

Project Title: Advanced Customer Churn Prediction and Retention Strategy

Submitted by Sree Sankaran Chackoth

Project Phase 1: Planning & Data Collection

1. Define Project Goals and Scope

Objective: The primary goal of this project is to accurately predict customer churn and enhance customer retention through data-driven strategies.

Deliverables:

- **Project Statement:**
 - To predict customer churn effectively and implement personalized retention strategies using insights from historical data.
- **High-level Objectives:**
 - Develop a predictive churn model.
 - Identify key factors influencing churn.
 - Suggest personalized retention strategies.

2. Define Data Requirements / Requirement Gathering

Objective: Clearly define and gather the necessary datasets for analyzing churn, including:

- Demographic data
- Socio-economic indicators
- Service usage statistics
- Customer interaction and satisfaction data

Deliverables:

- **Data Specifications:**
 - Demographics (Age, Gender, SeniorCitizen)
 - Socio-economic status (CLTV, Monthly Charges)
 - Service details (Contract Type, Payment Method, Internet Service)

- Customer churn reasons and categories

3. Data Collection and Data Sources

Objective: Gather reliable data sources to support thorough churn analysis.

Deliverables:

- **Data Sources:**
 - Churn Prediction Dataset (Kaggle)
 - Telco Customer Churn Dataset (Kaggle)
- **Data Collection Method:**
 - Direct CSV download
- **Data Cleaning Steps:**
 - Identify and handle missing values
 - Detect and manage outliers

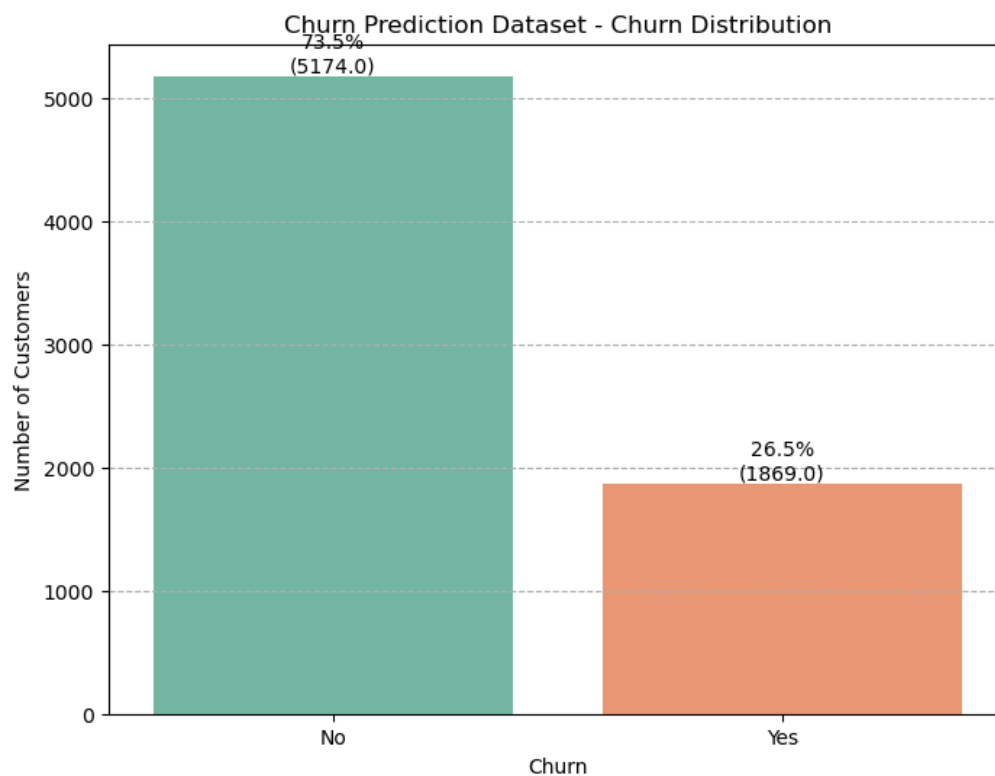
4. Data Exploration and Preliminary Analysis

Objective: Conduct an initial exploration to identify trends, anomalies, and potential data quality issues.

