# Customer Churn Analysis and Retention Strategy

**Program:** Digital Egypt Pioneers Initiative (DEPI) - Microsoft Data Engineering Track

**Project Supervisor:** Prof. Ahmed Azab

**Team Members (The "Churn Champions" Team):**

| Name | Role |
|---|---|
| Nada Ahmed Ahmed | Data Exploration & EDA |
| Abd Elrahman Ayman Abdo | Data Preprocessing & Cleaning |
| Yakoot Shaker Naseem | Team Leader; Model Building & Evaluation |
| Nancy Nabil Mohamed | Model Tuning & Interpretation |
| Ahmed Rashad Elsaid | Deployment & Retention Strategy |

**Date:** November 2025

---

## 1. Project Planning & Management

### 1.1. Project Overview

This project helps students analyze customer behavior and the factors contributing to customer attrition, and use the derived insights to inform a data-backed customer retention strategy. The core deliverable is an end-to-end machine learning solution designed to predict customer churn proactively. This initiative is critical as customer acquisition costs are significantly higher than retention costs, making the ability to accurately forecast and mitigate churn a key competitive advantage for any subscription-based or service-oriented business. The project demonstrates the application of advanced data science techniques—from exploratory data analysis and

rigorous preprocessing to sophisticated model development and practical deployment—to solve a high-value, real-world business problem.

## 1.2. Project Objectives

The project is guided by a dual set of objectives, balancing academic rigor with practical business application.

| Objective Type | Description |
|---|---|
| **Academic** | To successfully apply the full lifecycle of data science principles (Exploratory Data Analysis, preprocessing, modeling, evaluation, and deployment) to a real-world problem, thereby fulfilling the graduation requirements for the Digital Egypt Pioneers Initiative (DEPI) program and demonstrating technical competence in the Microsoft Data Engineering Track. |
| **Business** | To create a robust, predictive tool that empowers companies to proactively identify customers at high risk of churning. Furthermore, to develop data-backed, actionable strategies that can be immediately implemented to reduce attrition, save costs, and maximize Customer Lifetime Value (CLV). |

## 1.3. Project Team & Responsibilities

The project’ s success was a result of a structured, collaborative effort, with each member of the “Churn Champions” team taking ownership of a critical phase of the data science lifecycle.

| Team Member | Core Responsibility | Contribution Detail |
| --- | --- | --- |
| **Yakoot Shaker Naseem** (Team Leader) | Model Building & Evaluation | Led the project, established the technical stack, oversaw the data science workflow, and was directly responsible for the selection, training, and initial evaluation of all predictive models. |
| **Nada Ahmed Ahmed** | Data Exploration & EDA | Conducted the initial data quality assessment, performed comprehensive Exploratory Data Analysis (EDA), generated key visualizations, and identified initial data patterns and class imbalance issues. |
| **Abd Elrahman Ayman Abdo** | Data Preprocessing & Cleaning | Managed all data preparation tasks, including handling missing values, encoding categorical features, feature engineering, and applying necessary feature scaling (StandardScaler) to prepare the data for modeling. |
| **Nancy Nabil Mohamed** | Model Tuning & Interpretation | Focused on optimizing the selected final model through hyperparameter tuning (e.g., using `GridSearchCV`), performing rigorous model validation, and conducting feature importance analysis to derive business-critical insights. |
| **Ahmed Rashad Elsaid** | Deployment & Retention Strategy | Responsible for the practical application, including developing the Streamlit-based UI/UX, deploying the final model into the application, and formulating the three-pronged, data-backed customer retention strategy. |

## 1.4. Project Timeline & Key Milestones

The project followed a disciplined timeline, structured into three main phases to ensure timely completion and adherence to the DEPI program’ s requirements.

