ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar416

Comparitive Analysis of Gradient Boosting and Transformer Based Models for Binary Classification in Tabular Data

A Customer Churn Case Study

Jebaraj Vasudevan[1]

[1]Visa Inc., Atlanta, GA, USA

Publication Date: 2025/03/20

Abstract: This study compares the classification performance of the Gradient Boosting (XGBoost), and Transformer based model with multi-head self-attention for Tabular Data. While the methods exhibit broadly similar performance, the Transformer model particularly excels in Recall by about 8% showing that it would be better suited to applications such as Fraud Detection in Payment processing and Medical Diagnostics.

Keywords: *Transformer, Gradient Boosting, XGBoost, Tabular Data.*

How to Cite: Jebaraj Vasudevan (2025). Comparitive Analysis of Gradient Boosting and Transformer Based Models for Binary Classification in Tabular Data. *International Journal of Innovative Science and Research Technology*, 10(3), 466-470. https://doi.org/10.38124/ijisrt/25mar416

I. INTRODUCTION

Tabular data is ubiquitous in industry because it is inherently structured, easily interpretable, and compatible with a wide range of analytical and reporting tools. Its organization in rows and columns simplifies the process of data storage, retrieval, and manipulation, which is why relational databases, spreadsheets, and data warehouses predominantly use this format.

Industries such as finance, healthcare, retail, telecommunications, and manufacturing heavily rely on tabular data. In finance, for instance, transaction records, market data, and risk assessments are typically stored in structured tables, facilitating quantitative analyses and regulatory reporting. In healthcare, patient records, laboratory results, and treatment histories are maintained in tabular formats to support clinical decision-making and research. Retail and e-commerce sectors use tabular data for inventory management, sales tracking, and customer behavior analysis, while telecommunications companies employ it for billing, service usage, and churn prediction.

The prevalence of tabular data across these sectors highlights its role in enabling robust, data-driven decision-making and operational efficiency. Its simplicity and versatility make it a cornerstone of analytical workflows in both traditional and modern digital enterprises.

This case study shows a comparative analysis of XG Boost [1] and Tab Transformer [2], two of the most popular supervised learning algorithms for Tabular data. We chose the task of evaluating their performance on a Binary Churn prediction problem using the Telco Customer Churn data [3]. The algorithms exhibit a similar level of performance on multiple classification metrics while the Tab Transformer outperforms the XG Boost on Recall by +8%.

Comparing XG Boost and Tab Transformer reveals distinct methodologies that cater to different aspects of tabular data modeling. XG Boost, a gradient boosting framework, is lauded for its efficiency in handling structured data. It builds ensembles of decision trees using gradient statistics and regularization, resulting in robust models that mitigate overfitting and offer clear interpretability. This algorithm has been refined over years and is widely adopted in industry and research due to its computational speed and ease of deployment. In contrast, Tab Transformer harnesses the power of transformer architectures originally designed for natural language processing. By applying self-attention mechanisms, Tab Transformer captures complex, non-linear interactions among features, providing a deep representation of data relationships. While XG Boost excels in scenarios where model transparency and speed are paramount, Tab Transformer demonstrates potential in situations with intricate feature dependencies that require nuanced contextual understanding. The choice between these methods depends on the problem domain, computational resources, and the need for model interpretability versus expressive power. Both approaches offer complementary strengths; combining them might even enhance performance in hybrid systems. Ultimately, their continued development reflects the dynamic evolution of machine learning techniques for structured data analysis. This

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar416

comparative review highlights the importance of aligning algorithm selection with specific data challenges

II. SCOPE

➢ *XGBoost*

XG Boost is a highly efficient, scalable gradient boosting algorithm that has revolutionized machine learning practices across various domains. It constructs an ensemble of decision trees in a sequential manner, optimizing each new tree based on the residual errors of previous iterations. By employing both first-order and second-order gradient statistics, XG Boost effectively minimizes loss functions while integrating regularization techniques to prevent overfitting. This algorithm is well-known for its speed and performance, especially on large and complex datasets. Its implementation supports parallel processing and distributed computing, enabling the analysis of massive datasets with ease. Additionally, XG Boost provides robust handling of missing values and sparse data through innovative approaches such as weighted quantile sketch. The framework is highly customizable, accommodating various objective functions, including regression, classification, and ranking. As a result, it has become a favored choice in data science competitions and industry applications. With a strong emphasis on interpretability and computational efficiency, XG Boost has significantly contributed to the advancement of predictive analytics and remains a critical tool for researchers and practitioners aiming to extract meaningful insights from data. Furthermore, its design enables seamless integration with various programming languages and data processing libraries, making it a versatile solution for research and industry applications.

