data distributions as tends to be the case in this field of financial service. However, they are very time consume, and the final output is not as easy to analyze as that of pure decision trees [42].

Gradient Boosting (e.g., XGBoost, LightGBM): A primary case study of a structured data set and an imbalanced dataset using gradient boosting techniques. Thus, in a number of cases they are described as more accurate compared to comparatively simple models. However, they depend upon parameters, cannot easily be interpreted and are therefore not suitable for the applications to regulated environments [43].

Support Vector Machines (SVM): SVM are good for big values of dimensional space or feature space and is able to reparative class with nonlinear boundaries. However, they are used far more rarely in such areas as bank and other sides of financial service due to the calculation costs and the inability to work with big data [44].

Deep Learning Techniques

Neural Networks: Deep learning algorithm that contains neural network category is used for modelling complexity non-linear relation. They are opportune in the areas that contain large data base structure with many attributes, for example in fraud detection. However, due to their black box structure and the fact that most of them are computationally intensive they cannot be implemented in situations where interpretable models are desirable [45].

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): These are for data that exist in the temporal domain for that which holds the transactions logs. They are effective for modelling temporal structure ^ though they are high-demand in terms of data and computation [46].

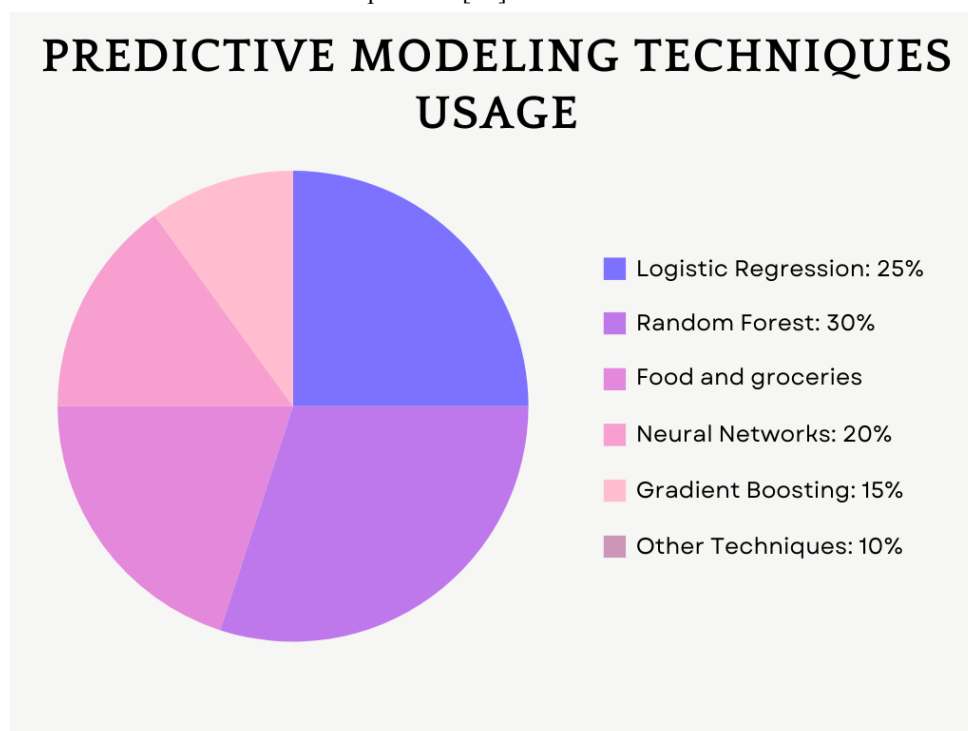


Figure: 4 showing predictive modeling technique usage



STRENGTHS AND LIMITATIONS

Interpretability: All the new fashionable names in the field of machine learning are opaque and hence can only be implemented in very controlled settings such as logistic regression, decision trees, and naive bays. What has been observed more often than not is that any enhancements in the creation of the machine learning and deep learning models seems to correlate with a decrease in interpretability at the ground level [47].