| Phase | Duration | Key Activities |
| --- | --- | --- |
| **Project Inception** | Aug 21 - Sep 13 | Graduation project announcement, team formation, project ideation, and final selection of “Customer Churn Analysis.” |
| **Phase 1: Planning & Data Acquisition** | Sep 13 - Sep 26 | Intensive dataset research and selection (Kaggle), and development of the detailed project schedule and task allocation among team members. |
| **Phase 2: Project Execution (Milestones 1-3)** | Sep 26 - Nov 6 | The core development period, encompassing all technical work from data analysis to final deployment and strategy formulation. |

**Detailed Milestone Breakdown:**

| Milestone | Date Range | Responsible Team Member | Deliverable |
| --- | --- | --- | --- |
| **Milestone 1** (Data Analysis) | Sep 26 - Oct 12 | Nada Ahmed & Abd Elrahman Ayman | Comprehensive EDA report and a clean, preprocessed dataset ready for modeling. |
| **Milestone 2** (Model Development) | Oct 12 - Oct 29 | Yakoot Shaker & Nancy Nabil | Optimized predictive model (XGBoost) with a full evaluation report and feature importance analysis. |
| **Milestone 3** (Deployment & Strategy) | Oct 29 - Nov 6 | Ahmed Rashad | Functional Streamlit application and a finalized, data-backed retention strategy document. |

---

# 2. Stakeholder Analysis

A thorough understanding of the project’ s stakeholders was essential to tailor the documentation and the final product to meet diverse requirements, ranging from academic assessment to practical business utility.

## 2.1. Academic & Program Stakeholders

These stakeholders are primarily concerned with the project's methodological soundness, technical execution, and adherence to program standards.

| Stakeholder | Primary Interest | Requirement from Documentation |
|---|---|---|
| **DEPI / MCIT** | Tracking student progress, ensuring high-quality skill development, and archiving successful project outcomes as a testament to the program's effectiveness. | A formal, well-structured document that clearly outlines the project's scope, methodology, and results, suitable for public archiving and program review. |
| **Prof. Ahmed Azab** (Project Supervisor) | Assessing the team's technical depth, problem-solving abilities, the correctness of the methodology, and the originality of the insights derived. | Detailed sections on data preprocessing, model selection justification, and rigorous evaluation metrics (Section 3 & 4). |

## 2.2. Hypothetical Business Stakeholders (Target Client)

These stakeholders are focused on the project's real-world value, return on investment (ROI), and ease of integration into existing business processes.

| Stakeholder | Primary Interest | Requirement from Final Product |
|---|---|---|
| **Business/Marketing Managers** | Reducing customer churn to protect revenue streams, minimizing customer acquisition costs, and gaining a competitive edge through proactive retention. | A simple, high-level tool that provides clear, actionable "Churn / No Churn" indicators and a strategic report on how to leverage the model's insights (Section 5). |
| **Customer Service Agents** | Easily identifying high-risk customers during routine interactions before they leave, and being guided on appropriate retention offers or intervention scripts. | An on-demand, user-friendly interface to check a customer's status and receive guidance on retention offers (e.g., discounts, plan upgrades) based on the predicted risk level. |

---

# 3. Data Analysis and Preparation (Milestone 1)

The foundation of any successful data science project is the quality and preparation of its data. This phase was critical for establishing a robust dataset for modeling.

## 3.1. Data Source

The project utilized a publicly available dataset to ensure reproducibility and focus on the application of techniques rather than proprietary data access.

| Attribute | Detail |
|---|---|
| **Dataset** | “Customer Churn Dataset” |
| **Source** | Kaggle |
| **Link** | https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset |

**Description:** Customer churn refers to the phenomenon where customers discontinue their relationship with a company or service. This dataset provides a historical record of customer attributes and behavior, allowing for the development of a predictive model to forecast which customers are likely to churn in the near future. The features encompass demographic, service usage, and complaint history, which are all critical inputs for a comprehensive churn analysis.