Deliverables:

Key Trends and Insights:

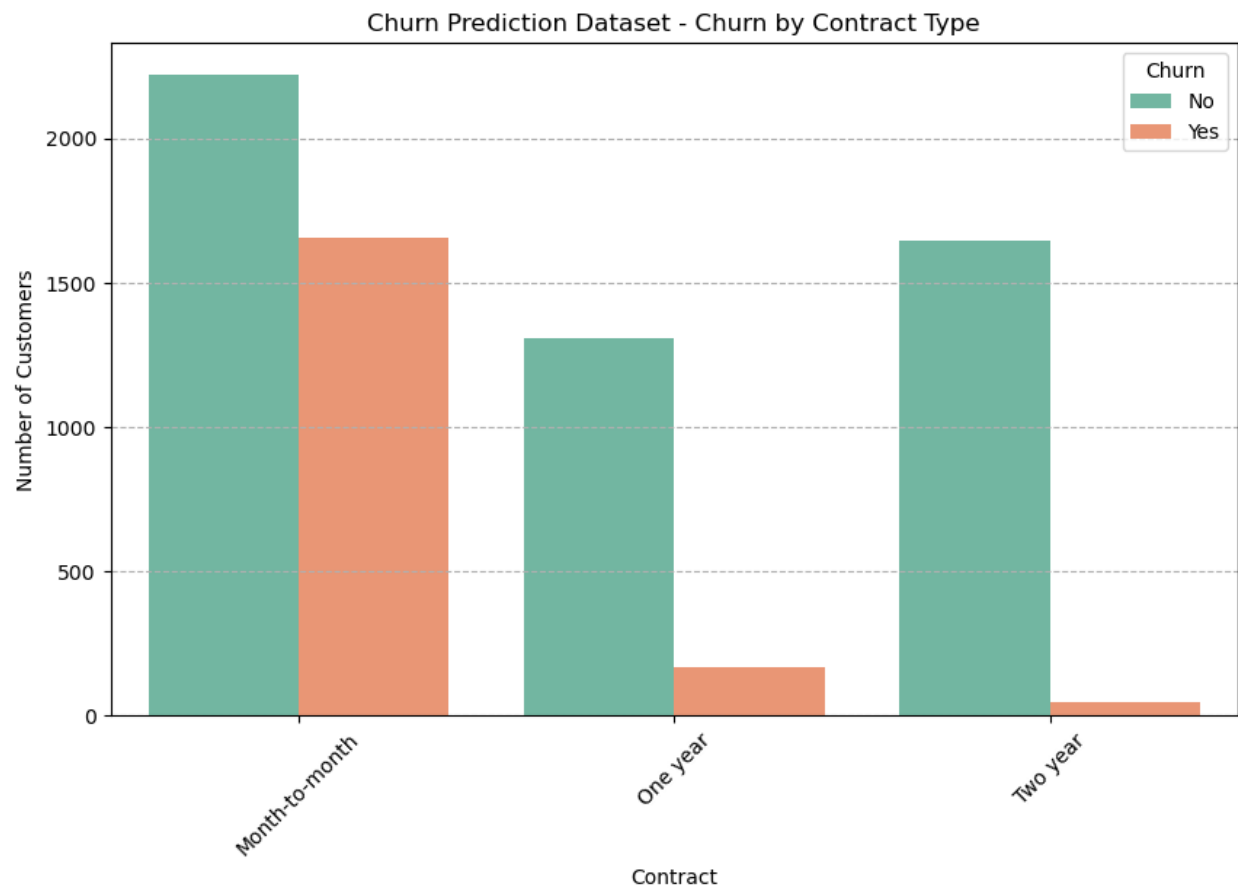
- **Churn Distribution:**
 - Both datasets exhibit a churn rate of approximately **26.5%**, indicating potential



challenges with class imbalance.

- **Contract Types & Churn Relationship:**

- Highest churn observed among **Month-to-month contracts**, suggesting shorter contracts are significantly more prone to churn.