➢ *TabTransformer*

Tab Transformer is an innovative neural architecture designed specifically for tabular data analysis by leveraging the principles of transformer models. It extends the self-attention mechanism, which is central to transformers, to capture intricate relationships among features in structured datasets. Transformers, initially introduced for natural language processing, utilize multi-head self-attention to assess the significance of each input element, regardless of their order. In Tab Transformer, categorical features are first transformed into dense embeddings, which are then processed through a series of transformer layers. These layers enable the model to learn complex, non-linear interactions among variables, facilitating superior feature representation. The self-attention mechanism allows the model to dynamically weigh contributions from different features, thus enhancing predictive accuracy and robustness. Moreover, the architecture seamlessly integrates with traditional deep learning frameworks, making it adaptable to various data science tasks. By combining the strengths of transformer architectures with specialized adaptations for tabular data, Tab Transformer offers a novel approach to overcome limitations of conventional methods. Its design represents a convergence of ideas from natural language processing and structured data modeling, offering promising potential in fields requiring high interpretability and performance. This approach not only enhances model efficiency but also paves the way for future innovations in data representation.

➢ *Data*

Telco Customer Churn Data from Kaggle contains real-world data collected from a telecommunications company, capturing various aspects of customer behavior and account characteristics. The dataset includes demographic details, account information, service subscriptions, billing data, and usage metrics. The primary target variable is a binary indicator representing whether a customer has discontinued their service ("Churn"), making it a popular benchmark for binary classification tasks focused on customer attrition.

The dataset's structure—with a mix of categorical features (e.g., gender, contract type, payment method) and numerical features (e.g., tenure, monthly charges, total charges)—requires robust preprocessing and feature engineering. Researchers and practitioners have leveraged this dataset to test various data transformation and modeling approaches, as its inherent challenges, such as handling missing values and imbalanced classes, reflect real business scenarios.

Due to its practical significance, the Telco Customer Churn dataset is frequently used in both academic studies and industrial applications. It helps organizations develop predictive models aimed at understanding and mitigating churn, ultimately supporting customer retention strategies through data-driven insights.

III. IMPLEMENTATION

The Tab Transformer [2] is organized into three principal components: a dedicated column embedding layer, a succession of N Transformer layers, and a concluding multilayer perception. Each Transformer layer, as described by [4] integrates a multi-head self-attention mechanism that dynamically models inter-feature dependencies, followed by a position-wise feed-forward network that refines the learned representations. This configuration facilitates the extraction of complex interactions within categorical data while seamlessly integrating numerical inputs, ultimately enhancing predictive performance on tabular datasets.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar416

