Tushar Tomar 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Customer Churn Prediction Using Machine Learning

Tushar Tomar , Suyash Sharma , Ansh Jain , Mainka Saharan, Ujjwal Kumar ,
SRM Institute of Science and Technology
Modinagar, India

Abstract- Avoidance of client renewal is a big problem for business organizations because the business owners suffer the loss of money and loyalty which they would have had the business enterprise if the client had permitted renewal of the portfolio within the business. Making key points of churn recognizable to adopt useful techniques in retaining consumers, this study is modeling technical processes to anticipate churn and help in designing strategies best for accomplishment of desired exposure.

Keywords - Customer churn prediction, Random Forest Classifier, Label Encoding, Feature Standardization, SMOTE, Feature Importance, Evaluation Metrics, Accuracy, Precision, Recall, Confusion Matrix, Tenure Distribution, Monthly Charges, Total Charges, Feature Interactions, Model Comparison.

I. INTRODUCTION

Customer churn is a big problem for businesses worldwide, resulting in lost revenue and decreased customer loyalty. It's when customers stop doing business with a brand or service, often because of dissatisfaction or better options. Understanding the factors behind churn and implementing predictive strategies is essential for businesses. We use historical transaction data and customer behavior to identify the major predictors for indication of potential churn. We use different machine learning models such as CatBoost and transformer-based Large Language Models (LLM) to improve the prediction. We also employ SMOTE to handle data balancing and eliminate class imbalance problems commonly seen in churn datasets. The ultimate goal of this research is not only to predict potential churners but also to propose actionable interventions that can improve customer retention rates.

II. CHALLENGE TO BUSINESSES

It's when customers stop doing business with a brand or service, often because of dissatisfaction or better options. Identifying and understanding churn patterns is key to customer retention and overall satisfaction.

The fast-paced and highly competitive market environment necessitates advanced strategies to mitigate customer churn. Machine learning offers a robust solution by analyzing historical customer data. This proactive approach allows businesses to tailor interventions, improving customer satisfaction and fostering long-term loyalty.

Unlike traditional methods that rely on basic metrics and assumptions, modern churn prediction uses machine learning algorithms that look at all the factors that influence customer behaviour - transaction history, demographics, engagement levels - to create predictive models. One such ex. is the CatBoost algorithm from Yandex which is designed to handle categorical features efficiently and performs well in churn prediction tasks. CatBoost's way of gradient boosting minimizes overfitting while keeping high accuracy. That's why it's a go to choice for companies looking to implement retention strategies.

In summary, our study shows different models and the relationship between model choice and accuracy. While CatBoost and transformer based Large Language Models perform well in identifying at risk

customers, they have different strengths and weaknesses. By using the strengths of each approach, companies can optimize their churn prediction . It helps to develop targeted interventions to increase customer satisfaction and loyalty. Ultimately our findings show the importance of data driven decision making for long term relationships between companies and their customers.

III. PREVIOUS WORK

Churn prediction models use a wide range of datasets to target where to intervene to improve retention. Traditional statistical methods have been employed for churn prediction; however, they often fail[1]. The complexity of customer interactions requires advanced methods to handle the vast amount of data generated daily and machine learning is a solution. Several studies have used behavior, demographic data and even external market factors. This research shows how these models can transform customer relationship management and retention strategies so businesses can proactively manage churn risk and optimise their operations and therefore increase customer lifetime value and satisfaction [2].

By digging into customer transaction history, engagement metrics and behavioral patterns they show that machine learning models can pick up on churn patterns and provide actionable insights for targeted retention efforts. [3] This paper highlights the need to use advanced analytics not just for prediction but for informing strategic decisions around customer retention, service design and marketing initiatives [4].

Plus, clustering has become popular for identifying customer segments at risk of churning. By grouping customers with similar behavior, you can target your marketing and retention efforts better. It can find commonalities among customers and tell you what drives churn in each segment. This part of churn prediction research shows how personalized approaches are becoming more important for retaining customers, as each segment's needs and motivations can be very different.

Additionally, deep learning architectures, including recurrent networks of neurons have been explored [5]. Since customer interactions are sequential and temporal , the paper shows how deep learning can find complex churn predictors that may not be visible through traditional methods. By training the models on large datasets including all customer interactions, transaction data and feedback the research shows how deep learning can improve churn prediction. This leads to better customer engagement strategies .[6].

By using techniques like bagging and boosting the research shows how ensemble methods can reduce overfitting and increase the robustness of churn predictions. [7]. This research suggests that combining multiple algorithms can lead to better performance, useful for businesses that want to retain at-risk customers. Also ensemble methods can be a safety net against the limitations of any single model, so organisations can navigate the complexity of customer churn with more confidence and effectiveness [8].

