Customer Churn Prediction and Analysis

Track: **Data Science – Digital Egypt Pioneers Initiative (DEPI)**

Organization: **CLS ONL3_AIS4_G1 – Team 2**

Instructor: **Eng. Khaled Ellithy**

Project Type: **Graduation Project**

Team Members

1. Shahd Ashraf Ramadan Hassan
2. Aya Anwar Mahmoud Mohammad
3. Aya Hesham Saleh Nasr
4. Ayat Mohammad Mekky
5. Drothy Gerges Dawod Samoeel
6. Mohammad Naser Ibrahim Abd Elrahman
7. Ahmad Mohammad Ahmad Elshazly

1. Problem Definition

Customer churn represents a significant challenge for telecommunication companies.
High churn rates cause revenue loss and increase customer acquisition costs.

The goal of this project is to develop a **machine learning–based churn prediction system** that identifies customers likely to leave based on their demographics, subscription details, and usage patterns.

By predicting churn early, the company can take **preventive retention actions**, reduce revenue loss, and strengthen customer loyalty.

---

2. Objectives

- To analyze customer behavior and identify the main factors contributing to churn.
- To preprocess, engineer, and select the most relevant features for modeling.
- To train and compare six classification models:
  Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost.
- To evaluate models using performance metrics such as accuracy, precision, recall, F1-score, and AUC.
- To extract business insights and recommendations that can guide customer retention strategies.

---

3. Dataset Source

Dataset Name: Telco Customer Churn Dataset
Source: Kaggle – Telco Customer Churn

Description:
The dataset contains customer demographic information, account details, and service usage patterns from a telecommunications company.

Key Features Include:

- **CustomerID** – Unique customer identifier
- **Demographic Info** – Gender, SeniorCitizen, Partner, Dependents
- **Account Info** – Tenure, Contract, PaymentMethod
- **Service Usage** – InternetService, OnlineSecurity, TechSupport
- **Billing Info** – MonthlyCharges, TotalCharges
- **Target Variable** – Churn (Yes/No)

---

4. Data Understanding and Exploration

Exploratory data analysis (EDA) showed:

- Around **26.5%** of customers have churned. *"This indicates a moderately imbalanced dataset requiring SMOTE balancing before training."*
- **Month-to-month contracts** are strongly correlated with churn.
- Customers with **higher monthly charges** and **lower tenure** are more likely to leave.
- Lack of technical support and online security significantly increases churn likelihood.

Visualizations such as **correlation heatmaps, histograms, and churn distribution plots** were used to better understand these relationships.

---

5. Statistical Tests for Feature Relevance

To validate the strength of relationships between features and churn:

- **Chi-Square Test** for categorical variables:
  Features such as **Contract**, **PaymentMethod**, and **InternetService** had p-values < 0.05, confirming strong associations with churn.
- **Mann–Whitney U Test** for numerical variables:
  Features such as **Tenure** and **MonthlyCharges** showed significant differences between churned and non-churned customers.

Insights:

- Contract type and payment behavior strongly influence churn.
- Spending and tenure differences are statistically significant churn indicators.

---

6. Feature Engineering

New features were engineered to enhance model performance:

1. **Binary Encoding:** Converted all "Yes/No" columns to 1/0.
2. **Tenure Groups:** Created tenure-based segments (0–12m, 13–24m, 25–48m, 49–72m).
3. **Average Monthly Spending:**
   AvgChargesPerMonth = TotalCharges / tenure
4. **Service Count:** Number of active subscribed services.
5. **One-Hot Encoding:** Applied to Contract, PaymentMethod, and InternetService.
6. **Interaction Feature:**
   Tenure_x_Charges = tenure * MonthlyCharges

Insights:

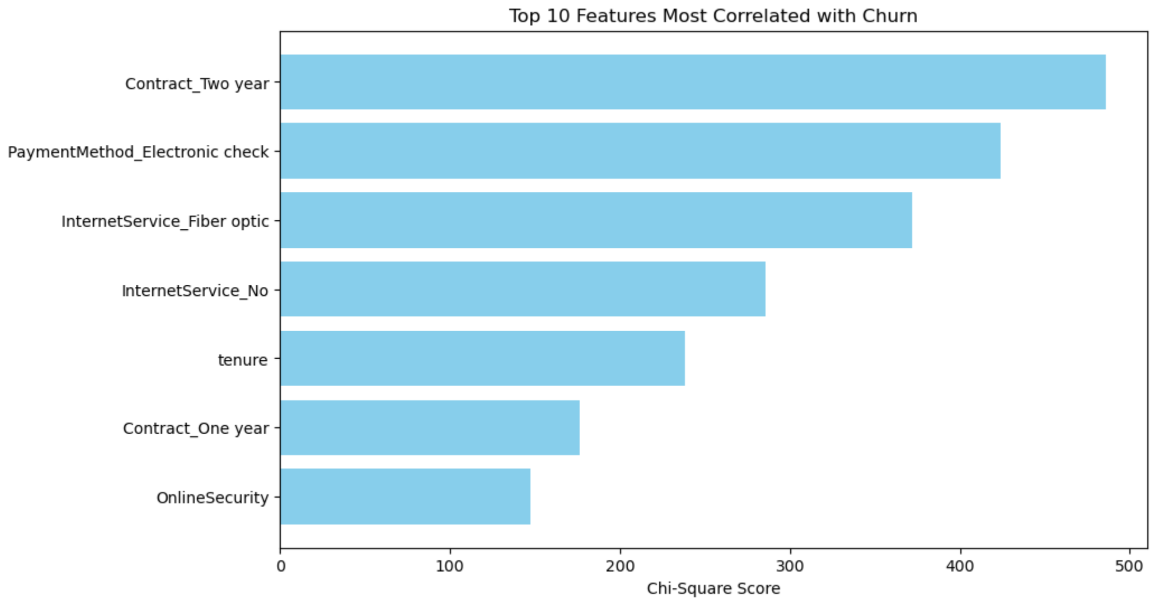
- ServiceCount and spending behavior reflect customer engagement.
- Interaction features reveal high-paying short-tenure customers, typically churn-prone.
- Grouped tenure segments improved interpretability and feature importance clarity.

7. Feature Selection

Using **Chi-Square (SelectKBest)**, the top 10 most influential features were identified:

Visualization:

A bar chart of Chi-Square scores ranked the most churn-correlated features.



8. Model Development and Evaluation

After completing feature selection, six classification models were trained, tuned, and evaluated on the processed dataset.
The evaluation metrics used include **Accuracy**, **AUC**, **Precision**, **Recall**, and **F1-score** for both **Class 0 (Non-Churn)** and **Class 1 (Churn)**.
The goal was to assess how effectively each model predicts customer churn while maintaining balanced performance across both classes.

8.1 Logistic Regression

Metric	Class 0	Class 1
Precision	0.90	0.49
Recall	0.71	0.77
F1-score	0.79	0.60
Accuracy	**0.7235**	

Observation:

The baseline model provides solid interpretability and a balanced recall for churners but struggles with precision for minority class (churn).

8.2 Random Forest

Metric	Class 0	Class 1
Precision	0.86	0.49
Recall	0.76	0.65
F1-score	0.80	0.56
Accuracy	**0.7285**	

Observation:
Improved balance compared to Logistic Regression. Better at detecting churners, but still moderate precision for Class 1.

8.3 Gradient Boosting

Metric	Class 0	Class 1
Precision	0.90	0.49
Recall	0.71	0.77
F1-score	0.79	0.60
Accuracy	**0.7235**	

Observation:
Similar to Logistic Regression, but with more stable training dynamics and reduced overfitting.

8.4 XGBoost

Metric	Class 0	Class 1
Precision	0.83	0.60
Recall	0.88	0.50
F1-score	0.85	0.55
Accuracy	**0.7776**	

Observation:
Strong predictive capability with balanced precision and recall.
Handles complex patterns effectively while maintaining interpretability.

8.5 LightGBM

Metric	Class 0	Class 1
Precision	0.84	0.60
Recall	0.87	0.55
F1-score	0.85	0.57
Accuracy	**0.7832**	

Observation:
Achieved the highest accuracy across all models.
Excellent overall generalization and efficient training on large data.

8.6 CatBoost

Metric	Class 0	Class 1
Precision	0.85	0.53
Recall	0.81	0.60
F1-score	0.83	0.56
Accuracy	**0.7500**	
AUC (Test)	**0.815**	
Cross-Validation Mean AUC	**0.931**	

Observation:
Excels in handling categorical features with minimal preprocessing.
High AUC scores indicate robust and consistent generalization across folds.

8.7 Model Comparison Summary

Model	Accuracy	AUC	Precision (Class 0)	Recall (Class0)	F1 (Class0)	Precision (Class 1)	Recall (Class1)	F1 (Class1)
Logistic Regression	0.7235	-----	0.90	0.71	0.79	0.49	0.77	0.60
Random Forest	0.7285	-----	0.86	0.76	0.80	0.49	0.65	0.56
Gradient Boosting	0.7235	0.825	0.90	0.71	0.79	0.49	0.77	0.60
XGBoost	0.7776	-----	0.83	0.88	0.85	0.60	0.50	0.55
LightGBM	0.7832	0.822	0.84	0.87	0.85	0.60	0.55	0.57
CatBoost	0.7500	0.931	0.85	0.81	0.83	0.53	0.60	0.56