Scalability: The large scale data can also be solved when the higher form models such as gradient boosting like the neural network unlike the normal models of learning.

Accuracy vs. Complexity: Nevertheless, complicated models, which are said to be less sensitive to the problems mentioned above, are more precise, require much more data, computing resources and experience than simple models. This is the decision that has to be taken on which of the predictive modeling techniques to use: It is, therefore, unclear if high accuracy is required, if high interpretability is required or if computational power makes it feasible. Where compliance and ethical questions are inherent in activities conducted in the financial services industry, this is where further development of ontological layers using apparatus such as SHAP or LIME to retain explain ability remains possible [48]. In particular, the paper demonstrates that right choice of research method and construction of precise prediction models enable some institutions to identify potential defectors or misbehaving customers.

He also mentioned that mainly in the financial service domain the field of predictive modeling has taken a significant amount of gestation period of just few years due to reason of implementation of higher technologies, new algorithms and new methodologies. They have also enhanced the efficiency, effectiveness and flexibility with which financial personalities are able to identify the risky customers [49].

Machine Learning Enhancements

Ensemble Methods: It seems now it is possible to use some of the most powerful methods that can be applied to financial risk assessment including random forests, gradient boosting (XGBoost, LightGBM), and stacking. These models are developed from simple algorithms because integrate the results of other models as they try to arrive at better result [50].

Automated Machine Learning (AutoML): Automated machine learning tools assist in determining whether the number of models should be necessary and how many features are effective and important for selection, in addition to the tuning of the hyper parameters of the models that specialists can apply the accurate high-power predictive models with efficiency [51].

Deep Learning Innovations

Chambers, Lee, and Qiu (2016) have noted that this financial risk modeling has begun using deep learning especially in the complicated patterns.

Neural Networks: In the tiered models the feed forward and convolutional neural networks are implemented in the transactional data and the customer feedback for the risk estimation [52].

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): They become most valuable where data is time-stamped: perhaps checking account information that might include prediction of quantic defaults or frauds [52].

Transformers: However, transition techniques such as the transformers are as BERT initially conceived for Natural Language Processing are if applied to the finance area, for instance, for continuous detection of fraudulent behaviors in unstructured data sets [53].

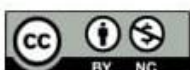
Hybrid and Ensemble Models: Rather frequently use is made of this, which are middle-aged, basic, regular statistical methods with the new generation of artificial neural networks. For instance when working with logistic regression together with gradient boosting learning here the institutions can learn from the outcome that while logistic regression has numerical interpretability, on the other hand, gradient boosting learning has highly numerical capabilities [54].

Real-Time Predictive Analytics: While big data ultimately leads to real time risk assessment, so does cloud computing. Through this technology, financial institutions are able to analyze big transactional streams with a view to reducing on risks in real time. This is advantageous especially in fraud and credit risk because they are real time as soon as the results of that analysis is needed is given [55].

Explainable AI (XAI): In particular it is stated that the legal and ethical regulations regarding the analysis of customers' data, coming from the field of financial services, must be associated with explainable models. Customer At-Risk identification is under the process of being described by the use of the SHAP and LIME tools to explain features in classification [56].

Alternative Data Sources: The application for example of spying feeds using social media update, location, and behavior data has grown beyond risk analysis. Here the sources are used conjugated with standard financial data to get the better figures of the customer behavior [57].

Federated Learning: Therefore, during the training of models based on decentralized data through the use of federated learning a financial institution it does not report specific information relating to the customers. This limits the privacy of its users although offers results from distributed datasets at the same time. They are revolutionizing how situations in the financial service industry are predicted so as to identify troublesome clients to institutions [58]. During the process of further development of the framework, integration of AI, the result of the implemented real-time analytics, and the usage of the other type of data source will improve the risk assessment functions, even as it points at direction for regulation





and ethic consideration. They say that those institutions, which have adopt these innovations, are well-placed to handle risk AND, arguably more importantly, sustain customer confidence where digital finance solutions are gradually assuming acceptability.