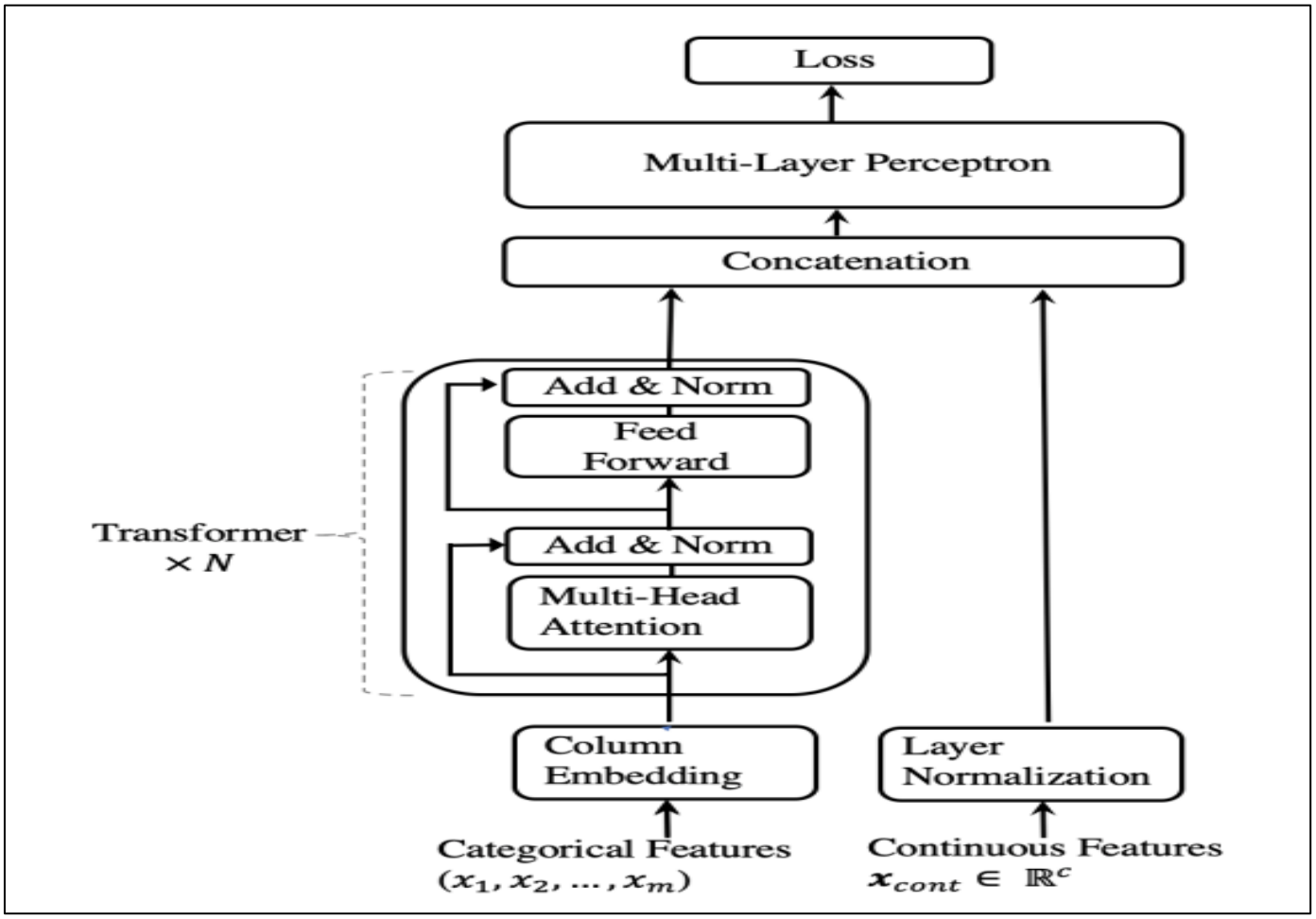

Fig 1 Tab Transformer Architecture [2]

➢ *Forward Pass*

$$L(x, y) \equiv C\left(g_\psi\left(f_\theta\left(E_\varphi(x_{cat})\right), x_{cont}\right), y\right) \quad (1)$$

- *Embedding Categorical Inputs*

In the forward method, each column of the categorical input x_{cat} is passed through its corresponding embedding layer. These embeddings are stacked along a new dimension to form a tensor of E with shape (batch, num_cat, embed_dim).

- *Embedding Categorical Inputs*

The stacked embeddings E are passed through the transformer encoder. This layer applies multi-head self-attention (explained in detail below), allowing the model to learn complex interdependencies between different categorical features

- *Concatenation and Prediction*

The output E′ from the previous layer is flattened to a vector and concatenated with the numerical features $x_{cont} \in \mathbb{R}^c$ denotes all the c continuous features. The resulting vector is processed by the MLP to yield the final prediction logits

For our classification task, let C be the cross-entropy for and we want to minimize the following loss function L(x, y) to learn all the parameters in an end-to-end learning gradient descent. The Tab Transformer parameters include φ for column embedding, θ for Transformer layers, and ψ for the top MLP layer.

