Customer Churn Analysis

Project by Parth Chaudhari.

1. Introduction

Customer churn, the phenomenon of customers discontinuing their subscription or services, is a significant concern for businesses. Predicting and understanding churn is crucial to improving customer retention and ensuring sustainable growth. This study aims to analyze a customer churn dataset using machine learning models to predict churn and derive actionable insights. By identifying key patterns and contributing factors, businesses can formulate strategies to mitigate churn effectively.

The analysis employs a structured approach, beginning with a detailed data exploration and cleaning process, followed by model training and evaluation. The report highlights the challenges faced during data preprocessing and provides comprehensive insights from the models and visualizations.

2. Dataset Overview

The dataset contains customer-related data, subscription details, and their churn status. The key features include:

- **Customer Demographics:**

  a. gender: Gender of the customer.

  b. SeniorCitizen: Indicates whether the customer is a senior citizen.

  c. Partner and Dependents: Information about the customer's family status.

- **Subscription Details:**

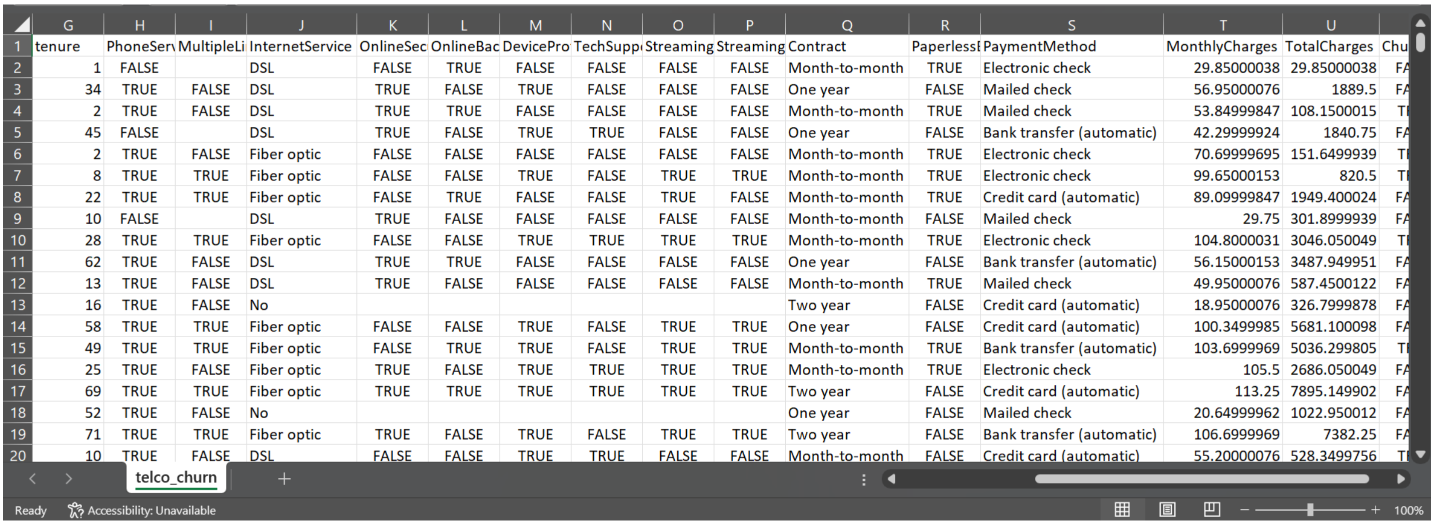
  a. Contract: Type of contract (e.g., month-to-month, one-year, two-year).

  b. MonthlyCharges and TotalCharges: Monetary details of the subscription.

  c. tenure: Number of months the customer has been with the company.

- **Services Used:**

  a. Features like InternetService, OnlineSecurity, TechSupport, and streaming services such as StreamingTV and StreamingMovies.



