Literature Review

Machine Learning for Customer Churn Prediction: A Review of Models, Data Challenges, and Interpretability

Introduction

Customer churn refers to the number of customers lost over a given period. Understanding and predicting customer churn is essential for companies to understand customer satisfaction and implement strategies to retain customers (O'Brien and Downie, 2024).

Depending on the industry the business is set, acquiring a new customer can cost between 5 and 25 times more than customer retention (Gallo, 2014). Companies that have revenue based on subscription services such as telecommunication, insurance, streaming services and banking can be greatly affected by customer churn as their revenue and profit margins are directly affected by fixed and regular payments (Geiler, Affeldt and Nadif, 2022).

Machine Learning (ML) is a branch of Artificial Intelligence and is widely used in customer churn prediction. ML is great at learning patterns from complex and large sets of data to make predictions and is useful in industry for identifying early warning signs of customer churn autonomously (Pecan, 2024).

This review evaluates existing literature in ML algorithms used for churn prediction and how their corresponding models perform. It also explores strategies to combat class imbalance when training models and the importance of model interpretability in churn prediction applications.

ML Models for Churn prediction

Customer Churn prediction is a binary classification machine learning problem, and the following list contains common examples of algorithms that are used to create models for this use case.

- Logistic regression (LR)
- Support Vector Machines (SVM)
- Naïve Bayes Classification (NB)
- K-Nearest Neighbors (KNN)
- Decision Trees (DT)

- Random Forest (RF)
- Gradient boosting (AdaBoost, XGBoost)
- Neural Networks

Evaluation Metrics

Once the model is trained, its performance needs to be evaluated on a test set of data using suitable evaluation metrics. As churn prediction can be addressed as a binary classification problem, where customers that churn are identified as positive (1) and customers that do not churn are identified as negative (0), the following metrics are used (Databricks, 2020).

- **Confusion Matrix**

  This is a table that summarizes how successful the model is predicting classes, an example is shown in Table 1. Other metrics are based on this table.

	Churn (predicted)	Does not Churn (predicted)
Churn (actual)	True Positive (TP)	False Negative (FN)
Does not Churn (actual)	False Positive (FP)	True Negative (TN)

Table 1 Example of a confusion matrix layout.

- **Accuracy** shows the proportion of correctly classified examples and is useful when errors in predicting all classes are of equal importance.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** is the ratio of correct positive prediction to all positive predictions.

$$precision = \frac{TP}{TP + FP}$$

- **Recall** is the ratio of correct positive predictions to the total number of positive examples in the data (Burkov, 2019).

$$recall = \frac{TP}{TP + FN}$$

- **F1-score** is the harmonic mean of precision and recall, ranging from 1 and 0 (scikit-learn, 2019).

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

- **Area Under the Curve (AUC)** is the area represented under a Receiver Operating Characteristic (ROC) curve. It represents the probability that a model will rank a positive example higher than a negative example, with a theoretically perfect model having an AUC value of 1.0 (Google Developers, 2019).

Model Performance Comparisons

Many studies have been done comparing the performance of different ML algorithms for churn prediction.

Pulkundwar et al. (2023) examined multiple machine learning algorithms for customer churn, comparing accuracy score and training time. In the study it was found that gradient boosting algorithms such as AdaBoost and XGBoost, Random Forest and Decision Trees performed the best, striking a good balance between accuracy and training time. The Artificial Neural Network compared to these provided the best accuracy, however, took a substantial amount of extra time to train.

Vafeiadis et al. (2015) compare boosted and non-boosted machine learning techniques for churn prediction in the Telecomm industry focusing on the F1-Score. In non-boosted tests, the neural network, DT and SVM classifiers performed the best with F1-scores of 77.48%, 77.04% and 73.16% respectively. When applying AdaBoost.M1 algorithm on these classifiers, the F1-scores for the neural network, DT and SVM increased to 80.97%, 83.87% and 84.22% respectively.

Khodabandehlou and Zivari Rahman (2017) compared the performance of boosting, bagging and simple version of an Artificial Neural Network (ANN), SVM and Decision Trees algorithm. Overall Boosting versions outperformed the other versions, and out of these the ANN got an Accuracy of 97.07% and F1-score of 97.92%, with SVM having 93.48% and 95.37% respectively. DT performed the worst in this study although not poorly.

Ahmad et al. (2019) further contribute by comparing Decision Tree, Random Forest, Gradient Boost Machine Tree (GBM) and XGBoost algorithms. XGBoost and the GBM algorithm performed the best with AUC results of 93.3% and 90.89% respectively.