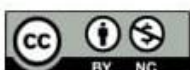
CONCLUSION

This is true in the financial services field where; predictive modeling has relieved institutions of carrying out client management in the traditional ways as there are far better approaches to identifying clients who are most likely to turn into delinquent. These models through identical data and statistical analysis and other intelligent algorithms including data mining help the institutions to manage various risks such as loans, frauds and customers' desertion. However, the use also has a number of challenges, which cannot ignore when using such type of information. Effective predictive modeling therefore relies on the data-transactional, demographic, credit and behavioral data collected from other sources. Such data preprocessing steps make it more suitable to be used; it handle some of the problems like missing data, imbalance, outliers and so on Simple algorithm models that are widely used are Logistic Regression, Decision Tree and so on they are easy, they are not as complex as a high complex model such as Gradient boosting and, neural networks they may be complex, but they are more accurate and more scalable. However, the decision in implementing a specific technique in the modeling also has to consider level of Sophistication of the technique in order to meet the regulatory and to converse the skepticism of the involved stakeholders.

It is traditional standard for verifying such models and guarantees that corresponding predictions are not over-optimistic; Accuracy or $(1 - \text{error rate})$; F statistic; ROC AUC or area under the receiver operating characteristic curve. But there is another aspect, and more importantly, it is connected with the implementation of the ethical issues including biasness or discrimination. Such concerns are being met by fairly recent developments in explainable AI, real time processing, along with federated learning to give the predictive modeling a missing ethical compass. The new innovations based on those listed above have been made possible by the recent advancement in elements of deep learning and ensemble techniques to improve the predictive models. For example, SHAP and LIME elucidate what the algorithm takes and real-time and a mixed approach make the risks always and variably evaluated by the institutions. Moreover, the offered solutions such as federated learning do not disclose a customer's data and correspond to the legislation norms. The bureaucrats hammer home the point that insofar as the territories of finances are changing, the exemplary predictive scales will remain necessary and as faculties of the risk management and the customer relation. In addition to fortifying their protection, good ethicalities represent the customer confidence and customer loyalty along with the dangers which are eliminated by utilizing the strong technologies by the financial institutions. Finally I posit that predictive modeling may potentially claim one of the most effective approaches toward fostering the flexibility, innovation and resilience for the financial services sector.

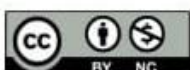
REFERENCES

1. Gilbert R, Widom CS, Browne K, Fergusson D, Webb E, Janson S. Burden and consequences of child maltreatment in high-income countries. *Lancet* 2009; 373(9657):68–81.
2. Anda RF, Butchart A, Felitti VJ, Brown DW. Building a framework for global surveillance of the public health implications of adverse childhood experiences. *Am J Prev Med* 2010; 39(1):93–8.
3. Putnam-Hornstein E, Webster D, Needell B, Magruder J. A public health approach to child maltreatment surveillance: evidence from a data linkage project in the U.S. *Child Abuse Rev* 2011; 20(4):256–73
4. WHO. Prevention of child abuse and neglect: making the links between human rights and public health. Geneva: Department of Injuries and Violence Prevention, WHO, 2001.
5. Mikton C, Butchart A. Child maltreatment prevention: a systematic review of reviews. *Bull WHO* 2009; 87(5):353–61.
6. Butchart A. Epidemiology: the major missing element in the global response to child maltreatment? *Am J Prev Med* 2008; 34(4S):S103–5?
7. Gilbert R, Fluke J, O'Donnell M, et al. Child maltreatment: variation in trends and policies in six developed countries. *Lancet* 2012; 379(9817): 758–72
8. Coulton CJ, Crampton DS, Irwin M, Spilsbury JC, Korbin JE. How neighborhoods influence child maltreatment: a review of the literature and alternative pathways. *Child Abuse Negl* 2007; 31(11–2):1117–42.
9. Stith SM, Liu T, Davies LC, et al. Risk factors in child maltreatment: a metaanalytic review of the literature. *Aggress Violent Behav* 2009; 14(1):13–29.
10. Prinz RJ, Sanders MR, Shapiro CJ, Whitaker DJ, Lutzker JR. Population-based prevention of child maltreatment: the U.S. Triple P System Population Trial. *Prev Sci* 2009; 10(1):1–12.
11. Bilukha O, Hahn RA, Crosby A, et al. The effectiveness of early childhood home visitation in preventing violence: a systematic review. *Am J Prev Med* 2005; 28(2S1):11–39?
12. Billings J, Dixon J, Mijanovich T, Wennberg D. Case findings for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ* 2006; 333:327.