## 3.2. Data Schema and Feature Analysis

The dataset was provided as a single flat file (CSV), simplifying the initial data ingestion process as no complex relational joins were required. The schema is composed of a mix of identifiers, numerical, and categorical features.

| Feature | Data Type | Description |
|---|---|---|
| `customer_id` | String | Unique identifier for the customer (non-predictive). |
| `tenure` | Integer | Number of months the customer has been with the company. |
| `usage` | Float | Data usage, call minutes, or other service metrics. |
| `complaints` | Integer | Number of complaints filed by the customer. |
| `plan_type` | Category | The customer’ s subscription plan (e.g., Basic, Premium). |
| `monthly_charges` | Float | The amount charged to the customer monthly. |
| `support_calls` | Integer | Number of times the customer contacted technical support. |
| `churn_status` | Binary | **1 (Churned) or 0 (Not Churned)** - This is the target variable. |

### 3.3. Exploratory Data Analysis (EDA) Summary

Nada Ahmed’ s comprehensive EDA revealed several critical characteristics of the data that directly informed the subsequent preprocessing and modeling phases.

1. **Class Imbalance:** EDA revealed a significant class imbalance, with the majority of customers belonging to the ‘Not Churned’ (0) class. This necessitated the selection of appropriate evaluation metrics (Recall, F1-Score, ROC-AUC) over simple Accuracy, as well as considering techniques like Synthetic Minority Over-sampling Technique (SMOTE) or class weighting during modeling.
2. **Feature vs. Churn Correlation:** Visualizations, such as bar charts comparing categorical features against the churn rate, showed strong correlations. For instance, customers on the ‘Basic’ plan type exhibited a higher churn rate than those on ‘Premium’ plans.
3. **Tenure Distribution:** Histograms of the `tenure` feature showed a bimodal distribution, with a high concentration of churn among customers with very low tenure (less than 6 months) and a smaller spike among long-term customers, suggesting different churn drivers for different customer segments.

4. **Complaint Impact:** A direct visualization of `complaints` versus churn rate demonstrated a near-monotonic relationship: customers with a higher number of complaints were significantly more likely to churn.

**Key Insight:** EDA revealed a significant class imbalance and strong correlations between features like `complaints`, `tenure`, and `plan_type` with the target variable, confirming the dataset' s suitability for predictive modeling.

## 3.4. Data Preprocessing Summary

Abd Elrahman Ayman executed a rigorous preprocessing pipeline to transform the raw data into a format suitable for machine learning algorithms, ensuring model stability and performance.

| Preprocessing Step | Detail | Rationale |
|---|---|---|
| **Handling Missing Data** | Used mean imputation for the `usage` feature, as the missingness was determined to be Missing At Random (MAR) and the mean provided a robust estimate without introducing significant bias. | Ensures all records are usable and prevents model failure due to null values. |
| **Encoding Categorical Features** | Applied One-Hot Encoding to the nominal categorical feature `plan_type`. | Converts non-numeric data into a binary format that machine learning models can process, avoiding the assumption of ordinality. |
| **Feature Scaling** | Utilized `StandardScaler` on all numeric features, including `tenure`, `usage`, `monthly_charges`, and `support_calls`. | Standardizes the features by removing the mean and scaling to unit variance, which is crucial for distance-based algorithms and to prevent features with larger magnitudes from dominating the model training process. |

---

# 4. Predictive Model Development (Milestone 2)

This phase focused on selecting, training, and optimizing the machine learning model to achieve the highest predictive performance, with a strong emphasis on interpretability for business application.

## 4.1. Model Selection

The team adopted a comparative approach, selecting a diverse set of models to establish a performance baseline and identify the most suitable advanced algorithm.

**Model Choices:**

1. **Baseline Model: Logistic Regression** (for its simplicity and interpretability).
2. **Advanced Ensemble Models: Random Forest** and **XGBoost** (Selected for their high performance, ability to handle non-linear relationships, and built-in feature importance capabilities).

**Justification:** The final selection focused on **XGBoost (Extreme Gradient Boosting)**. This model consistently outperformed the others due to its superior handling of the class imbalance and its ability to capture complex interactions between features, making it the most robust predictor of customer churn.

## 4.2. Training and Evaluation

The model training process was executed with a focus on rigorous validation to ensure the model’ s generalization capability.