From fundamental algorithms LR, KNN and NB are weakest in the application of churn prediction while decision trees and SVM algorithms perform reasonably. All studies agree that boosting algorithms considerably improves their performance with churn prediction and gradient boosting algorithms such as AdaBoost, XGBoost and GBM provide the best results.

Neural Networks also perform well although training times were longer and therefore require more computing power.

Dealing with Imbalanced datasets

In churn prediction the dataset classes are normally imbalanced, as churn is generally a rarer instance than not, the churn class is however the class of most interest. The class with the most data is normally called the majority class, and the class with the least data is called the minority class.

Imbalanced data in classification modelling creates challenges, this is because most ML algorithms are designed around the assumption of equal distribution of classes in the data. The model therefore struggles to learn the characteristics of the minority class leading to poor predictive performance (Brownlee, 2020a).

Various techniques can be used to overcome class imbalance such as over-sampling, under-sampling and selecting algorithms less sensitive to the problem (Burkov, 2019).

Sampling methods

Under-sampling aims to equalize the data classes by eliminating instances of the majority class, these instances are normally eliminated randomly. Over-sampling aims to increase the importance of the minority class by creating new samples, popular methods are Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling Approach (ADASYN) (Fernández et al., 2018).

Cost-sensitive learning

Cost-sensitive learning methods work by associating greater cost with false negatives than with false positives, this improves performance of the model with respect to the positive minority class. It is sometimes implemented in algorithms where the class weighting can be tuned as a parameter in which a higher weight is given to the minority class, such as decision trees (Burez and Van den Poel, 2009).

Algorithm selection and techniques

Some algorithms are better suited to problems with imbalanced datasets, so selecting and comparing these can be beneficial in combination to other methods.

Decision trees in combination with cost-sensitive learning is normally a good approach to imbalanced datasets. Boosting algorithms such as AdaBoost and XGBoost are also ideal as through training iterations higher weight is given to the minority class (Feki, 2022).

One-class classification techniques can also be used where the negative case (majority class, non-churn) can be treated as normal and the positive case (minority, churn) can be treated as an outlier such as One-Class SVM algorithm (Brownlee, 2020b).

Performance of different techniques

Shumaly et al. (2020) compared customer churn prediction performance using random over-sampling, random under-sampling and SMOTE techniques. In this study it was shown that in most cases random over-sampling produced the best AUC results which produced an AUC of 90.1% for gradient boosting algorithm, random under-sampling gave the next best results with an AUC of 89.2%.

Burez and Van den Poel (2009) compared sampling, boosting and cost-sensitive learning techniques. In this study under-sampling was shown to improve prediction AUC and weighted Random Forests performed much better than the base model. It was stated however at the end of the study that there is no general answer to class imbalance and that improving prediction is case dependent.

Ahmad et al. (2019) compared XGBoost, GBM, RF and DT algorithms with no balancing as well as over-sampling and under-sampling. It was found that XGBoost and GBM had better performance overall even with no balancing techniques used, but for RF and DT improvements were seen when the balancing techniques were implemented.

Interpretability and Explainability

Understanding why models make the predictions they do helps to evaluate the model's performance and build trust in its deployment to make data driven decisions for businesses (Ribeiro, Singh and Guestrin, 2016).

The aim of customer churn prediction is to be able to accurately and preemptively predict customers at risk of churning, however another key aspect for a business is interpretability of the model's decision making. Being able to explain the model helps to identify the most impactful features indicating risk of churn.

Some algorithms such as linear regression and decision trees are inherently interpretable, however other models such as neural networks and ensemble methods are harder to explain and are sometimes referred to as "black box" models. Explanation mechanisms can then be implemented in these cases to provide better insight (Caigny, W. De Bock and Verboven, 2024).

Explainable Artificial Intelligence (XAI) refers to these mechanisms to provide increased ML transparency and some popular methods are SHAP and LIME, both identifying the most important and influential factors in a model's prediction through different means (Özkurt, 2025).

While XAI methods can help to explain more complex "black box" models, the simpler models with inherent explainability usually offer increased interpretability as all parameters are visible, whereas XAI methods analyze the inputs and predictions made by the model (Roelenga, 2021).

Discussion

Ensemble models and Neural Networks seem to produce the highest performing models with regards to F1-score and AUC metrics, however due to their complex nature they are less interpretable. For churn prediction, interpretation of model predictions to understand reasons for customer churn help businesses compile targeted strategies to retain these at-risk customers. XAI techniques help bridge the gap to explain these more powerful models.

While ensemble and boosted algorithms are generally performing higher and are less resource intensive than Neural Networks, some studies showed that decision trees with boosting combined with class balancing techniques can perform very well. This could give better transparency for business stakeholders without sacrificing much performance.

