Assessment 2: Big data project – Phase 1

Mathias Henry Morgan

35087056

**Proposal**

**Business Case:**

My chosen data set (discussed below) and project will address the challenge of customer churn prediction. Customer churn is a critical problem for businesses, particularly those that rely on subscription revenue.

**Motivation:**

Customer churn has significant impacts on business revenue, particularly in the telecommunications industry where there are many competitors with costs for switching providers are low. With a variety of competitors offering similar services, customers can easily switch which makes it difficult to retain a customer base.

**Data Set:**

\*\*\* see attached source file in zip folder

Source: <https://huggingface.co/datasets/d0r1h/customer_churn/viewer> (see references for full citation)

Dataset – Telecommunications Data:

* Volume: 36,992 rows
* Type: CSV
* Features – Data Dictionary
  + Age: Age of the customer
  + Gender: Gender of the customer
  + Security\_no: Unique client identifier
  + Region\_category: Type of location the customer lives (e.g City, Town)
  + Membership\_category: Member type if any
  + Joining\_date: Date customer joined the company
  + Joined\_through\_referral: Binary field if they joined through referral
  + Referral\_id: Unique identifier for referral
  + Preferred\_offer\_types: Customer preferred offers
  + Medium\_of\_operation: Type of device customer uses
  + Internet\_option: Type of Internet (e.g Wi-Fi, Fiber Optic)
  + Last\_visit\_time: Last time client used services
  + Days\_since\_last\_login: Days since last logging on
  + Avg\_time\_spent: Average time spent using services (minutes)
  + Avg\_transaction\_value: Average amount spent
  + Avg\_frequency\_login\_days: Average days spent using services
  + Points\_in\_wallet: Points accrued
  + Used\_special\_Discount: Binary field if special discount has been used
  + Offer\_application\_preference: Binary field for application preference
  + Past\_complaint: Binary field if a past complaint has been logged
  + Complaint\_status: Current complain status
  + Feedback: Customer feedback
  + Churn\_risk\_score: Binary churn risk score (Dependent Variable)

Assumptions

* Data errors could re-occur, wrangling/cleaning process will only handle **visible** and **foreseeable** data issues and not unknown or nuanced edge cases.
* Columns and values within fields that often repeat will remain static allowing for some hardcoded mappings in the wrangling phase.

Owner:

* Pawan Trivedi (d0r1h): <https://huggingface.co/d0r1h>

Date:

* Last Updated May 7, 2022

Past Projects

* N/A

**Approach:**

The general approach is (as the assignment suggests) divide the work across two phases:

**Wrangling** – Create a re-usable Jupyter notebook script that will clean and wrangle the data based on certain conditions. Separating each task into discrete functions with one main function to execute the pipeline/script for maximum re-usability and automation. Ensuring all erroneous data is removed, relevant data is imputed, and values are transformed and encoded as needed for the modelling phase.

**Modelling** – Create a re-usable Juypter notebooks script, use gradient boosting model (e.g XG Boost) to predict customers at risk of churn and assess model performance using a subset of the data for training and other subset for testing. The process will begin with all features mentioned and identify those which are the most statistically significant to achieve better performance, will reduce features to those with greatest predictive power to predict churn risk score and compare predicted and actual results.

**References**

d0r1h. (2022, May 7). *customer\_churn* [Data set]. Hugging Face. https://huggingface.co/datasets/d0r1h/customer\_churn

Khandelwal, N., & Sakalle, V. (2025). Customer churn prediction using XGBoost classifier. *Journal of Information Systems and Enterprise Management, 10*(19s). https://doi.org/10.52783/jisem.v10i19s.3111