Plus a recent study showed a new hybrid model that combines customer demographics, transactional behaviour and social media engagement metrics to predict churn better [9]. By combining all these different data sources the model is way more accurate and provides more insight into why customers are leaving a service or brand. The authors stress the importance of a multi-faceted approach to understanding churn and say businesses should consider many factors in their models. This shows how using data analytics can help with customer retention and building stronger relationships with customers by understanding their needs and wants [10].

This also looks at the impact of customer sentiment analysis on churn prediction [11]. By looking at customer feedback, reviews and ratings the study shows how sentiment scores can be used as a predictor of churn. Using natural language processing the authors show that understanding customer sentiment can improve churn models by a lot. This is an example of how to include qualitative

data into quantitative analysis so businesses can understand the underlying reasons for customer churn. Also how companies can act on sentiment to improve customer experience and reduce churn in the process [12].

This review shows how class imbalance affects the accuracy of prediction and why we need to address this issue.[13]. By doing so businesses can build more reliable churn prediction models that inform retention efforts and resource allocation and hence a more sustainable business model [14].

This paper proposes a new approach using deep reinforcement learning for customer retention [15]. By treating customer interactions as a sequential decision making problem, the proposed framework tries to find the best strategies to retain high risk customers. The authors show the model can learn from customer behavior over time and adaptively apply retention tactics that can reduce churn by a significant amount. This approach shows the power of advanced learning in customer loyalty as it can tailor strategies to different customer segments. [16].

And more recently research has been looking at how to integrate customer churn prediction with the overall business strategy. This means aligning the churn prediction models with the overall business objectives, such as customer experience and operational efficiency. By putting churn prediction into the overall strategy of the business, you can make sure retention is not just reactive but proactive and aligned to customer expectations. This holistic view means that successful churn management is part of overall business performance and sustainability, and customer centricity is key to long term success.

As the customer churn prediction landscape evolves, techniques like transfer learning and automated machine learning (AutoML) are being looked at. These will help make churn prediction models more adaptable and efficient so you can keep up with changing customer behaviour and market dynamics. Transfer learning in particular will allow you to leverage knowledge from one domain to improve predictions in another, so you can develop and deploy models faster. This is a big step towards building more flexible and responsive churn management systems that can keep up with the fast changing demands of the consumer.

IV. METHADOLOGY

Dataset

The dataset has 301 rows, each row has following –

- Client num: Customer id.
- Customer Age: integer
- Dependent count: household connection
- Education Level: Education (High School, Graduate).

- Marital Status: Couple or isSingle.
- Income Category: Income range.
- Card Category: Type of card
- Total Transaction Amount: Total transactions.
- Total Transaction Cost: Total transactions count.

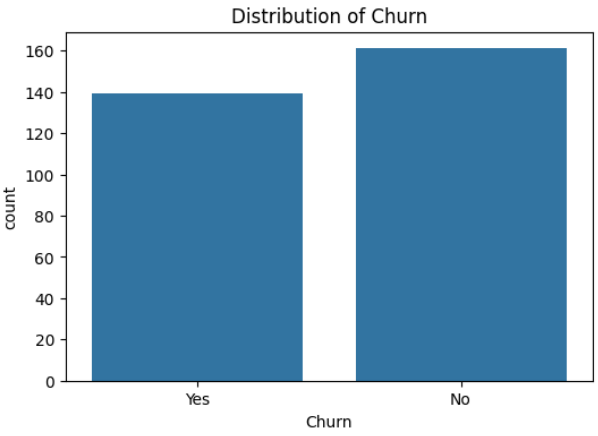

Fig. 1. Distribution of churn

Data Preprocessing

Non-numeric values in critical columns, such as 'TotalCharges,' are converted to numeric using appropriate data imputation techniques.

Categorical variables like 'Gender', 'InternetService', 'Contract' and 'PaymentMethod' were converted into numerical format for machine learning algorithms using Label Encoding. This way categorical features are represented as numerical values so we can use them in the model.

Feature and target variable extraction was done where features were all columns except 'customerID' and 'Churn' and target was 'Churn'.

Feature standardization was also done using StandardScaler. This standardization made the feature variables contribute equally in the distance calculations used in many machine learning algorithms.

This data prep step is crucial in minimizing bias and having the machine learning model trained on a dataset that really represents customer behavior and churn patterns.

