Sri Lanka Institute of Information Technology

Statement of work (SOW)

Mini Miners

Fundamentals of Data Mining Project (IT3051)

Year 3, Semester 1, 2025

Batch ID : Y3S1- Group 1.2 (WE)

Group ID : FDM_MLB_G16

IT Number	Student Name	Student E-mail Address	Contact Number
IT23256378	Kalubowila K. S. U	IT23256378@my.sliit.lk	0716529676
IT23209152	Liyanage D S	IT23209152@my.sliit.lk	0768186991
IT23242418	Bogahawatta M O	IT23242418@my.sliit.lk	0769105626
IT23232136	Gunasekara P A H I	IT23232136@my.sliit.lk	0758580439

Table of Contents

1. Background

Customer churn is a major challenge in the banking sector, where losing clients leads to significant revenue loss and higher acquisition costs. With global churn rates averaging 10–15% annually, and even higher in digital-first banks, retaining existing customers is crucial, especially in developing markets with limited growth opportunities.

This project addresses the issue by applying data mining and machine learning techniques to predict churn based on customer demographics, financial behavior, and transaction patterns. It is academically motivated by the need to apply theoretical knowledge in data science to real-world challenges, covering data mining, machine learning, business analytics, and software development.

The churn prediction problem is well-suited for supervised learning, offering clear classification targets, strong business relevance, and scalability across institutions and markets.

2. Scope of work

This research aims to design and implement a churn prediction system that leverages historical customer records to identify individuals at risk of leaving a banking service. The study will be conducted through the following stages:

1. Data Collection and Preparation

- Review customer demographic, financial, and behavioral attributes.
- Handle missing values, outliers, and duplicate records to ensure data quality.
- Encode categorical variables, scale numerical features, and partition the dataset into training, validation, and test subsets.

2. Exploratory Analysis

- Conduct univariate, bivariate, and multivariate analysis to understand customer behavior and churn drivers.
- Segment customers by demographics and activity patterns.
- Apply statistical summaries and visualizations to reveal trends and correlations.

3. Feature Engineering

- Create new indicators such as account tenure, transaction frequency, and variations in financial activity.
- Identify the most influential variables contributing to churn through feature selection techniques.

4. Model Development

- Implement multiple supervised learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and others).
- Apply cross-validation, hyperparameter tuning, and class balancing methods to address data imbalance.

5. Model Validation

- Evaluate predictive performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- Compare models to select the most effective solution for churn prediction.

6. System Deployment

- Build a RESTful API to generate churn predictions.
- Develop an interactive dashboard for monitoring churn risk scores and business insights.
- Package the solution in a containerized environment for ease of deployment.

7. Business Intelligence Application

- Provide churn risk scoring and customer lists for batch processing.
- Deliver trend analysis and actionable insights to support customer retention strategies.
- Outline a framework for assessing potential ROI from predictive insights.

8. Anticipated Outcomes

- Accurate identification of high-risk customers.
- Data-driven support for retention initiatives.
- Improved decision-making through predictive analytics and business intelligence reporting.

3. Activities

- **Literature Review**

The project will begin with a comprehensive review of existing research on machine learning techniques for customer churn prediction. Key focus areas include identifying the most effective algorithms (e.g., logistic regression, ensemble methods, neural networks), understanding evaluation criteria, and analyzing challenges such as class imbalance and fairness. The review will also highlight gaps in prior research, providing justification for the proposed methodology.

- **Data Exploration & Understanding**

The dataset will be examined to gain a thorough understanding of its structure, variable types, and distribution. Descriptive statistics and visualization techniques will be applied to uncover correlations, churn patterns, and demographic or behavioral trends. Potential data quality issues such as missing values, duplicates, or imbalances between churned and retained customers will be identified during this stage.

- **Data Preprocessing**

Data cleaning techniques will be applied to handle missing values, duplicates, and inconsistencies. Numerical features will be normalized, and categorical variables will be encoded using suitable strategies. To address class imbalance, advanced resampling methods such as SMOTE (Synthetic Minority Oversampling Technique) and under-sampling will be employed, ensuring the models are trained on balanced datasets.

- **Model Development**

A variety of classification models will be implemented, including Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks. The Fairlearn library will be integrated to evaluate and mitigate bias across sensitive attributes such as gender or geography. Fairness constraints will be applied to ensure equitable predictions while maintaining accuracy, aligning the system with ethical AI practices.

- **Model Evaluation & Selection**

Models will be evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and fairness indicators from Fairlearn. Comparative analysis will be performed to identify the most balanced model in terms of both predictive performance and fairness. Trade-offs between fairness and accuracy will be documented to guide decision-making.

- **Software Development**

A prototype churn prediction system will be developed to integrate the chosen machine learning model. The system will include features for data input, real-time and batch churn predictions, and visualization of results (e.g., churn risk scores, performance dashboards). The software design will emphasize scalability, maintainability, and usability for banking stakeholders.