➢ *Muti-head Self Attention*

In the formulation presented by [4], the Transformer architecture is structured around a multi-head self-attention mechanism followed by a position-wise feed-forward network, with both sub-layers augmented by residual connections and layer normalization. The self-attention mechanism operates via three learnable projection matrices—namely, Key, Query, and Value. Each input embedding is projected onto these matrices to produce its corresponding key, query, and value vectors. Formally, let $K \in \mathbb{R}^{m\,x\,k}, Q \in \mathbb{R}^{m\,x\,k}, V \in \mathbb{R}^{m\,x\,v}$ denote the matrices containing the key, query, and value vectors for m input embeddings, where k , v represent the dimensions of the key and value vectors, respectively. Each embedding then computes attention over all embeddings via an attention head defined by

$$Attention(K, Q, V) = A \,.\, V \quad (2)$$

with the attention weights are given by,

$$A = softmax\left(\frac{QK^T}{\sqrt{k}}\right) \quad (3)$$

Here, the matrix $A \in \mathbb{R}^{m\,x\,m}$ quantifies the degree to which each embedding attends to every other embedding, thereby producing contextually enriched representations. Following the attention operation, the output—originally of

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar416

dimension v is re-projected to the embedding dimension d via a fully connected layer. This is then processed sequentially by two position-wise feed-forward layers, where the first layer expands the dimensionality to four times the original size and the second layer subsequently reduces it back to d.

IV. ANALYSIS

➢ *Feature Engineering*

We used several new features to enhance model performance by providing additional context and capturing non-linear relationships within the data. Below is an explanation of the key engineered features and their potential impact:

- *Average Monthly Charge*

$$AvgMonthlyCharge = \frac{TotalCharges}{Tenure}$$

This feature normalizes the total spending by the length of the customer's relationship, highlighting customers who incur higher charges relative to their engagement duration. It may indicate dissatisfaction or financial stress, both of which can correlate with churn.

- *Service Count*

By summing binary indicators for various service features (e.g., OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies), we created a feature:

$$ServiceCount = \sum_{i=1}^{n} 1\,(Service_i = "Yes")$$

This aggregation provides a measure of customer engagement with additional services, which can be a proxy for loyalty. A higher count may imply a deeper investment in the ecosystem, potentially reducing churn risk.

- *Tenure Binning*

Instead of using the continuous tenure variable directly, we segmented tenure into categorical bins (e.g., 0–12 months, 13–24 months, etc.). This transformation captures non-linear effects, as churn likelihood may change drastically at different stages of a customer's lifecycle.

- *Interaction Features*

We explored interaction terms such as the product of Monthly Charges and Contract type, which can reveal combined effects where, for example, high charges paired with a month-to-month contract might be a stronger churn signal than either feature in isolation.

These engineered features enrich the dataset by providing more nuanced signals for the learning algorithms. For XG Boost, the additional numerical variables enhance tree-splitting decisions, while for Tab Transformer, they offer extra context that complements the embedded representations of categorical data. Overall, these features aim to improve the models' ability to detect subtle patterns and relationships that contribute to customer churn.

➢ *Methodology and Metrics*

Both the models were trained using the same set of features and the training was stopped as soon as the loss of the unseen data did not improve (early stopping)

The models are compared using several performance metrics in **Error! Reference source not found.** that provide a comprehensive view of their classification abilities. These include:

- Accuracy: Measures the overall proportion of correct predictions.
- Precision: Evaluates the correctness of positive predictions, indicating how many predicted positives are true positives.
- Recall (Sensitivity): Assesses the model's ability to identify all actual positive cases.
- F1 Score: The harmonic mean of precision and recall, offering a balance between them.
- Area Under the ROC Curve (AUC): Captures the trade-off between true positive and false positive rates across different thresholds.

Table 1 Metrics Comparing the Model Performance on Unseen Data

	Accuracy	Precision	Recall	F1	AUC
XGBoost	79.4%	64.3%	50.2%	56.5%	84.1%
TabTransformer	79.5%	63.1%	54.8%	58.6%	83.6%

As evident from the table shown above, the models have very similar overall performance similar to what [2] had also noticed in their results. But what we also see here is that the Transformer model outperforms the Boosting method in Recalling the positive examples by about 8%. So, in scenarios, when the cost of missing a true positive far outweighs the inconvenience or cost of incorrectly flagging a negative instance as positive. For instance, in medical diagnostics—such as screening for cancer or infectious diseases—failing to identify a diseased patient (a false negative) can have severe or even fatal consequences, whereas a false positive might lead to further testing that, while potentially anxiety-inducing and costly, is comparatively less harmful. Similarly, in fraud detection systems, overlooking a fraudulent transaction could result in substantial financial loss, making it preferable to flag more transactions for review even if some are false alarms. In these circumstances, the Transformer based model can be preferred over the Gradient Boosting XG Boost.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar416

