Customer Churn Predictive Modeling

William Sparks

EKU - CSC 746

October 8, 2024

Introduction

- Predicting customer churn helps businesses retain valuable customers.
- We are using the Telco Churn dataset to train Random Forest model.
- Objective: Identify customers likely to churn based on features such as `MonthlyCharges`, `TotalCharges`, and `Contract`.

Data Preprocessing

- Cleaned dataset by removing the `customerID` column and handling missing values.
- Replaced categorical values like `No phone service` and `No internet service` with `No` for consistency.
- Replaced yes/no with 1/0.
- Applied one-hot encoding to categorical features such as `InternetService`, `Contract`, and `PaymentMethod`.

Cleaned Data Info

```
Index: 7032 entries, 0 to 7042
Data columns (total 27 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   gender                                   7032 non-null   int64
 1   SeniorCitizen                            7032 non-null   int64
 2   Partner                                  7032 non-null   int64
 3   Dependents                               7032 non-null   int64
 4   tenure                                   7032 non-null   int64
 5   PhoneService                             7032 non-null   int64
 6   MultipleLines                            7032 non-null   int64
 7   OnlineSecurity                           7032 non-null   int64
 8   OnlineBackup                             7032 non-null   int64
 9   DeviceProtection                         7032 non-null   int64
 10  TechSupport                              7032 non-null   int64
 11  StreamingTV                              7032 non-null   int64
 12  StreamingMovies                          7032 non-null   int64
 13  PaperlessBilling                         7032 non-null   int64
 14  MonthlyCharges                           7032 non-null   float64
 15  TotalCharges                             7032 non-null   float64
 16  Churn                                    7032 non-null   int64
 17  InternetService_DSL                      7032 non-null   bool
 18  InternetService_Fiber optic              7032 non-null   bool
 19  InternetService_No                       7032 non-null   bool
 20  Contract_Month-to-month                  7032 non-null   bool
 21  Contract_One year                        7032 non-null   bool
 22  Contract_Two year                        7032 non-null   bool
 23  PaymentMethod_Bank transfer (automatic)  7032 non-null   bool
 24  PaymentMethod_Credit card (automatic)    7032 non-null   bool
 25  PaymentMethod_Electronic check           7032 non-null   bool
 26  PaymentMethod_Mailed check               7032 non-null   bool
```

Random Forest Model

- Used Random Forest for churn prediction due to its robustness with large datasets and high accuracy.
- Split the data into 80% training and 20% testing sets.
- Applied Random Forest with 100 estimators and balanced class weights.

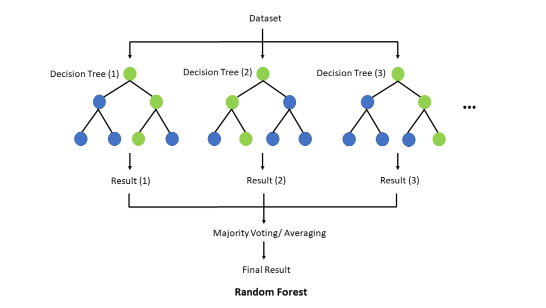

Figure: Random Forest Structure

Initial Results

- Achieved an accuracy of 79% on the test set.
- Precision, Recall, and F1-scores indicate the model is better at predicting non-churners (Class 0) than churners (Class 1).

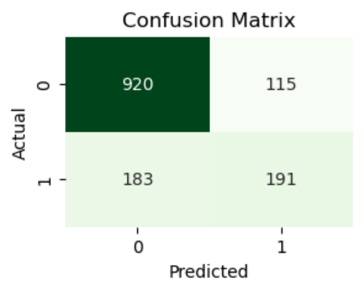

Figure: Confusion Matrix for Initial Model

Feature Importance

- Identified the most important features for predicting churn.
- Key features include: `TotalCharges`, `MonthlyCharges`, and `Tenure`.
- Less important features to the right in figure below were filtered out to improve model robustness. Acceptance threshold used: > 0.02.

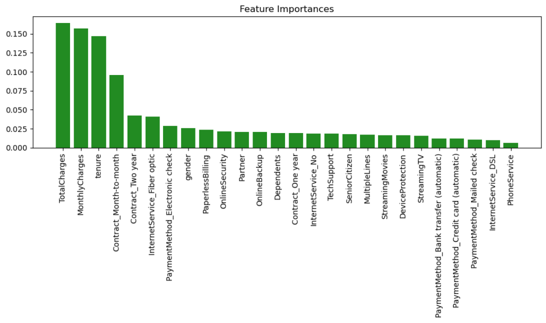

Figure: Top Features by Importance

Optimized Results

- After feature reduction, retrained the model with the most important features and maximized F1-score for class 1 (churn).
- Model performance: Accuracy dropped slightly to 78%, but F1-score for class 1 (churners) improved.
- Optimal threshold found to be 0.31.

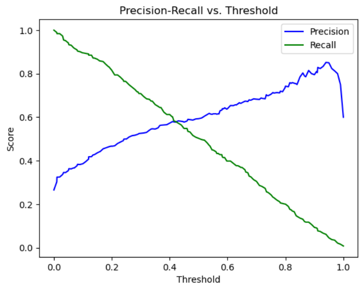

Figure: Precision-Recall Curve

Conclusion

- Random Forest is effective at predicting churn, especially for non-churners.
- Optimizing the threshold and feature set can improve performance for identifying churners if we accept a greater risk of false positives.
- Next steps: Explore further hyperparameter tuning and alternative models such as XGBoost, CatBoost, LightGBM.

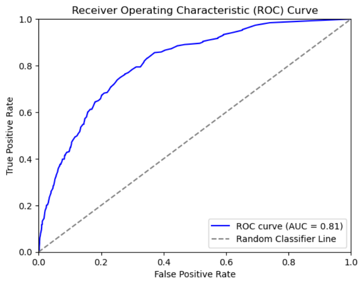

Figure: ROC Curve ($AUC = 0.81$)

Sources

- 1. https://www.flickr.com/photos/piro007/26178213572
- 2. https://commons.wikimedia.org/wiki/File:Random_forest_explain.png
- 3. https://www.kaggle.com/datasets/blastchar/telco-customer-churn