- **Testing & Validation**

The system will undergo thorough testing, including unit tests for individual modules and integration testing for the complete pipeline. Robustness will be validated across multiple datasets and scenarios to ensure consistent performance. Additionally, fairness-aware testing will be conducted to confirm that predictions meet ethical AI standards.

- **Documentation & Presentation**

Comprehensive documentation will be prepared, covering methodology, design decisions, results, and system usage. Business-focused reports and interactive dashboards will be developed for stakeholders to interpret findings. Finally, a structured presentation and live demonstration will be delivered, showcasing the problem statement, methodology, model performance, system functionality, and business value.

4. Approach

The project will follow a structured and phased methodology aligned with the defined objectives. Each phase ensures a systematic transition to the next, maintaining accuracy, efficiency, and practical applicability in churn prediction.

1. Literature Review

To establish a strong foundation, the project will begin with a review of scholarly articles, industry reports, and case studies related to customer churn prediction in banking. This stage will inform the selection of machine learning algorithms, evaluation metrics, and feature engineering techniques.

2. Data Exploration & Understanding

The historical customer dataset will be explored to understand its structure, distribution, and churn-related trends. This stage will highlight key attributes, detect class imbalances, identify missing values, and uncover correlations that may influence churn behavior.

3. Data Preprocessing

The dataset will undergo systematic preprocessing, including handling missing values, detecting and treating outliers, encoding categorical variables, scaling numerical features, and balancing churn classes. These steps ensure data integrity and consistency, enabling accurate modeling.

4. Feature Engineering

Domain knowledge and exploratory insights will be used to construct new features such as account tenure, transaction frequency, and changes in spending activity. Feature selection methods will then identify the most influential predictors of churn.

5. Model Development

Multiple supervised machine learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Neural Networks) will be implemented and iteratively improved. Hyperparameter tuning and cross-validation will be applied to optimize performance.

6. Model Evaluation & Selection

The developed models will be assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Comparative analysis will identify the model that provides the best trade-off between predictive performance and generalizability.

7. Software Development

The selected churn prediction model will be integrated into a functional prototype. A RESTful API will be built to serve predictions, and a user-friendly dashboard will be developed to visualize churn risk scores, customer insights, and business intelligence reports.

8. Testing & Validation

The prototype will undergo rigorous testing, including unit, integration, and system-level tests. Model stability will be validated against unseen datasets to ensure reliability and robustness. Performance optimization will also be carried out to ensure scalability.

9. Documentation & Presentation

All project stages will be thoroughly documented to ensure reproducibility and transparency. Final outcomes, insights, and the working prototype will be presented through a comprehensive report and presentation tailored for both technical and business stakeholders.

10.Hosting & Development

The churn prediction prototype will be deployed on a cloud platform (e.g., Heroku, AWS, or Azure) with optional Docker support for scalability. The frontend will be integrated with the backend model for real-time predictions, accessible only to authorized users. Performance will be monitored, models evaluated using key metrics, and feature importance analyzed to provide actionable insights through an interactive dashboard.

5. Deliverables

The following items will be submitted for the final evaluation of the project:

1. Statement of Work (SOW)

A formal document outlining the project background, scope, objectives, methodology, approach, timeline, and team responsibilities. This will serve as the foundational plan for the project execution.

2. Final Report

A comprehensive report detailing the problem statement, dataset description, preprocessing steps, exploratory analysis, feature engineering, model development, evaluation results, business insights, and final conclusions.

