**Comprehensive Report on Customer Churn Prediction**

**Author**: Pankaj Patil  
**Date**: 08/09/2024

**1. Problem Definition and Planning**

The objective is to build a predictive model that identifies telecom customers at a high risk of churn. Using synthetically generated customer data, we aim to detect churn patterns and develop strategies for customer retention. The dataset consists of 5000 records with features categorized into customer demographics, account information, service details, and derived features. A target churn rate of approximately 20% is maintained throughout the analysis.

**2. Exploratory Data Analysis (EDA)**

**2.1 Dataset Overview**

The synthetic dataset includes the following features:

* **Customer Demographics**: CustomerID, Age, Gender.
* **Account Information**: ContractType, MonthlyCharges, TotalCharges, PaymentMethod.
* **Service Information**: TechSupport, InternetService, PaperlessBilling, Tenure.
* **Derived Features**: AverageMonthlyCharges, CustomerLifetimeValue.
* **Target Variable**: Churn (0 for no churn, 1 for churn).

**2.2 Data Cleaning and Preprocessing**

* **Handling Missing Values**: As this dataset was synthetically generated, it contained no missing values. If there had been, strategies such as imputation would have been applied.
* **Encoding Categorical Variables**: Binary attributes (like Gender, TechSupport, PaperlessBilling, Churn) were encoded using label encoding. Multiclass variables (ContractType, InternetService, PaymentMethod) were handled with one-hot encoding.
* **Feature Engineering**: Additional features were engineered to capture meaningful relationships:
  + Binary indicators for factors like auto-pay and long tenure.
  + Logarithmic and polynomial transformations to better model nonlinear relationships.
  + Interaction terms to capture combined effects between features.

**3. Bivariate and Multivariate Analysis**

**3.1 Distribution and Summary Statistics**

* **Age**: Appears uniformly distributed across most customers, with no bias toward a specific age group, aiding in model generalization.
* **TotalCharges**: Positively skewed, indicating that most customers have lower charges, while a few have higher charges (likely due to premium services or extended usage).

**3.2 Imbalanced Categorical Variables**

Categorical features like Churn, ContractType, PaymentMethod, and InternetService showed imbalanced distributions. The original churn rate was 19.5%, making it necessary to handle class imbalance. Using SMOTE (Synthetic Minority Over-sampling Technique), we balanced the dataset, raising the churn class to 50%. This helped mitigate the effects of imbalance and improved model robustness.

**3.3 Outlier Detection**

Boxplots of CustomerLifetimeValue and TotalCharges revealed upper-side outliers, particularly among customers who churned. These customers typically had high value or spending patterns, suggesting service quality or pricing model issues. Investigating these outliers provided actionable insights for improving customer retention.

**3.4 Correlation Analysis**

* **Strong Correlations (0.78)**: Between Tenure and both CustomerLifetimeValue and TotalCharges. Longer-tenured customers contribute more revenue and have higher lifetime value.
* **Moderate Correlations (0.55)**: Between MonthlyCharges, AverageMonthlyCharges, and CustomerLifetimeValue. Higher spending correlates with longer tenure, but not as strongly as with total charges.

**4. Model Development and Evaluation**

Given the categorical nature of the target variable (Churn), we employed classification algorithms. We experimented with both simple and complex models, comparing their performance based on accuracy, precision, recall, F1 score, and AUC (Area Under the ROC Curve).

**4.1 Baseline Models**

* **Logistic Regression**: Achieved an accuracy of 73.09% and AUC of 0.8007. Precision (73.45%) and recall (71.97%) were balanced, making it effective for minimizing false positives and negatives.
* **Decision Tree**: Slightly outperformed logistic regression, with an accuracy of 74.66% and precision of 74.86%. However, its AUC (0.7466) was lower, indicating slightly less discriminatory power.

**4.2 Advanced Models**

To improve accuracy, we explored more complex models:

* **Random Forest**: Achieved the highest accuracy (80.48%) and a strong AUC (0.8684). Its precision (83.18%) and recall (76.73%) provided a good balance between correctly identifying churn and minimizing errors.
* **Gradient Boosting**: Accuracy was slightly lower at 76.93%, with an AUC of 0.8341. High precision (82.99%) but lower recall (68.12%) suggested that it missed more true churn cases.
* **XGBoost**: Delivered robust performance with 80.06% accuracy and the highest AUC (0.8709). It balanced precision (82.33%) and recall (76.87%), making it effective at reducing misclassification.

**5. Model Selection and Insights**

After reviewing the models, **Random Forest** was selected due to its superior performance across evaluation metrics. It demonstrated strong generalization with higher accuracy and precision, ensuring better churn prediction.

After hyperparameter tuning with GridSearchCV, the Random Forest model showed significant improvements. The best parameters were max\_depth of 30, min\_samples\_leaf of 1, min\_samples\_split of 2, and n\_estimators set to 300. This resulted in an accuracy of 80.84%, precision of 83.41%, and recall of 77.29%, leading to an F1 score of 80.23%. The model's AUC reached 0.8727, indicating strong discriminative ability between churn and non-churn customers. These metrics reflect the model's balance between predictive power and generalization.

**Key Features**

* **Financial Features**: TotalCharges, MonthlyCharges, and AverageMonthlyCharges were highly significant predictors of churn.
* **CustomerLifetimeValue**: Highlighted the importance of customer value in assessing churn risk.
* **Age**: Provided insights into how demographic differences affect churn behavior.

**Conclusion**

This analysis identified key drivers of customer churn and demonstrated how machine learning models can predict at-risk customers with a high degree of accuracy. Financial factors, tenure, and customer value were crucial in predicting churn, and the Random Forest model proved the most effective at generalizing predictions. These insights can be used by the telecom company to implement targeted retention strategies, improving customer satisfaction and reducing churn rates.