

## **1. Problem Formulation**

### **1.1 Business Problem Statement for Customer Churn Prediction**

Our business seeks to proactively identify customers at high risk of churning—those who may stop using our products or services in the near future—even though interventions could retain them. By analysing customer data on demographics, behaviour, and engagement, we aim to predict which customers are likely to leave and develop targeted retention strategies (e.g., personalized offers, improved support) to decrease churn rate, increase customer lifetime value, and maintain a profitable, loyal customer base.

### **1.2 Key Business Objectives**

Reduce the addressable churn rate through timely interventions.

Increase customer retention and loyalty.

Boost revenue and overall business health by minimizing loss of customers.

Enable data-driven and targeted retention campaigns for at-risk customer segments.

### **1.3 Key data sources and their attributes**

HuggingFace API - <https://datasets-server.huggingface.co/rows/scikit-learn/churn-prediction>

Kaggle API - <https://www.kaggle.com/api/v1/datasets/download/blastchar/telco-customer-churn>

### **1.4 Expected Outputs from the Pipeline**

1. Machine learning-ready, transformed feature dataset.
2. Churn prediction model performance report (including metrics such as accuracy, recall, precision, ROC-AUC).
3. Well-organized and documented source code split into stages (ingestion, cleaning, engineering, modelling, evaluation, orchestration, reporting).

### **1.5 measurable evaluation metrics**

- a. models are saved on each day when new training data is run and each version is saved so that we can revert anytime we want.