BIG DATA AND DATA SCIENCE

FINAL WRITE-UP WITH IMPLEMENTATION WORKFLOW

Name: Jhansi Balla

Student Id: 11801681

**Title**: Forecasting Customer Churn in Subscription-based Businesses

**Research Question:**

This project is concerned with developing a machine learning model to predict when the customer will churn and identify important demographic and behavioral factors. It will draw upon a dataset of customer demographics, subscription information, billing history, usage patterns, and customer support interactions. Logistic Regression, Random Forest, and XGBoost will be applied to predict the outcomes and analyze data to identify trends. It would include data from public datasets, CRM APIs, product analytics, and surveys. “*How can machine learning models, combined with demographic, behavioral, and subscription data, be used to accurately predict customer churn and identify key factors influencing retention in subscription-based businesses?*” The solution should be able to grow and handle big data with tools like Apache Spark, so it will help businesses cut losses, improve customer engagement, and enhance profitability.

**DataSet:**

In the prediction of customer churn for subscription businesses, the dataset should contain demographic data such as age, gender, and location to understand behavioral trends, while it should also have subscription information regarding the type and length of a subscription to verify loyalty. It should contain billing history, including payment behavior and failed transactions, to show the risk of churn, and usage patterns such as login frequency and feature usage that would indicate engagement. Customer support interactions include the number of support requests and the resolution time, which indicate the customer's satisfaction. For our prediction target variable, a simple churn indicator- either retained or churned-will be used. The detailed dataset will explain what the causes of churn are and allow for an accurate predictive modeling.

The synthetic customer dataset contains 1,000 entries with 10 numerical features and a binary target variable (`churn`). These features are labeled `Feature\_1` to `Feature\_10`, which can stand for customer attributes or behaviors, while the column named `churn` identifies whether a customer has churned (1) or been retained (0). The dataset has no missing values, and therefore direct exploration and modeling are possible. However, further analysis would be required to map these features to meaningful business metrics.

**Methodologies:**

The methodology for forecasting **customer churn** involves **data preparation**, **exploratory analysis**, **modeling**, and **deployment**. **Data preparation** includes **cleaning**, **feature engineering**, and splitting into **training**, **validation**, and **test sets**. Models such as **Logistic Regression**, **Random Forest**, and **XGBoost** are trained and optimized using **cross-validation**. Evaluation metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC** guide **model selection**. **Feature importance** techniques provide **actionable insights** to address churn. This, in turn, deploys the best model and its continuous **monitoring** for **performance**, while enabling focused **retention strategies** against **high-risk customers**.

**Conclusion:**

The expected outcome for a customer churn prediction project is a robust machine learning model, capable of foreseeing which customers are likely to leave a subscription-based business. The model is supposed to be trained and validated on historical data for good generalization on unseen cases and evaluated using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.

The key deliverable is a churn probability score for each customer, which can be used to classify customers into risk categories such as High, Medium, and Low. This actionable insight enables targeted interventions, such as offering discounts or personalized engagement to high-risk customers. Additionally, the project outputs reproducible code, scalable pipelines, visualizations (e.g., confusion matrix, ROC curves), and a feature importance analysis to explain the factors driving churn, empowering informed decision-making for churn reduction.