**Data Split:** The preprocessed dataset was split into an 80% training set and a 20% testing set to evaluate the model on unseen data.

**Evaluation Metrics:** Given the significant class imbalance identified in the EDA, the team prioritized metrics that provide a more accurate picture of performance than simple accuracy:

- **Precision:** The proportion of positive predictions that were actually correct (minimizing false positives).
- **Recall (Sensitivity):** The proportion of actual positive cases that were correctly identified (minimizing false negatives, which is critical for identifying all potential

churners).

- **F1-Score:** The harmonic mean of Precision and Recall, providing a balanced measure.
- **ROC-AUC (Area Under the Receiver Operating Characteristic Curve):** Measures the model's ability to distinguish between the two classes.

**Results:** The final, optimized XGBoost model demonstrated superior performance on the test set.

| Metric | Logistic Regression | Random Forest | XGBoost (Final Model) |
|---|---|---|---|
| Accuracy | 0.82 | 0.87 | **0.89** |
| Precision (Churn) | 0.65 | 0.78 | **0.81** |
| Recall (Churn) | 0.70 | 0.75 | **0.80** |
| F1-Score (Churn) | 0.67 | 0.76 | **0.80** |
| ROC-AUC | 0.84 | 0.88 | **0.91** |

## 4.3. Tuning and Interpretation

Nancy Nabil's work focused on extracting maximum performance and business value from the selected XGBoost model.

**Hyperparameter Tuning:** The final XGBoost model was rigorously optimized using `GridSearchCV` with a 5-fold cross-validation strategy. Key hyperparameters tuned included `max_depth`, `learning_rate`, and `n_estimators`. This process resulted in a model that was both highly accurate and less prone to overfitting.

**Feature Importance:** A critical step for business application was the analysis of feature importance, which explains *why* the model makes its predictions. This was achieved using the model's built-in importance scores, which revealed the primary drivers of customer churn.

**Analysis revealed the top 3 drivers of churn are:**

1. **High `complaints`:** This was the single most influential factor, indicating that unresolved or frequent issues are the strongest predictor of customer dissatisfaction and subsequent attrition.

2. **Low `tenure`:** Customers who have been with the company for a short period are inherently more volatile and likely to churn, confirming the "honeymoon period" risk.
3. **`plan_type` (specifically the 'Basic' plan):** The type of subscription plan significantly impacts churn, suggesting that the 'Basic' plan may not meet customer needs or that 'Premium' customers are more satisfied due to better service or features.

---

# 5. Deployment and Business Application (Milestone 3)

The final phase bridged the gap between the technical model and its practical application, ensuring the solution delivers tangible business value.

## 5.1. UI/UX Design (Streamlit Application)

The deployment strategy centered on creating a simple, accessible interface that caters directly to the non-technical business user.

**User Persona:** The primary user is a non-technical business manager or customer service agent. They require a tool that is intuitive, fast, and provides clear, unambiguous results.

**User Story:** "As a user, I want to enter a customer's data (e.g., tenure, plan, complaints) into a simple form and click 'Predict' to immediately see if they are a 'High Churn Risk' or 'Low Churn Risk,' so that I can take proactive retention action."

**Interface Design:** The UI is built with **Streamlit** for its simplicity, rapid prototyping capabilities, and ease of deployment. The design features a clean input form on a sidebar where the user enters the customer's attributes. The main panel provides a clear, color-coded prediction output (e.g., **RED** for High Risk, **GREEN** for Low Risk) and a brief explanation of the result, fulfilling the requirement for a tool that is actionable without needing technical expertise.

## 5.2. Model Deployment

The deployment process ensured that the production environment accurately replicates the training environment, preventing data leakage or transformation errors.