	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	tenure	PhoneServ	MultipleLi	InternetService	OnlineSec	OnlineBac	DevicePro	TechSupp	Streaming	Streaming	Contract	Paperless	PaymentMethod	MonthlyCharges	TotalCharges	Chu
2	1	FALSE		DSL	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	Month-to-month	TRUE	Electronic check	29.85000038	29.85000038	FA
3	34	TRUE	FALSE	DSL	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	One year	FALSE	Mailed check	56.95000076	1889.5	FA
4	2	TRUE	FALSE	DSL	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	Month-to-month	TRUE	Mailed check	53.84999847	108.1500015	TI
5	45	FALSE		DSL	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	One year	FALSE	Bank transfer (automatic)	42.29999924	1840.75	FA
6	2	TRUE	FALSE	Fiber optic	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Month-to-month	TRUE	Electronic check	70.69999695	151.6499939	TI
7	8	TRUE	TRUE	Fiber optic	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Month-to-month	TRUE	Electronic check	99.65000153	820.5	TI
8	22	TRUE	TRUE	Fiber optic	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	Month-to-month	TRUE	Credit card (automatic)	89.09999847	1949.400024	FA
9	10	FALSE		DSL	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	Month-to-month	FALSE	Mailed check	29.75	301.8999939	FA
10	28	TRUE	TRUE	Fiber optic	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	Month-to-month	TRUE	Electronic check	104.8000031	3046.050049	TI
11	62	TRUE	FALSE	DSL	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	One year	FALSE	Bank transfer (automatic)	56.15000153	3487.949951	FA
12	13	TRUE	FALSE	DSL	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	Month-to-month	TRUE	Mailed check	49.95000076	587.4500122	FA
13	16	TRUE	FALSE	No							Two year	FALSE	Credit card (automatic)	18.95000076	326.7999878	FA
14	58	TRUE	TRUE	Fiber optic	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	One year	FALSE	Credit card (automatic)	100.3499985	5681.100098	FA
15	49	TRUE	TRUE	Fiber optic	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	Month-to-month	TRUE	Bank transfer (automatic)	103.6999969	5036.299805	TI
16	25	TRUE	FALSE	Fiber optic	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	Month-to-month	TRUE	Electronic check	105.5	2686.050049	FA
17	69	TRUE	TRUE	Fiber optic	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Two year	FALSE	Credit card (automatic)	113.25	7895.149902	FA
18	52	TRUE	FALSE	No							One year	FALSE	Mailed check	20.64999962	1022.950012	FA
19	71	TRUE	TRUE	Fiber optic	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	Two year	FALSE	Bank transfer (automatic)	106.6999969	7382.25	FA
20	10	TRUE	FALSE	DSL	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	Month-to-month	FALSE	Credit card (automatic)	55.20000076	528.3499756	TI

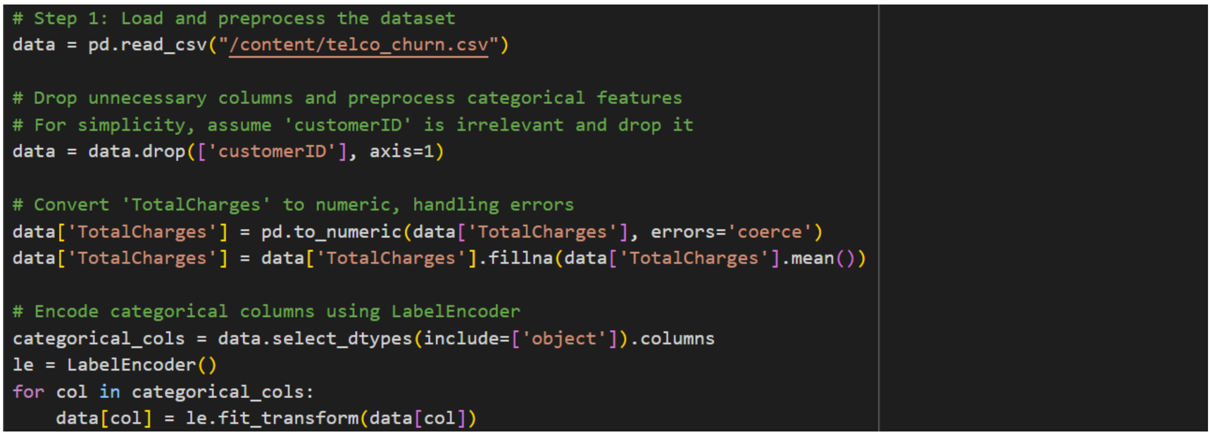
Observations in the Initial Dataset

1. **Missing Data:**
   - The TotalCharges column contained missing or non-numeric values for some entries. These entries were primarily customers with zero tenure.
2. **Class Imbalance:**
   - The churn status was significantly imbalanced, with a majority of the dataset belonging to major classes (0 and 1) and minimal representation for minor classes (e.g., Class 4).
3. **Inconsistent Data Types:**
   - Several features, such as TotalCharges, were stored as objects instead of numeric types.
4. **Redundant Information:**
   - The customerID column was deemed irrelevant to the analysis and removed.

