Customer Churn Prediction Report

1. Introduction

This report presents the analysis and modeling process for predicting customer churn for a telecom company. We generated a synthetic dataset with 5,000 customer records, including features such as demographic information, contract details, and service usage. The goal is to build a predictive model that can effectively identify customers who are likely to churn.

2. Exploratory Data Analysis (EDA)

2.1 Summary Statistics

- **Age:** The customers' ages range from 18 to 69 years, with a median age of approximately 44 years.
- **Monthly Charges:** The monthly charges vary between $20 and $120, with an average of around $69.70. A small percentage (1%) of the customers have higher-than-average monthly charges due to introduced outliers.
- **Total Charges:** Total charges range from $100 to $5000, with a median of approximately $2,546. Some missing values were introduced, which were handled during data preprocessing.
- **Tenure:** The tenure spans from 1 to 71 months, with a median of around 36 months.

2.2 Distribution Visualizations

- **Age Distribution:** The age distribution is fairly uniform, with a slight concentration of customers in their mid-40s.
- **Gender Distribution:** The dataset is balanced in terms of gender, with almost equal representation of male and female customers.
- **Monthly Charges by Churn Status:** Customers who churn tend to have slightly higher monthly charges compared to those who don't.

Avni Burman

2.3 Correlation Matrix

- The correlation matrix shows a moderate positive correlation between Tenure and TotalCharges, as expected.
- MonthlyCharges and customer_lifetime_value are also positively correlated, indicating that higher monthly charges contribute to a higher customer lifetime value.
- No single feature has a very strong correlation with churn, suggesting that churn is influenced by a combination of factors.

3. Data Preprocessing

3.1 Missing Values

- The missing values in TotalCharges (5% of the data) were imputed using the median of the available data.

3.2 Categorical Features Encoding

- Categorical features such as Gender, ContractType, TechSupport, InternetService, PaperlessBilling, and PaymentMethod were encoded using one-hot encoding. This resulted in a total of 19 features after encoding.

3.3 Feature Scaling

- Numeric features were scaled using StandardScaler to standardize the range of the data for model training.

4. Model Development

4.1 Logistic Regression

- **Grid Search Parameters:** Regularization strength (C) was tuned over the values [0.01, 0.1, 1, 10].
- **Best Parameters:** The model performed best with C = 1.
- **Precision:** The cross-validated precision score was 81.3%.

4.2 Decision Tree Classifier

- **Grid Search Parameters:** The model was tuned with min_samples_split values [2, 10, 20] and min_samples_leaf values [1, 5, 10].
- **Best Parameters:** The best model used min_samples_split = 10 and min_samples_leaf = 5.
- **Precision:** The cross-validated precision score was 79.5%.

4.3 Model Evaluation

- The best-performing model was the Logistic Regression model, achieving a higher precision score than the Decision Tree model.

5. Model Performance

5.1 Classification Report

- **Precision:** 81.1% (proportion of true positive predictions out of all positive predictions)
- **Recall:** 63.4% (proportion of true positive predictions out of all actual positives)
- **F1-Score:** 70.9% (harmonic mean of precision and recall)

5.2 ROC AUC

- **ROC AUC:** 0.84, indicating a good ability of the model to distinguish between churn and non-churn customers.

5.3 Confusion Matrix

- The confusion matrix shows that the model correctly identified a significant number of both churn and non-churn customers, with some false positives and false negatives.

6. Feature Importance

- The top features contributing to churn prediction include ContractType_One year, ContractType_Two year, and MonthlyCharges, suggesting that customers on month-to-month contracts with higher monthly charges are more likely to churn.

7. Conclusion

- The Logistic Regression model was selected as the best model for predicting customer churn. The model performed well, with a good balance between precision and recall. The analysis suggests that contract type and monthly charges are key factors influencing customer churn. Further tuning and model improvements could focus on better handling class imbalance or exploring more complex models like ensemble methods.

Avni Burman