SAP – Customer Churn Prediction

Author: Venkata Sandeep Kumar Reddy
Date: May 26, 2025

A machine learning project to predict customer churn using logistic regression and random forest models.

Abstract

Customer churn prediction is a critical task in the subscription-based software industry. This project models the churn behavior of customers using real-world telecommunications data, simulating a scenario similar to SAP's B2B SaaS customer base. We use Logistic Regression and Random Forest classifiers to predict churn with an emphasis on accuracy, recall, and ROC AUC.

Business Objective

The goal is to accurately predict which customers are likely to churn, allowing SAP to proactively engage at-risk clients and reduce customer attrition. The business benefit lies in reducing churn-related revenue loss.

Dataset

We used the Telco Customer Churn dataset, which contains 7043 customer records with features such as contract type, monthly charges, tenure, and payment method. The target variable is 'Churn' (Yes/No).

Data Cleaning

Missing values were handled and columns like customerID were dropped. TotalCharges was converted to numeric and invalid entries were removed. The final dataset was encoded for modeling.

Feature Engineering

All categorical variables were encoded using LabelEncoder. The 'Churn' column was binary encoded. Numerical features were scaled using StandardScaler.

Train-Test Split

The dataset was split into 80% training and 20% testing sets using stratified sampling to maintain the churn ratio.

Models Used

1. Logistic Regression: A linear baseline model.

2. Random Forest: An ensemble model with 100 decision trees.

Model Performance

Logistic Regression:

- Accuracy: 78.5%

- ROC AUC: 0.83

Random Forest:

- Accuracy: 79.0%

- ROC AUC: 0.81

Both models achieved strong predictive performance with ROC AUC > 0.80.

ROC Curve Comparison

The ROC curve clearly shows both models perform better than random chance. Random Forest had slightly better accuracy, while Logistic Regression had marginally higher ROC AUC.

Interpretation

The models correctly identified the majority of churners with reasonable precision and recall. Random Forest captured more complexity in the data, which can be beneficial in nonlinear cases.

Business Impact

By deploying this model, SAP could proactively identify at-risk customers and intervene with offers, loyalty programs, or better support, improving retention and reducing revenue churn.

Limitations

The model assumes past behavior predicts future churn. It also doesn't factor in external market factors or customer satisfaction surveys, which could enhance predictions.

Future Work

Integrate satisfaction scores, product usage logs, and support ticket sentiment. Deploy model to a real-time dashboard for sales and support teams.

Deployment Options

The model can be deployed via Streamlit for internal use. Sales or support staff could upload CSV files and view churn risk scores with explanations.

Conclusion

This project demonstrates the value of machine learning in customer retention strategies. With ~79% accuracy and strong interpretability, this solution could help SAP reduce churn and boost retention.

Appendix: Code Snippet

```
model = RandomForestClassifier(n_estimators=100)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
roc_auc_score(y_test, model.predict_proba(X_test)[:, 1])
```