3. Data Cleaning and Preprocessing

Detailed Data Cleaning Process

1. **Handling Missing Values:**
   - Missing entries in TotalCharges were replaced with the mean value of the column. This ensured consistency and avoided introducing bias due to deletion.
2. **Encoding Categorical Variables:**
   - Categorical features, such as gender, InternetService, and Contract, were transformed into numeric formats using Label Encoding to make them compatible with machine learning algorithms.
3. **Scaling Numerical Features:**
   - Continuous variables, including MonthlyCharges and TotalCharges, were normalized using StandardScaler to enhance model convergence and performance.
4. **Class Balancing Awareness:**
   - While the class imbalance was not fully resolved in this analysis, it was acknowledged as a critical challenge affecting model predictions, especially for underrepresented classes.



```
# Step 1: Load and preprocess the dataset
data = pd.read_csv("/content/telco_churn.csv")

# Drop unnecessary columns and preprocess categorical features
# For simplicity, assume 'customerID' is irrelevant and drop it
data = data.drop(['customerID'], axis=1)

# Convert 'TotalCharges' to numeric, handling errors
data['TotalCharges'] = pd.to_numeric(data['TotalCharges'], errors='coerce')
data['TotalCharges'] = data['TotalCharges'].fillna(data['TotalCharges'].mean())

# Encode categorical columns using LabelEncoder
categorical_cols = data.select_dtypes(include=['object']).columns
le = LabelEncoder()
for col in categorical_cols:
    data[col] = le.fit_transform(data[col])
```

Challenges Faced

1. **Class Imbalance:**
   - The dataset's imbalance posed challenges for algorithms, particularly in predicting minor classes accurately. Techniques like oversampling (e.g., SMOTE) or ensemble methods were considered for future iterations.
2. **Data Type Inconsistencies:**
   - Converting non-numeric entries in TotalCharges required careful handling to prevent data loss.
3. **Identifying Relevant Features:**
   - Determining which features significantly impacted churn required exploratory analysis, as some features exhibited weak correlations with the target variable.

4. Methodology

Workflow

1. **Exploratory Data Analysis (EDA):**
   - Visualizations were created to understand feature distributions, correlations, and relationships with churn.
2. **Model Training and Evaluation:**
   - Models were trained on a 70-30 train-test split.
   - Algorithms used include Logistic Regression, SVM, KNN, XGBoost, and Random Forest.
3. **Evaluation Metrics:**
   - Accuracy and ROC-AUC scores provided insights into model performance.
   - Precision, recall, and F1-scores were analyzed for a detailed assessment of multi-class predictions.

5. Results

Model Performance Summary

Model	Accuracy	ROC-AUC	Weighted Precision	Weighted Recall	Weighted F1-score
Logistic Regression	0.79	0.95	0.78	0.79	0.78
SVM	0.78	0.94	0.77	0.78	0.77
KNN	0.76	0.92	0.76	0.76	0.76
XGBoost	0.76	0.94	0.75	0.76	0.75

Model	Accuracy	ROC-AUC	Weighted Precision	Weighted Recall	Weighted F1-score
Random Forest	0.79	0.95	0.78	0.79	0.78

Observations

- Logistic Regression and Random Forest achieved the highest accuracy (0.79) and ROC-AUC (0.95), making them the most effective models for predicting churn.
- Class imbalance resulted in low recall and precision for minor classes (e.g., Class 4), reducing overall model effectiveness.

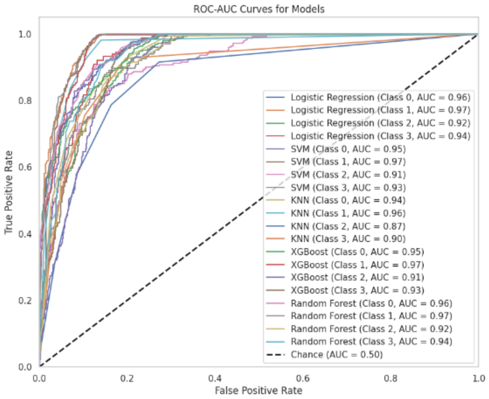
Confusion Matrices

- Most models had high recall for major classes (Class 0 and Class 1) but struggled with smaller classes, reflecting the dataset's imbalance.

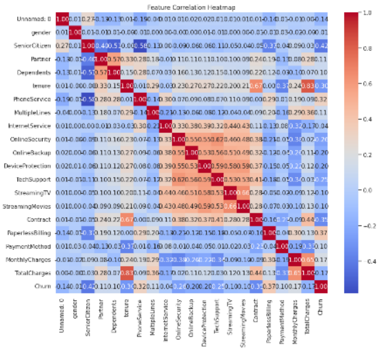
6. Data Insights

Key Visualizations and Insights

1. **ROC-AUC Curves:**
   - Models performed exceptionally well for major classes, with AUC values close to 0.97.
   - Minor classes exhibited lower AUC values, emphasizing the need for advanced balancing techniques.