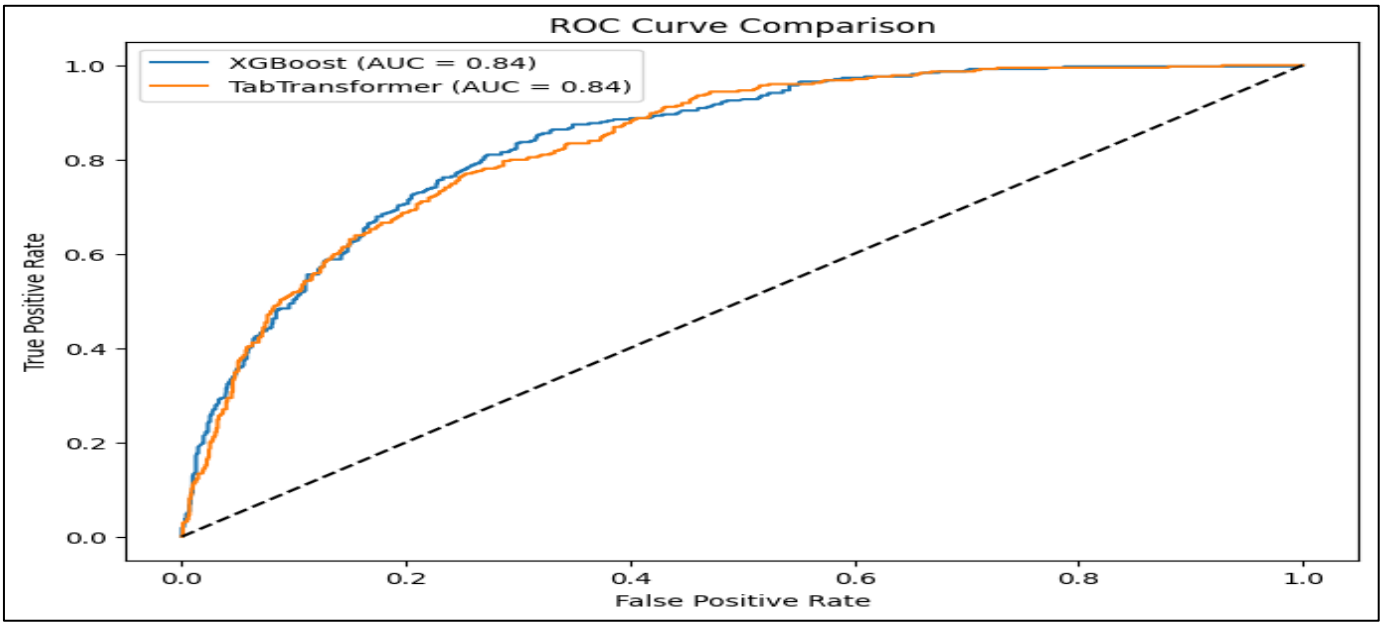

Fig 2 Roc Curve

Additionally, ROC curves as shown in Fig 2 are plotted to visually analyze the distribution of classification errors and to assess model discrimination capabilities. These combined metrics allow for a detailed scientific comparison between the XGBoost and TabTransformer models, highlighting strengths and potential trade-offs in different aspects of performance.

V. CONCLUSTION

In conclusion, this study provides a comprehensive comparative analysis of Gradient Boosting (XGBoost) and Transformer-based models for binary classification in tabular data. Both models exhibit similar performance across various metrics, with the Transformer model demonstrating a notable advantage in recall. This suggests that the Transformer model may be better suited for applications where the cost of false negatives is high, such as fraud detection and medical diagnostics. The findings underscore the importance of aligning model selection with specific data challenges and application requirements. Future research could explore hybrid approaches that combine the strengths of both models to further enhance performance. Overall, this study contributes valuable insights into the evolving landscape of machine learning techniques for structured data analysis.

REFERENCES

[1]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016.
[2]. X. Huang, A. Khetan, M. Cvitkovic and Z. Karnin, "TabTransformer: Tabular Data Modeling Using Contextual Embeddings," 2020.
[3]. "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data.
[4]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones and A. Gomez, "Attention is all you need," 2017.