Customer Churn Prediction: One-Page Report

Sean W. Ryan

Introduction: Customer churn is a critical issue for businesses, leading to the loss of customers who discontinue their services. Predicting churn allows companies to take proactive measures to retain customers, thus improving business performance. This project uses the "Telco Customer Churn" dataset from Kaggle to build models that predict whether a customer will churn.

Data Exploration and Cleaning: The dataset contains 7,043 entries with 21 columns, including customer demographics, account information, and service usage details. Initial exploration required converting TotalCharges to numeric and dropping the customerID column. One-hot encoding was performed on categorical features, and new features like tenure groups and average monthly charges were created to enhance model performance.

Feature Selection: Recursive Feature Elimination (RFE) with a RandomForestClassifier was used to select the top 10 significant features: tenure, MonthlyCharges, TotalCharges, avg_monthly_charges, gender_Male, InternetService_Fiber optic, Contract_One year, Contract_Two year, PaperlessBilling_Yes, and PaymentMethod_Electronic check.

Model Building: Starting with a baseline Logistic Regression model (accuracy: 79%), I explored Random Forest and Gradient Boosting models. Hyperparameter tuning for Gradient Boosting identified the best parameters: learning_rate=0.1, max_depth=3, and n_estimators=200.

Model Improvement: Addressing class imbalance with SMOTE improved recall for the churn class. The tuned Gradient Boosting model with SMOTE achieved a recall of 0.65 for the churn class, enhancing identification of churned customers.

Stacking Classifier: I implemented a Stacking Classifier combining Random Forest and Gradient Boosting models with a Logistic Regression meta-model. Hyperparameter tuning for the Stacking Classifier identified the best C parameter for the Logistic Regression meta-model.

Results: The Gradient Boosting model with SMOTE provided the best recall for the churn class, balancing accuracy and churn identification. Performance of models: Logistic Regression (Accuracy: 0.79, Precision: 0.62, Recall: 0.52, F1-Score: 0.56), Random Forest (Accuracy: 0.78, Precision: 0.63, Recall: 0.47, F1-Score: 0.54), Gradient Boosting (Accuracy: 0.79, Precision: 0.64, Recall: 0.48, F1-Score: 0.55), Tuned Gradient Boosting (Accuracy: 0.79, Precision: 0.63, Recall: 0.50, F1-Score: 0.56), Tuned GB with SMOTE (Accuracy: 0.77, Precision: 0.56, Recall: 0.65, F1-Score: 0.60), and Stacking Classifier (Accuracy: 0.79, Precision: 0.64, Recall: 0.47, F1-Score: 0.54).

SHAP Analysis: SHAP (SHapley Additive exPlanations) analysis revealed that tenure is the most critical factor, with longer tenure significantly reducing churn likelihood. Fiber optic internet service increases churn risk, while two-year contracts significantly reduce it. Payment methods like electronic checks are linked to higher churn, and monthly and total charges have complex interactions with churn risk.

Conclusion: The Gradient Boosting model with SMOTE provided the best recall for the churn class. The Stacking Classifier also showed promising results. Future work could explore further feature engineering, advanced ensemble methods, and more sophisticated hyperparameter tuning. This project demonstrates a comprehensive approach to solving a real-world problem using data science techniques.