- **Senior Citizen & Monthly Charges Impact:**

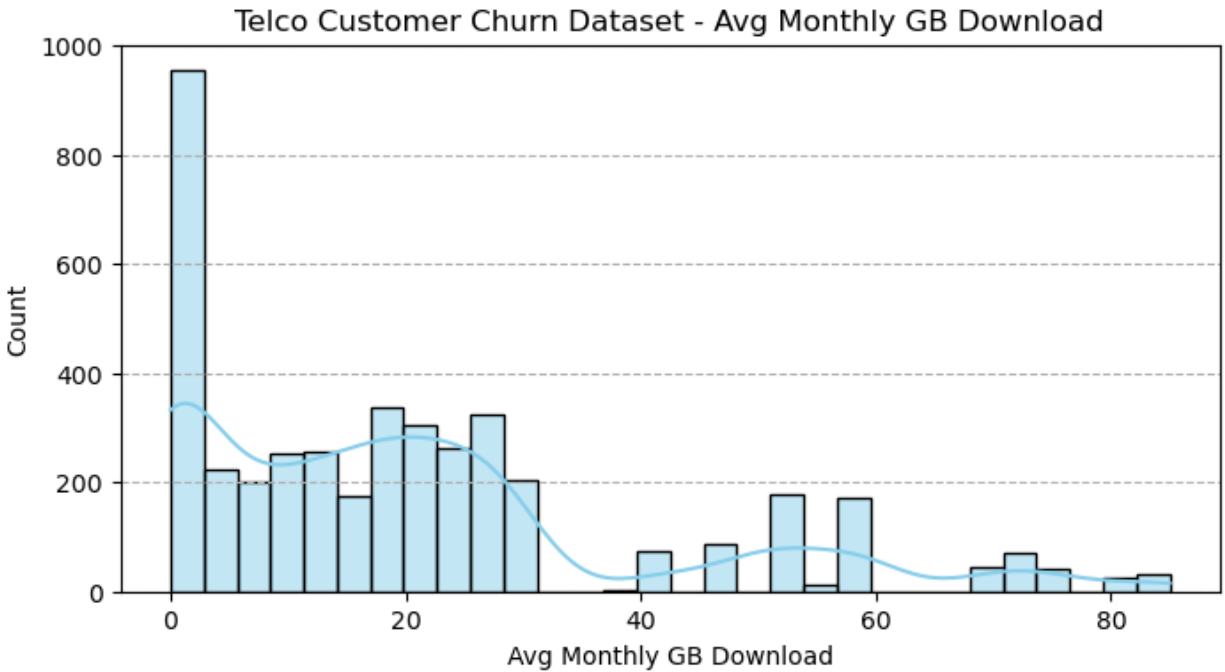
- Customers who are senior citizens and those with higher monthly charges exhibit noticeable churn rates, requiring targeted strategies.

2. Data Issues & Handling Methods:

- **Missing Values:**

- Significant missing data found in fields like Churn Reason, Internet Type, and Offer.

- **Action:** Consider imputation for Internet Type and Offer; categorize missing Churn Reason separately or employ NLP techniques to infer reasons.
- **Outliers Detected:**
 - Identified outliers, especially in Avg Monthly GB Download.
 - **Action:** Use robust statistical methods (winsorization or trimming) to mitigate the impact of outliers on model performance.



2. Define Tools and Technologies

Objective: Select appropriate technologies and tools that facilitate efficient data handling, visualization, modeling, and reporting.

Deliverables:

- **Technology Stack:**
 - **Data Processing & Analysis:** Python (Pandas, NumPy, SciPy)
 - **Visualization:** Matplotlib, Seaborn
 - **Machine Learning:** XGBoost, PyCaret
 - **Database & Storage:** ChromaDB (for embeddings and case retrieval)
 - **Embeddings & NLP:** Sentence Transformers
 - **LLM & RAG:** LangChain and Mistral

- **Dashboarding:** Power BI
- **Integration Plan:**
 - Python for data analysis and cleaning
 - Connect cleaned datasets to visualization and modeling tools
 - Use Python-based NLP for handling textual data and explanations

Next Steps:

- Address identified data issues through cleaning and preparation.
- Initiate modeling with identified key features.
- Implement robust methodologies to handle class imbalance and outliers for accurate model performance.