**Deployment Mechanism:** The trained XGBoost model was saved (pickled) and loaded directly into the Streamlit application's environment. The application functions as follows:

1. **Input Collection:** The Streamlit front-end collects the user input (e.g., tenure, plan, complaints) via the form.
2. **Real-Time Preprocessing:** The application includes the exact same preprocessing pipeline (imputation, One-Hot Encoding, `StandardScaler`) used during training. This ensures the user's raw input is transformed correctly.
3. **Prediction:** The preprocessed data is fed to the loaded XGBoost model to generate a probability score.
4. **Output:** The probability is converted into a clear "High Churn Risk" or "Low Churn Risk" label and displayed to the user.

This architecture ensures a seamless, end-to-end prediction service that is both accurate and easy to maintain.

## 5.3. Data-Backed Retention Strategy

The project's ultimate value lies in the strategic actions it enables. Based on the model's predictions and feature importance analysis, the following three-pronged retention strategy is proposed:

**Strategy 1: Proactive Outreach (Targeting Low Tenure/High Usage)**

- **Action:** Use the application to identify customers who are in the low `tenure` category (e.g., < 6 months) but have high `usage` metrics. These customers are valuable but volatile.
- **Intervention:** Offer them a surprise loyalty discount or a free service upgrade after their third month. This preemptive action addresses the low-tenure risk factor before a complaint arises, turning a potential churner into a loyal customer.

**Strategy 2: Complaint Intervention (Targeting High-Risk/High Complaints)**

- **Action:** Flag any customer in the "High-Risk" category who also has >0 `complaints` for immediate follow-up from a senior retention specialist.
- **Intervention:** The specialist's goal is not just to solve the current issue but to perform a "service recovery" that rebuilds trust. This directly addresses the highest-ranking churn driver (`complaints`) with a high-touch, personalized approach.

**Strategy 3: Feedback Loop (Targeting Plan Optimization)**

- **Action:** Use the model's feature importance (e.g., the high churn rate associated with the 'Basic' `plan_type`) to inform business decisions.
- **Intervention:** The product development team should investigate the 'Basic' plan. Is it under-featured? Is the price point too high for the value? The model provides the evidence needed to justify improving or retiring unpopular subscription plans, thereby reducing systemic churn.

---

# 6. Conclusion & Future Work

## 6.1. Project Summary

This project successfully developed an end-to-end churn prediction solution, from the initial stages of data cleaning and exploratory analysis to the creation of a highly accurate predictive model and its deployment in a functional Streamlit application. By adhering to a rigorous data science methodology and focusing on both academic requirements and practical business utility, the team has delivered a valuable asset. The solution not only meets all requirements for the DEPI graduation project but also provides a clear, actionable framework for companies to significantly reduce customer attrition, demonstrating the team's competence in translating complex data engineering and machine learning concepts into tangible strategic outcomes.

## 6.2. Future Work

To ensure the continued relevance and performance of the churn prediction system, the following improvements are recommended for future iterations:

1. **Integration with Live Database:** The current application uses static input. Future work should focus on integrating the Streamlit application with a live company database (e.g., PostgreSQL or SQL Server) to enable real-time, automated risk scoring for the entire customer base, moving from a reactive tool to a proactive monitoring system.
2. **Model Retraining Automation:** The model’ s performance will naturally decay as customer behavior evolves. A quarterly model retraining pipeline should be established, utilizing new data to ensure the model remains highly predictive and captures emerging churn drivers.
3. **Customer Dashboard Implementation:** Enhance the Streamlit application by adding a customer dashboard feature. This would allow business managers to visualize churn trends over time, track the success of retention campaigns, and view the feature importance scores dynamically, providing deeper business intelligence.

---

**References**


1. Kaggle. *Customer Churn Dataset*. https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset
2. DEPI. *Digital Egypt Pioneers Initiative*. https://www.mcit.gov.eg/en/initiatives/digital-egypt-pioneers-initiative-depi
3. XGBoost. *Extreme Gradient Boosting*. https://xgboost.readthedocs.io/en/stable/
4. Streamlit. *The fastest way to build and share data apps*. https://streamlit.io/
5. MCIT. *Ministry of Communications and Information Technology*. https://www.mcit.gov.eg/