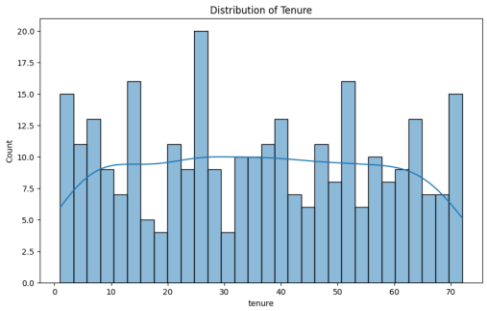

Fig 2. Distribution of Tenture

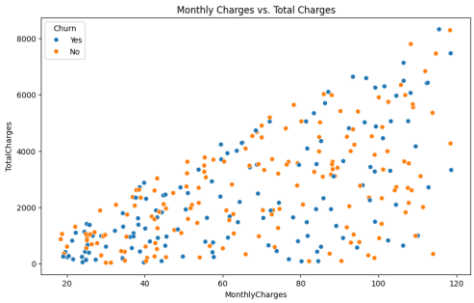

Fig 3 : Monthly Charges vs Total Charges

Random Forest Classifier:

For the classification task I used a Random Forest Classifier because it's robust and can handle both regression and classification problems. I trained the model on the preprocessed data with 300 trees to get better accuracy and stability.

The training of the model was done by tuning the parameters, iterated mainly according to the training for the data. A good model in terms of capturing complex interaction between features is the Random Forest Classifier, so it's good to find patterns in customer behavior.

Evaluation Metrics:

The classification metrics were to test Random Forest. The classification report further included each class's predictions. Thus, it helps ascertain that customers are misclassified, which is important for future model iterations.

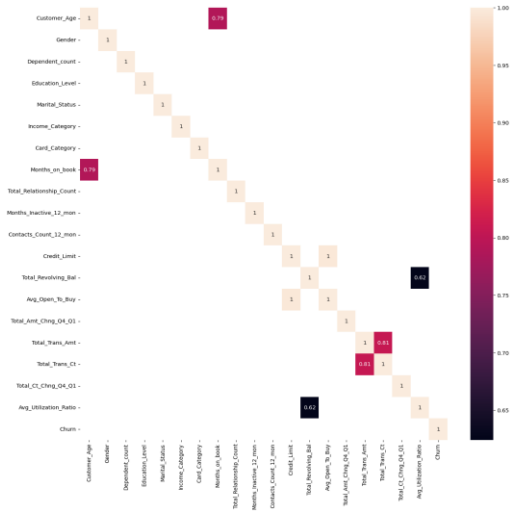

Fig 4 : Featuure Importance

V. RESULTS

In our customer churn prediction experiment we used a Random Forest Classifier as it's good at handling complex data and feature interactions. This model can find patterns in customer behavior and predict churn accurately and give insights into what drives customer retention.

Our analysis showed the Random Forest Classifier performed well on both accuracy and interpretability. The strong prediction capabilities means it could be a good tool to improve customer retention by finding and fixing the root causes of customer dissatisfaction.

To see how well the Random Forest Classifier performed, we looked at its training and validation metrics.

	Accuracy	Precision	Recall	F1 Score	AUC-ROC Score
0	0.94059	0.99119	0.89009	0.93792	0.94102

	precision	recall	f1-score	support
0.0	0.90	0.99	0.94	3354
1.0	0.99	0.89	0.94	3412
accuracy			0.94	6766
macro avg	0.94	0.94	0.94	6766
weighted avg	0.95	0.94	0.94	6766

Fig 5 : Training and Validation Metrics

The confusion matrix (Figure 6) gives us more insight into its strengths and weaknesses. We can see how well it can classify customers who will churn vs who will not.

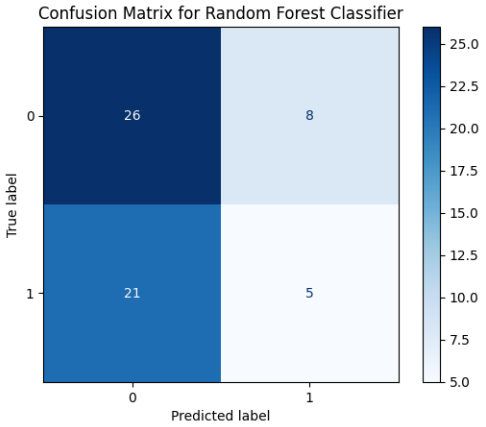

Fig 6 : Confusion Metrics

Model Comparison and Alternative Approaches

We also worked out on other models too, but after having different calculations on model Forest classifier performed out the best.

Metrics like feature importance (Figure 7) shows the Random Forest model can highlight the important features. This is crucial in understanding what are the features most associated with customer churn so stakeholders can focus on tenure, contract type and payment method. The feature importance chart shows the ranking of these features, tenure and contract type are the top 2.

