AIDI 1002 Final Project Report

Project Title: Customer Churn Prediction with TabNet and CatBoost

Group Members: Kabir Barot, Krishi Patel

Date: August 2025

1. Introduction / Problem Statement

Loss of customers or paying members means customer churn and it may have serious drawbacks on company revenues. The ability to predict churn beforehand allows businesses to take initiatives regarding customer retention.

In this project, the work is on developing a churn prediction model on Telco Customer Churn data. The baseline model is TabNet, a neural architecture to learn tabular data that is interpretable. The major addition that we have made is the inclusion of CatBoost as a comparative baseline and SHAP to explain the model.

The main goals are:

1. Use machine learning to come up with a model that classifies customers into likely to churn and not.

2. Contrast the performance of TabNet with CatBoost.

3. Make the models interpretable enough to determine influential factors of churn

2. Dataset Information

Source: Kaggle – Telco Customer Churn
(https://www.kaggle.com/datasets/blastchar/telco-customer-churn)
We used a GitHub-hosted CSV mirror for direct access in Colab without Kaggle authentication.

Details:
- Rows: 7043 customers
- Features: 21 columns including demographics, account info, and service details
- Target Variable: Churn (Yes=1, No=0)

3. Proposed Methodology

The methodology consists of the following steps:
1. Data preprocessing including handling missing values, encoding categorical features, and splitting into train/validation/test sets.
2. Training a baseline model using TabNet for tabular data.
3. Adding a significant contribution by introducing CatBoost as a comparative model.

4. Using SHAP explainability to identify important features affecting churn.
5. Evaluating both models using Accuracy, F1 Score, and ROC-AUC metrics.

4. Implementation

The implementation was carried out entirely in Google Colab using Python.
Key libraries used:
- pytorch-tabnet for TabNet model implementation
- catboost for the gradient boosting model
- shap for explainability analysis
- scikit-learn for preprocessing and evaluation metrics

Steps:
1. Loaded the Telco Customer Churn dataset from a GitHub URL.
2. Dropped the 'customerID' column, converted 'TotalCharges' to numeric, and handled missing values.
3. Applied Label Encoding to categorical variables.
4. Split the dataset into 70% train, 15% validation, and 15% test sets with stratification.
5. Trained TabNet with default hyperparameters and evaluated on the test set.
6. Trained CatBoost with depth=6, learning_rate=0.1, and n_estimators=500, then evaluated.
7. Generated SHAP summary plots for CatBoost to visualize feature importance.

5. Contribution

The major addition to the baseline technique is the entry of the CatBoost as a benchmark model. CatBoost was selected as a good representation in case of categorical variables and less demand in terms of preprocessing. We have also incorporated SHAP explainability to enable us to get interpretable insights as a way of model prediction.

6. Results

The performance of both models is summarized below:

Model	Accuracy	F1 Score	ROC-AUC
TabNet	0.77	0.47	0.65
CatBoost	0.79	0.55	0.69

Observations:
- CatBoost slightly outperformed TabNet in Accuracy, F1 Score, and ROC-AUC.
- SHAP analysis highlighted key churn drivers such as Contract, Tenure, and MonthlyCharges.

7. Conclusion

In this project, We illustrated how deep learning and gradient boosting models can be employed in prediction of churn. Here, CatBoost, SHAP will be added to improve the interpretability and give viable business recommendations.

Future Improvements:
- Hyperparameter tuning for TabNet
- Threshold optimization to maximize F1 score
- Testing the approach on additional datasets like Adult Census

8. References

- Arik, S. Ö., & Pfister, T. (2019). TabNet: Attentive Interpretable Tabular Learning. https://arxiv.org/abs/1908.07442

- DreamQuark-AI. pytorch-tabnet GitHub: https://github.com/dreamquark-ai/tabnet

- Telco Customer Churn Dataset: https://www.kaggle.com/datasets/blastchar/telco-customer-churn