TELECOM CUSTOMER CHURN PREDICTION PIPELINE

============================================================

Using Gradient Boosting Classifier with comprehensive ML pipeline

============================================================

============================================================

STEP 1: LOADING AND PREPROCESSING DATA

============================================================

Loading data from:(/content/WA\_Fn-UseC\_-Telco-Customer-Churn.csv)

Dataset shape: (7043, 21)

Columns: ['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn']

Dataset Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 7043 entries, 0 to 7042

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 customerID 7043 non-null object

1 gender 7043 non-null object

2 SeniorCitizen 7043 non-null int64

3 Partner 7043 non-null object

4 Dependents 7043 non-null object

5 tenure 7043 non-null int64

6 PhoneService 7043 non-null object

7 MultipleLines 7043 non-null object

8 InternetService 7043 non-null object

9 OnlineSecurity 7043 non-null object

10 OnlineBackup 7043 non-null object

11 DeviceProtection 7043 non-null object

12 TechSupport 7043 non-null object

13 StreamingTV 7043 non-null object

14 StreamingMovies 7043 non-null object

15 Contract 7043 non-null object

16 PaperlessBilling 7043 non-null object

17 PaymentMethod 7043 non-null object

18 MonthlyCharges 7043 non-null float64

19 TotalCharges 7043 non-null object

20 Churn 7043 non-null object

dtypes: float64(1), int64(2), object(18)

memory usage: 1.1+ MB

None

Handling missing values...

Missing values in TotalCharges: 0

Empty strings in TotalCharges: 0

Missing values after cleaning: 0

Data preprocessing completed successfully!

============================================================

STEP 2: EXPLORATORY DATA ANALYSIS

============================================================

Dataset Overview:

Total customers: 7043

Churned customers: 1869

Churn rate: 26.54%

EDA visualizations saved to 'eda\_analysis.png'

Key Insights from EDA:

Churn rate by Contract:

Month-to-month: 42.71%

One year: 11.27%

Two year: 2.83%

Churn rate by Internet Service:

Fiber optic: 41.89%

DSL: 18.96%

No: 7.40%

Average Monthly Charges:

No Churn: $61.27

Churn: $74.44

============================================================

STEP 3: FEATURE PREPARATION

============================================================

Categorical columns: ['gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod']

Numerical columns: ['SeniorCitizen', 'tenure', 'MonthlyCharges', 'TotalCharges']

Applying one-hot encoding to categorical variables...

Features before encoding: 19

Features after encoding: 30

Splitting data into training and testing sets...

Training set size: (5634, 30)

Testing set size: (1409, 30)

Training set churn rate: 26.54%

Testing set churn rate: 26.54%

============================================================

STEP 4: INITIAL MODEL TRAINING

============================================================

Training Gradient Boosting Classifier...

Initial model accuracy: 0.7991

Performing 5-fold cross-validation...

Cross-validation scores: [0.81987578 0.80479148 0.81810115 0.78793256 0.78241563]

Mean CV accuracy: 0.8026 (+/- 0.0305)

============================================================

STEP 5: HYPERPARAMETER OPTIMIZATION

============================================================

Parameter grid:

n\_estimators: [100, 200]

learning\_rate: [0.05, 0.1, 0.15]

max\_depth: [3, 4, 5]

min\_samples\_split: [10, 20]

min\_samples\_leaf: [5, 10]

Total combinations to test: 72

Performing GridSearchCV (this may take a while)...

Fitting 3 folds for each of 72 candidates, totalling 216 fits

Best parameters: {'learning\_rate': 0.1, 'max\_depth': 3, 'min\_samples\_leaf': 5, 'min\_samples\_split': 20, 'n\_estimators': 100}

Best cross-validation score: 0.8039

============================================================

STEP 6: MODEL EVALUATION

============================================================

Optimized model accuracy: 0.8013

Classification Report:

precision recall f1-score support

No Churn 0.84 0.90 0.87 1035

Churn 0.66 0.52 0.58 374

accuracy 0.80 1409

macro avg 0.75 0.71 0.73 1409

weighted avg 0.79 0.80 0.79 1409

Model evaluation visualizations saved to 'model\_evaluation.png'

Top 10 Most Important Features:

1. tenure: 0.2946

2. InternetService\_Fiber optic: 0.1939

3. PaymentMethod\_Electronic check: 0.1263

4. Contract\_Two year: 0.0750

5. TotalCharges: 0.0660

6. Contract\_One year: 0.0597

7. MonthlyCharges: 0.0529

8. PaperlessBilling\_Yes: 0.0225

9. OnlineSecurity\_Yes: 0.0167

10. TechSupport\_Yes: 0.0126

============================================================

BUSINESS INSIGHTS AND RECOMMENDATIONS

============================================================

Key Findings:

1. Customer tenure is a strong predictor. Implement retention programs for new customers.

2. Fiber optic customers may have different churn patterns. Analyze service quality.

3. Electronic check payment method may indicate higher churn risk.

4. Two-year contracts reduce churn significantly. Promote longer commitments.

5. Customers with higher total charges show different churn patterns. Focus on long-term value customers.

Recommended Actions:

1. Implement contract upgrade incentives

2. Develop early customer engagement programs

Model Performance Summary:

- The model can predict customer churn with 80.1% accuracy

- Focus on the top 5 features for maximum impact

- Implement targeted retention strategies based on risk factors

============================================================

PIPELINE COMPLETED SUCCESSFULLY!

============================================================