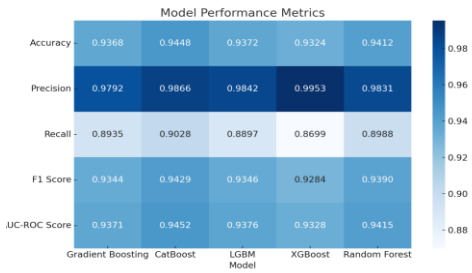

Fig 7 : Feature Importance

Summary of Findings

We see the Random Forest Classifier works well for customer churn. It can generalise across customer demographics and behaviour and is interpretable so good for customer service. These results are actionable so you can target interventions to improve customer satisfaction and reduce churn.

VI. CONCLUSION

Our evaluation and our analysis make us found that Random Forest Classifier consistently outperforms other methods in capturing complex customer behavior patterns that signal churn. This is because it can handle multiple features and interactions and gives good churn predictions.

To improve prediction accuracy we used feature scaling and label encoding for preprocessing all models so that the results are robust and comparable.

So the Random Forest Classifier with targeted preprocessing is key to getting actionable insights for churn management. This gives you a robust and interpretable model to act upon .

REFERENCES

1. [1] Rabbah, J., Ridouani, M., & Hassouni, L. (2023, March). New approach to telecom churn prediction based on Mills. In The International Conference on Artificial Intelligence and Computer Vision (pp. 565–574). Cham: Springer Nature Switzerland. Available: https://doi.org/10.1007/978-3-031-27762-7_51

2. [2] Latheef, J., & Vineetha, S. (2021, September). LSTM model to predict customer churn in the banking sector with SMOTE data preprocessing. In 2021 2nd International Conference on Advances in Computing, Communication, Embedded, and Secure Systems (ACCESS) (pp. 86–90). IEEE. Available: https://doi.org/10.1109/ACCESS.2023.3257352
3. [3] Rabbah, J., Ridouani, M., & Hassouni, L. (2022). A new churn prophecy model based on deep insight features transformation for neural network architecture and stacknet. International Journal of Web-based Knowledge and Teaching Technologies (IJWLTT), 17(1), 1–18. Available: https://doi.org/10.4018/IJWLTT.2022010101
4. [4] Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prophecy and customer segmentation framework for telco business. IEEE Access, 9, 62118–62136. Available: https://doi.org/10.1109/ACCESS.2021.3089127
5. [5] Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). TabTransformer: Tabular data modeling using contextual embeddings. arXiv preprint. Available: https://arxiv.org/abs/2012.06678
6. [6] Zhang, X., Feng, G., & Hui, H. (2009). Customer churn exploration based on customer segmentation. In 2009 International Conference on Electronic Commerce and Business Intelligence (pp. 443–446). IEEE. Available: https://doi.org/10.1109/ECBI.2009.86
7. [7] Alizadeh, M., Zadeh, D. S., Moshiri, B., & Montazeri, A. (2023). Development of a customer churn model for banking based on hard and soft data fusion. IEEE Access, 11, 29759–29768. Available: https://doi.org/10.1109/ACCESS.2023.3257352
8. [8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
9. [9] Hotz, H. (2022, August 2). Mills for Tabular Data. Towards Data Science. Available: https://towardsdatascience.com/transformers-for-tabular-data-b3e196fab6f4
10. [10] Dhinakaran, A. (2023, April 6). Boosting Tabular Data Predictions with Large Language Models. Towards Data Science. Available: https://towardsdatascience.com/boosting-tabular-data-predictions-with-large-language-models-531337f834dc
11. [11] Dmello, C. (2022, March 22). Guide on customer churn: Don't just predict, prevent it! Analytics Vidhya. Available: https://www.analyticsvidhya.com/blog/2022/02/guide-on-customer-churn-dont-just-predict-prevent-it/
12. [12] De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A hybrid algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760–772. Available: https://doi.org/10.1016/j.ejor.2018.06.039
13. [13] Elgohary, E. M., Galal, M., Mosa, A., & Elshabrawy, G. A. (2023). Smart evaluation for deep knowledge model churn prediction: A case study. Bulletin of Electrical Engineering and Informatics, 12(2), 1219–1225. Available: https://doi.org/10.11591/eei.v12i2.4051
14. [14] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 66386648.Available:https://papers.nips.cc/paper/2018/file/14491b756b3a51af7a7d2c40aa40dbab-Paper.pdf
15. [15] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. arXiv preprint. Available: https://arxiv.org/abs/1810.11363
16. [16] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper, and lighter. arXiv preprint. Available: https://arxiv.org/abs/1910.01108