3. Software Solution

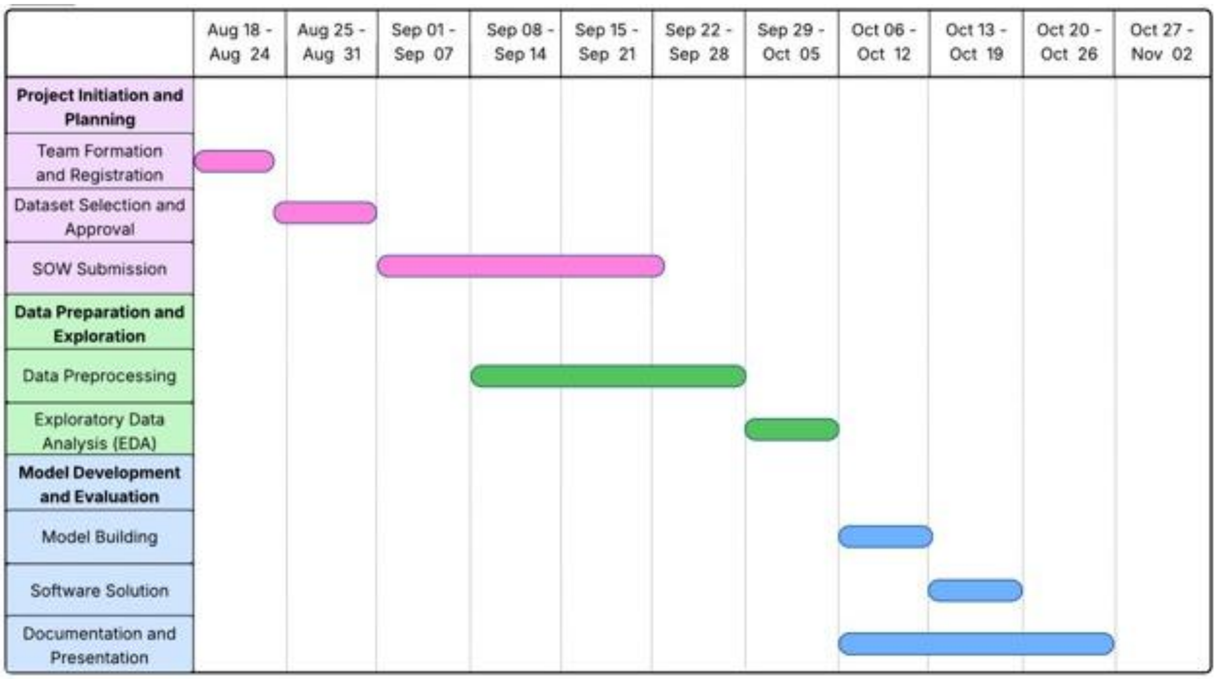
A functional churn prediction system comprising:

- A RESTful API for generating predictions,
- An interactive dashboard for visualizing churn risk and customer insights,
- A deployable containerized package ensuring scalability and ease of deployment.

4. Video Presentation

A recorded presentation summarizing the project's objectives, methodology, findings, and showcasing the final software solution through a live demonstration.

6. Project Plan & Timeline



	Aug 18 - Aug 24	Aug 25 - Aug 31	Sep 01 - Sep 07	Sep 08 - Sep 14	Sep 15 - Sep 21	Sep 22 - Sep 28	Sep 29 - Oct 05	Oct 06 - Oct 12	Oct 13 - Oct 19	Oct 20 - Oct 26	Oct 27 - Nov 02
Project Initiation and Planning											
Team Formation and Registration	■										
Dataset Selection and Approval		■									
SOW Submission			■	■	■						
Data Preparation and Exploration											
Data Preprocessing				■	■	■					
Exploratory Data Analysis (EDA)							■				
Model Development and Evaluation											
Model Building								■			
Software Solution									■		
Documentation and Presentation								■	■	■	

7. Assumptions

The project assumes that the provided dataset (“botswana_bank_customer_churn.csv”) is valid, sufficient, and representative of customer churn behavior, with missing values handled effectively during preprocessing. The dataset’s behavioral trends are stable, enabling accurate churn modeling. Python libraries (pandas, scikit-learn, matplotlib, seaborn, Flask) will support development, and predictions are understood as probabilistic outcomes based solely on the given dataset.

Key milestones include project foundation and dataset validation (Week 2), data preparation and EDA completion (Week 4), model development and evaluation (Week 6), system integration with API and dashboard (Week 8), and final submission with report and presentation (Week 10).

Risks such as dataset quality issues, underperforming models, API complexity, timeline delays, and resource constraints are acknowledged, with mitigation through preprocessing, multiple algorithms, incremental development, buffer time, and optimized computing resources.

Linkes to Dataset

Git Hub Link: https://github.com/SusheniUmayangana/Customer_churn.git

Kaggle Link: https://www.kaggle.com/datasets/sandiledesmondmfazi/bank-customer-churn

8. Project team, roles, and responsibilities

Student ID	Full Name	Primary Role	Responsibilities
IT23256378	Kalubowila K. S. U	Data Collection, Data Preprocessing Lead & Team Coordinator	• Collect the dataset(s). • Clean missing values, duplicates, and outliers. • Perform feature engineering (create new meaningful variables). • Apply data normalization and encoding.
IT23209152	Liyanage D S	Exploratory Data Analysis (EDA) & Visualization	• Analyze dataset structure and characteristics. • Summarize descriptive statistics. Visualize churn patterns (charts, heatmaps, correlation analysis). • Identify key factors influencing churn
IT23242418	Bogahawatta M O	Modeling & Prediction	• Apply machine learning models (e.g., Logistic Regression, Decision Tree, Random Forest, XGBoost). • Tune hyperparameters for better performance. • Compare model performance using metrics such as accuracy, precision, recall, F1-score, and AUC.
IT23232136	Gunasekara P A H I	Evaluation, Reporting & Deployment Prep	• Conduct final evaluation of the best-performing model. • Document results and insights clearly. • Prepare the final report and presentation slides. • Optionally, prepare a simple deployment (e.g., Streamlit app or API)