Conclusion

Few studies comprehensively examine all three aspects of model performance, class imbalance and interpretability together, highlighting a significant research gap.

Interpretability vs. Performance

Gradient Boosting algorithms such as XGBoost and AdaBoost consistently performed better than classical ML algorithms in examined studies on churn prediction applications, Neural networks have high performance however require a significant amount more computing resource to train efficiently. Improvements need to be made in the transparency of these algorithms and stakeholders need to closely examine how much transparency is needed in their application.

Future Research Directions

Decision Trees was the highest performing classical model, especially when class imbalance techniques such as SMOTE, over-sampling or under-sampling were implemented. There is not much research into the effect of these techniques on the transparency of models, which could be explored further in the future.

Current research is investigating hybrid models to increase transparency and model performance such as the study by Caigny, W. De Bock and Verboven (2024). Further research in this field would be beneficial.

References

O'Brien, K. and Downie, A. (2024). Customer churn. Available at: https://www.ibm.com/think/topics/customer-churn [Accessed 15 Jun. 2025].

Gallo, A. (2014). The Value of Keeping the Right Customers. Available at: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers [Accessed 15 Jun. 2025].

Geiler, L., Affeldt, S. and Nadif, M. (2022). A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, 14. doi:https://doi.org/10.1007/s41060-022-00312-5.

Pecan (2024). We Used AI to Predict Customer Churn — Now What? Churn Reduction Initiatives. Available at: https://www.pecan.ai/blog/churn-reduction-strategies-prediction-playbook/ [Accessed 15 Jun. 2025].

Databricks (2020). Churn 03: Model Selection - Databricks. Available at: https://www.databricks.com/notebooks/churn/3-model-selection.html [Accessed 15 Jun. 2025].

Burkov, A. (2019). THE HUNDRED-PAGE MACHINE LEARNING BOOK. Andriy Burkov.

scikit-learn (2019). sklearn.metrics.f1_score — scikit-learn 0.21.2 documentation. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html [Accessed 15 Jun. 2025].

Google Developers (2019). Classification: ROC Curve and AUC | Machine Learning Crash Course. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc [Accessed 15 Jun. 2025].

Pulkundwar, P., Rudani, K., Rane, O., Shah, C. and Virnodkar, S. (2023). A Comparison of Machine Learning Algorithms for Customer Churn Prediction. doi:https://doi.org/10.1109/icast59062.2023.10455051.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.Ch. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55(55), pp.1–9. doi:https://doi.org/10.1016/j.simpat.2015.03.003.

Khodabandehlou, S. and Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), pp.65–93. doi:https://doi.org/10.1108/jsit-10-2016-0061.

Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1). doi:https://doi.org/10.1186/s40537-019-0191-6.

Brownlee, J. (2020a). A Gentle Introduction to Imbalanced Classification. Machine Learning Mastery. Available at: https://machinelearningmastery.com/what-is-imbalanced-classification/ [Accessed 16 Jun. 2025].

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. and Herrera, F. (2018). Learning from Imbalanced Data Sets. Cham: Springer International Publishing. doi:https://doi.org/10.1007/978-3-319-98074-4.

Feki, R. (2022). Imbalanced data: best practices. Medium. Available at: https://rihab-feki.medium.com/imbalanced-data-best-practices-f3b6d0999f38 [Accessed 16 Jun. 2025].

Brownlee, J. (2020b). *One-Class Classification Algorithms for Imbalanced Datasets*. Available at: https://machinelearningmastery.com/one-class-classification-algorithms/ [Accessed 16 Jun. 2025].

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), pp.4626–4636. doi:https://doi.org/10.1016/j.eswa.2008.05.027.

Sajjad Shumaly, Pedram Neysaryan and Guo, Y. (2020). Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees. doi:https://doi.org/10.1109/iccke50421.2020.9303698.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). ‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, pp.1135–1144. doi:https://doi.org/10.1145/2939672.2939778.

De Caigny, A. W. De Bock, K. and Verboven, S. (2024). Hybrid black-box classification for customer churn prediction with segmented interpretability analysis. Decision Support Systems, 181, pp.114217–114217. doi:https://doi.org/10.1016/j.dss.2024.114217.

Özkurt, C. (2025). Transparency in Decision-Making: The Role of Explainable AI (XAI) in Customer Churn Analysis. Information Technology in Economics and Business. doi:https://doi.org/10.69882/adba.iteb.2025011.

Roelenga, B. (2021). Think outside the ‘black’ box | Towards Data Science. Available at: https://towardsdatascience.com/think-outside-the-black-box-7e6c95bd2234/ [Accessed 16 Jun. 2025].