Churn Prediction Analysis Report

1. Data Overview

The dataset consists of 7043 customer records with 21 features. The dataset provides information about each customer, including demographic details, service usage, and subscription information. Here's a summary of the dataset:

- **CustomerID**: Unique identifier for each customer.
- **Gender:** Customer's gender (Male/Female).
- **SeniorCitizen:** Whether the customer is a senior citizen (0 or 1).
- **Partner:** Whether the customer has a partner (Yes/No).
- **Dependents:** Whether the customer has dependents (Yes/No).
- **Tenure:** The number of months the customer has been with the company.
- **PhoneService:** Whether the customer has phone service (Yes/No).
- **MultipleLines:** Whether the customer has multiple lines (No phone service, No, Yes).
- **InternetService:** Type of internet service the customer has (DSL, Fiber optic, No).
- **OnlineSecurity:** Whether the customer has online security (Yes/No).
- **OnlineBackup:** Whether the customer has online backup (Yes/No).
- **DeviceProtection:** Whether the customer has device protection (Yes/No).
- **TechSupport:** Whether the customer has tech support (Yes/No).
- **StreamingTV:** Whether the customer has streaming TV (Yes/No).
- **StreamingMovies:** Whether the customer has streaming movies (Yes/No).
- **Contract:** Type of contract the customer has (Month-to-month, One year, Two year).
- **PaperlessBilling:** Whether the customer uses paperless billing (Yes/No).

- **PaymentMethod:** Payment method used by the customer (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).

- **MonthlyCharges:** The amount charged to the customer monthly.

- **TotalCharges:** The total amount charged to the customer.

- **Churn:** Whether the customer has churned (Yes/No).

2. Data Preparation

Missing Values and Data Types

The dataset contains no missing values across all columns. The data types are:

- **Numeric:** `SeniorCitizen`, `tenure`, `MonthlyCharges`, `TotalCharges`.

- **Categorical:** All other columns.

Feature Engineering

1. Encoding Categorical Variables: Categorical features were converted into numerical format using one-hot encoding. This process results in additional columns for each category (e.g., `gender_Male`, `Partner_Yes`, `InternetService_DSL`).

2. Standardization: Numerical features were standardized to have a mean of 0 and a standard deviation of 1. This step helps in improving the performance of machine learning models by ensuring that features are on the same scale.

3. Data Splitting: The dataset was split into training (80%) and testing (20%) sets. The training set contains 5634 samples, while the testing set contains 1409 samples.

3. Model Training and Evaluation

Models Trained

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosting
5. XGBoost
6. Support Vector Machine (SVM)

Performance Metrics

- Logistic Regression

- Confusion Matrix:
  [[934 102]
   [149 224]]

- Precision: 0.86 for class 0, 0.69 for class 1
- Recall: 0.90 for class 0, 0.60 for class 1
- F1-Score: 0.88 for class 0, 0.64 for class 1
- Accuracy: 0.82

-Decision Tree

- Confusion Matrix:
  [[825 211]
   [201 172]]

- Precision: 0.80 for class 0, 0.45 for class 1
- Recall: 0.80 for class 0, 0.46 for class 1
- F1-Score: 0.80 for class 0, 0.46 for class 1
- Accuracy: 0.71

-Random Forest

- Confusion Matrix:
  [[945 91]
   [201 172]]

- Precision: 0.82 for class 0, 0.65 for class 1
- Recall: 0.91 for class 0, 0.46 for class 1
- F1-Score: 0.87 for class 0, 0.54 for class 1
- Accuracy: 0.79

-Gradient Boosting

- Confusion Matrix:
  [[937 99]
   [168 205]]

- Precision: 0.85 for class 0, 0.67 for class 1
- Recall: 0.90 for class 0, 0.55 for class 1
- F1-Score: 0.88 for class 0, 0.61 for class 1
- Accuracy: 0.81

-XGBoost

- Confusion Matrix:
  [[923 113]
   [179 194]]

- Precision: 0.84 for class 0, 0.63 for class 1
- Recall: 0.89 for class 0, 0.52 for class 1
- F1-Score: 0.86 for class 0, 0.57 for class 1
- Accuracy: 0.79

-SVM

- Confusion Matrix:
  [[954 82]
   [184 189]]

- Precision: 0.84 for class 0, 0.70 for class 1
- Recall: 0.92 for class 0, 0.51 for class 1
- F1-Score: 0.88 for class 0, 0.59 for class 1
- Accuracy: 0.81

Best Parameters and Scores

- Logistic Regression:
  - Best Parameters: `{'C': 100, 'solver': 'liblinear'}`
  - Best Score: 0.80

- Gradient Boosting:
  - Best Parameters: `{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 50}`
  - Best Score: 0.80

- SVM:
  - Best Parameters: `{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}`
  - Best Score: 0.80

4. Churn Prediction and Recommendations

Top Predicted Churners

The top customers predicted to churn with the highest probabilities are listed below:

CustomerID	Churn Probability	Churn Prediction	Recommended Action
7216-EWTRS	0.94	1	Offer a discount on the next bill
6910-HADCM	0.91	1	Offer a discount on the next bill
3068-OMWZA	0.89	1	Offer a discount on the next bill
5192-EBGOV	0.89	1	Offer a discount on the next bill
1455-UGQVH	0.88	1	Offer a discount on the next bill

Recommendations to Reduce Churn

Based on the analysis, the following strategies are recommended to reduce churn:

1. Tenure: Customers with shorter tenure are more likely to churn. Implement strategies to improve customer experience early in their subscription.

2. InternetService_Fiber optic: Customers with Fiber optic internet service show higher churn rates. Consider offering tailored promotions or discounts to retain these customers.

3. PaymentMethod_Electronic check: Customers using electronic checks have a higher churn rate. Explore offering alternative payment methods or incentives for using other payment methods.

4. Contract Type: Customers with month-to-month or one-year contracts have a higher churn rate compared to those with two-year contracts. Consider promoting longer-term contracts with additional benefits.

5. Customer Support: Enhancing customer support services could address issues leading to higher churn rates.

5. Conclusion

The churn prediction models provide valuable insights into customer behavior and help identify those at high risk of leaving. Implementing the recommended actions and strategies based on these insights can aid in reducing churn and improving customer retention.