13. Cabeças, A. (2014). Comportamento dos clientes da banca portuguesa em período de crise económica [Universidade Autónoma de Lisboa].
14. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-Step Data Mining Guide. DaimlerChrysler.
15. Cohen, D., Gan, C., Yong, H. H. A., & Chong, E. (2007). Customer Retention by Banks in New Zealand. *Banks and Bank Systems*, 2(1), 40–55. Customer Retention by Banks in New Zealand
16. Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95, 27–36.
17. Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications*, 34(14), 11751–11768.
18. Ejeu, E. (2021). Analysis of customer churn in Ugandan commercial banks: a case study of Stanbic Bank [College of Business and Management Sciences].
19. Hadden, J. (2008). A Customer Profiling Methodology for Churn Prediction [School of Applied Sciences].
20. Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). A Multi-Layer Perceptron Approach for Customer Churn Prediction. *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213–222.
21. Oluwatoyin, A. M., Misra, S., Wejin, J., Gautam, A., Behera, R. K., & Ahuja, R. (2023). Customer Churn Prediction in Banking Industry Using Power Bi (pp. 767–774).
22. Oyeniyi, A. O., & Adeyemo, A. B. (2015). Customer Churn Analysis in Banking Sector Using Data Mining Techniques. *African Journal of Computing & ICT African Journal of Computing & ICT Reference Format: A. O.*
23. Hui, S. H., Khai, W. K., XinYing, C., & Wai, P. W. (2023). Prediction of customer churn for ABC Multistate Bank using machine learning algorithms/Hui Shan Hon...[et al.]. *Malaysian Journal of Computing (MJoC)*, 8(2), 1602-1619.
24. Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 1-20. DOI: <https://doi.org/10.1080/10494820.2021.1928235>
25. Berk, R. (2012). Criminal justice forecasts of risk: A machine learning approach. Springer Science & Business Media. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
26. Chandar, M., Laha, A., & Krishna, P. (2006, March). Modeling churn behavior of bank customers using predictive data mining techniques. In *National Conference on Soft computing techniques for engineering applications (SCT-2006)* (pp. 24-26).
27. Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, 192, 105361. Cica, D., Sredanovic, B., Tesic, S., & Kramar, D. (2020). Predictive modeling of turning operations under different cooling/lubricating conditions for sustainable manufacturing with machine learning techniques. *Applied Computing and Informatics*. DOI: <https://doi.org/10.1016/j.aci.2020.02.001>
28. Costache, R., Arabameri, A., Moayed, H., Pham, Q. B., Santosh, M., Nguyen, H., & Pham, B. T. (2022). Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree.
29. Geocarto International, 37(23), 6780-6807. Costache, R., Hong, H., & Wang, Y. (2019). Identification of torrential valleys using GIS and a novel hybrid integration of artificial intelligence, machine learning and bivariate statistics. *Catena*, 183, 104179.
30. Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, 54(6), 1-25.
31. Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, 34, 11751-11768.
32. Guan, D., Yuan, W., Lee, Y. K., Najeebullah, K., & Rasel, M. K. (2014). A review of ensemble learning based feature selection. *IETE Technical Review*, 31(3), 190-198.
33. Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal of Applied Microeconometrics*, 1(2), 85-99.
34. Gumus, M., & Kiran, M. S. (2017, October). Crude oil price forecasting using XGBoost. In *2017 International conference on computer science and engineering (UBMK)* (pp. 1100- 1103). IEEE.
35. Ha Thanh, N., & Vy, N. (2022). Building a proper churn prediction model for Vietnam's mobile banking service. *International Journal of Advanced and Applied Sciences*, 9(7), 139-149.





