Project Report: Customer Churn Prediction

Introduction

Customer churn prediction is a real-world machine learning problem that aims to identify customers who are likely to leave a service. Since retaining existing customers is significantly more cost-effective than acquiring new ones, accurate churn prediction plays a critical role in business decision-making for telecom companies, SaaS platforms, and subscription-based services.

Problem Definition & Key Decision

The problem was treated as a binary classification task because the target variable (Churn) has two discrete classes: Yes and No. Regression techniques were not suitable as the output is categorical in nature.

Dataset Selection

The Telco Customer Churn dataset from Kaggle was selected due to its realistic structure, industry relevance, and wide acceptance in academic and professional environments.

Data Cleaning Decisions

The TotalCharges column contained hidden missing values represented as empty strings. These values were safely converted to NaN using numeric coercion to avoid silent data errors and ensure proper handling during preprocessing.

Preprocessing Strategy

Separate preprocessing pipelines were created for numerical and categorical features. Numerical features were imputed using the median and scaled using StandardScaler. Categorical features were imputed using the most frequent category and encoded using One-Hot Encoding. All preprocessing was handled using Pipeline and ColumnTransformer to prevent data leakage.

Train–Test Split Strategy

An 80–20 train-test split with stratification was used to preserve the original churn distribution across training and testing data, ensuring fair and reliable evaluation.

Model Selection

Logistic Regression was used as a baseline model due to its simplicity and interpretability. A Random Forest classifier was then applied to capture non-linear patterns and handle class imbalance more effectively.

Evaluation Metrics

Since the dataset is imbalanced, accuracy alone was not sufficient. Precision, Recall, and F1-score were used, with F1-score selected as the primary evaluation metric due to its balanced consideration of false positives and false negatives.

Cross-Validation Strategy

Five-fold cross-validation with a custom F1 scorer was used to obtain a robust and reliable estimate of model performance while avoiding dependency on a single train-test split.

Results

The Random Forest model achieved a test F1 score of 0.57 and an average cross-validation F1 score of 0.60, indicating stable generalization and improved performance compared to the baseline model.

Final Model Decision

Random Forest was selected as the final model due to its higher and more stable F1-score, better handling of non-linear relationships, and ability to manage class imbalance effectively.

Conclusion

This project demonstrates a complete end-to-end machine learning workflow following industry best practices, focusing on robust preprocessing, appropriate metric selection, model comparison, and clear decision-making. The project is well-suited for internship-level roles in data science and machine learning.

Author

Archit Bankey