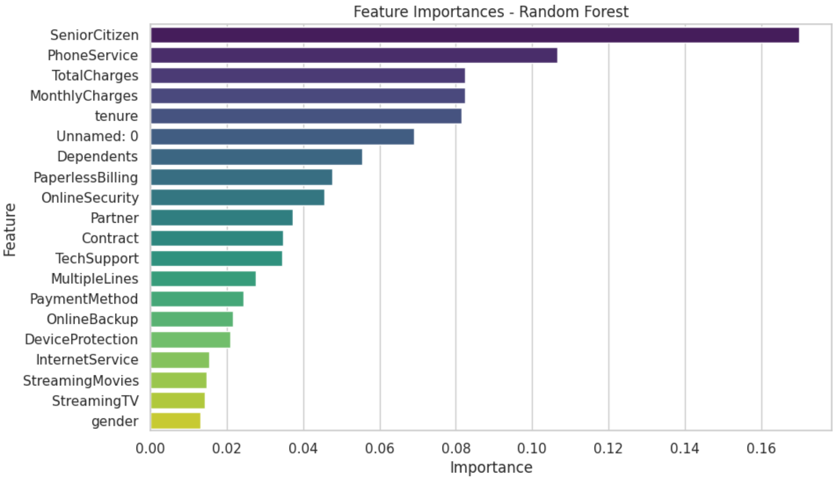


2. **Correlation Analysis:**
   - Strong correlations were observed between SeniorCitizen, PhoneService, MonthlyCharges, and TotalCharges with churn.
   - Features like Contract and InternetService showed moderate correlations, indicating their importance in prediction.



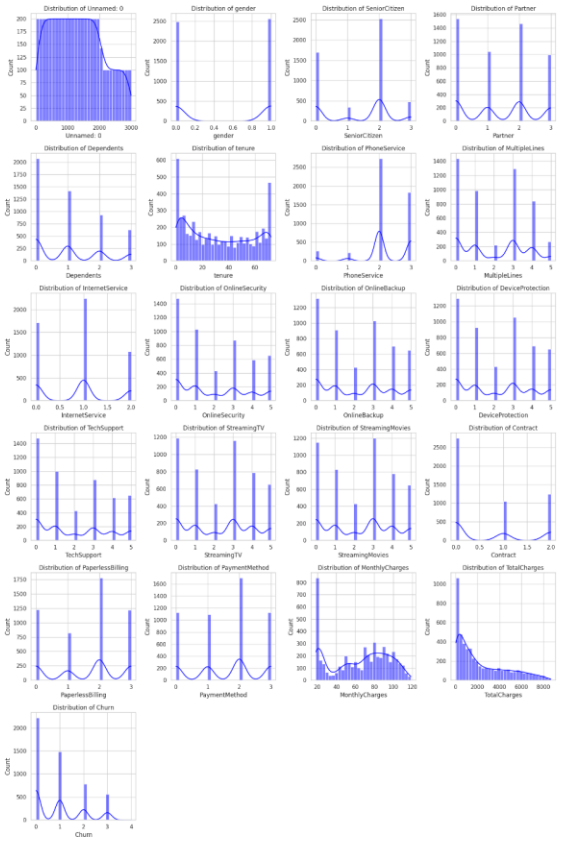
3. **Feature Importance:**

- The Random Forest model identified SeniorCitizen, PhoneService, and TotalCharges as the top contributors to churn.
- Features like StreamingTV and StreamingMovies had minimal impact on churn predictions.



4. **Feature Distributions:**

- Distributions of MonthlyCharges and TotalCharges revealed that customers with higher charges were more likely to churn.
- Tenure was inversely related to churn, with long-term customers being less likely to discontinue services.



7. Discussion

Strengths

- Random Forest and Logistic Regression provided robust predictions, balancing accuracy and interpretability.
- XGBoost effectively captured non-linear relationships, making it suitable for more complex datasets.

Weaknesses

- Class imbalance significantly impacted performance for minor classes, leading to poor recall and precision in those categories.
- The KNN model struggled with high-dimensional data, reducing its effectiveness.

Insights

- Customers with high MonthlyCharges and low tenure are at the highest risk of churn.
- Retention strategies targeting senior citizens and high-paying customers could substantially reduce churn rates.

8. Conclusion

This analysis demonstrated that Logistic Regression and Random Forest are the most effective models for churn prediction. Emphasizing tenure and payment behaviors, alongside addressing class imbalance, can significantly enhance predictive accuracy and business outcomes.

9. Recommendations

1. **Improving Models:**

- Implement oversampling techniques like SMOTE to address class imbalance.
- Explore hybrid models combining Random Forest and XGBoost to leverage their strengths.

2. **Business Strategies:**

- Introduce loyalty programs and personalized retention strategies for customers with high charges and low tenure.
- Enhance services related to Internet and TechSupport, as these features influence churn.

3. **Future Directions:**

- Collect additional data on customer satisfaction and complaints to enrich feature sets.
- Conduct time-series analysis to track behavioral trends over time, improving prediction accuracy further.