Churn Prediction with Advanced Feature Engineering and Ensemble Methods

Abstract

This study explores the practical application of advanced machine learning techniques to the problem of churn prediction in a tabular customer dataset. Emphasis is placed on extracting actionable insights through a combination of data preprocessing, engineered features, and ensemble modeling.

Methodology

The dataset underwent comprehensive exploratory data analysis to identify the most predictive attributes and underlying patterns related to customer churn. Data integrity was ensured by addressing missing values, outliers, and categorical inconsistencies. Feature engineering played a central role, with the introduction of interaction terms, ratio features, targeted binning of continuous variables, and mean encoding for selected categorical variables. This enriched the representational power of the data.

Preprocessing steps included one-hot encoding of categorical features and normalization of numerical values. Several machine learning models were implemented and benchmarked, including gradient boosting machines (LightGBM, CatBoost), regularized logistic regression, and a multilayer perceptron neural network. To capture complex relationships and reduce variance, a stacking ensemble was employed, leveraging the predictive strengths of individual models within a unified meta-learning framework. Model performance was evaluated using stratified k-fold cross-validation, with ROC AUC as the principal metric.

Results and Practical Impact

The final ensemble model achieved a ROC AUC of 75.07%, indicating strong discrimination capability for identifying customers at risk of churning. The integration of engineered features and ensemble learning not only improved overall predictive accuracy but also produced a solution robust to overfitting. Feature importance analysis revealed that newly constructed interaction and ratio features contributed significantly to model performance.

In a business context, such an approach enables early identification of customers with a high propensity to churn, supporting targeted retention strategies. The methodology presented is scalable and adaptable to similar problems in subscription-based industries, where reducing churn directly enhances revenue stability and customer lifetime value.

Conclusion

This work demonstrates the practical benefits of combining advanced feature engineering with ensemble machine learning techniques for churn prediction. The resulting model provides reliable, interpretable, and actionable outputs that are immediately valuable in a business analytics pipeline.