36. He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia computer science*, 31, 423-430.
37. Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3), 515-524.
38. Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 18, 90-100.
39. Kaur, I., & Kaur, J. (2020, November). Customer churn analysis and prediction in banking industry using machine learning. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 434-437). IEEE.
40. Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 2(1), 1-13.
41. Khodabandehlou, S., & Rahman, M. Z. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, 19, 65-93.
42. Hou, L., Sharma, B. Y., Li, H., Chen, C., Li, Y., & Chen, H. (2018). Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Computer Methods and Programs in Biomedicine*, 153, 211-225.
43. Malhotra, D. K., Malhotra, K., & Malhotra, R. (2020). Evaluating consumer loans using machine learning techniques. In *Applications of Management Science*. Emerald Publishing Limited.
44. Malik, E. F., Khaw, K. W., & Chew, X. Y. (2022). A new hybrid data preprocessing technique for fraud detection prediction. *Computing and Informatics*, 41, 981-1001.
45. Marcelino, P., de Lurdes Antunes, M., Fortunato, E., & Gomes, M. C. (2021). Machine learning approach for pavement performance prediction. *International Journal of Pavement Engineering*, 22(3), 341-354.
46. Meza Ramirez, C. A., Greenop, M., Ashton, L., & Rehman, I. U. (2021). Applications of machine learning in spectroscopy. *Applied Spectroscopy Reviews*, 56(8-10), 733-763.
47. Oh, G., Song, J., Park, H., & Na, C. (2022). Evaluation of random forest in crime prediction: Comparing three-layered random forest and logistic regression. *Deviant Behavior*, 43(9), 1036-1049.
48. Patgiri, R., Varshney, U., Akutota, T., & Kunde, R. (2018, November). An investigation on intrusion detection system using machine learning. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1684-1691). IEEE.
49. Patro, K. K., Jaya Prakash, A., Jayamanmadha Rao, M., & Rajesh Kumar, P. (2022). An efficient optimized feature selection with machine learning approach for ECG biometric recognition. *IETE Journal of Research*, 68(4), 2743-2754.
50. Pisal, N. S., Abdul Rahman, S., Hanafiah, M., & Kamarudin, S. I. (2022). Prediction of life expectancy for Asian population using machine learning ALGORITHMS. *Malaysian Journal of Computing (MJoC)*, 7(2), 1150-1161.
51. Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1196-1201). IEEE.
52. Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, 2(7), 1-17.
53. Shafie, S., Ooi, S. P., & Khaw, K. W. (2023). Prediction of employee promotion using hybrid sampling method with machine learning architecture. *Malaysian Journal of Computing*, 8(1), 1264-1286.
54. Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019, May). A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1255-1260). IEEE.
55. Tékouabou, S. C., Gherghina, Ș. C., Touluni, H., Mata, P. N., & Martins, J. M. (2022). Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. *Mathematics*, 10(14), 2379.
56. Wang, M., & Chen, H. (2020). Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. *Applied Soft Computing*, 88, and 105946.
57. Yan, C., Liu, X., Boota, M. W., & Pan, Z. (2022). A Comparison of Machine Learning Methods Applied to the Automated Selection of River Networks. *The Cartographic Journal*, 59(3), 187-202.
58. Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2020). Random forest prediction intervals. *The American Statistician*, 74(4), 392-406.

