Project Proposal Plan

Project Summary

This project aims to address operational inefficiencies in the loan application process by developing an automated loan eligibility prediction data product. The existing manual system leads to delays and inconsistencies in eligibility results, which costs Northern Bank to lose potential customers and waste resources that are best used performing other tasks.

By adopting this automated loan eligibility prediction system, Northern Bank can offer a more streamlined and transparent application process to its customers, ensure consistency across loan eligibility predictions, and conserve costly resources.

The machine learning prediction model will be delivered in the form of an intuitive, user-friendly application that benefit both Northern Bank and its potential customers by offering a streamlined loan approval workflow. In addition, a comprehensive user guide will be included that will provide a detailed breakdown of the application's workflow and ensure easy adoption for all users.

Data Summary

The raw data for this project will be sourced from a comprehensive historical loan dataset on Kaggle.com, called Eligibility Prediction for Loan, and incorporates a broad spectrum of demographic and financial variables to ensure relevance to real-world scenarios [1].

The development life cycle for the loan eligibility prediction system will take place in phases, to ensure effective data processing and management. The initial phase in the life cycle is the data collection phase, where the dataset will be loaded into the application, and brief analysis is done to gain familiarity with the dataset. Then comes the design phase, where exploratory data analysis (EDA) will be conducted to gain insights into dataset, such as relationships, potential patterns, and the distribution of features. This phase lays the foundational understanding of the data and informs decisions to be made in the development phase. Moving into the development phase, we will begin preprocessing of the dataset, including handling missing values, encoding categorical variables, and normalizing features of the dataset. By performing these steps, we will ensure quality data is input to the machine learning model, allowing it to provide reliable predictions.

Ethical considerations for this project will include encryption of personal identifiers to protect privacy and attempting to mitigate biases in the machine learning model. By adhering to data protection regulations and obtaining applicant consent for the use of data, we will ensure the ethical development of this loan eligibility prediction model.

Implementation

Implementation of the Loan Eligibility Prediction system, we will adhere to the industry-standard Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.

Business Understanding

The initial phase involves collaborating with stakeholders to define objectives and requirements for the project and establish key performance indicators (KPIs) for the Loan Eligibility Prediction system, ensuring a clear understanding of the desired outcomes.

Data Understanding

This phase focuses on the collection of data, exploration of the historical loan dataset to gain insights into feature distributions and relationships, data quality and anomaly issues, and sets the foundation for preprocessing.

Data Preparation

The data preparation phase focuses on the preprocessing of the dataset and is where we will address missing values, handle outliers, and perform any necessary encoding for categorical variables. With a clean dataset we will then split the dataset into two sets for use in training and testing the model.

Modeling

In the modeling phase, we will select a suitable machine learning algorithm for predicting loan eligibility, train the model with the training dataset, and aim for optimal accuracy.

Evaluation

Evaluate model comparisons and perform cross-validation to obtain evaluation metrics for each algorithm that is tested. Confusion matrices will be utilized to visualize model performance.

Deployment

The final phase is the deployment of the model, which will involve integrating the loan eligibility prediction model into the loan application. User training will be conducted, and the model will be deployed for public use.

Timeline

Milestone / Deliverable	Duration	Start Date	End Date
Proof of Concept	1 week	December 15, 2023	December 22, 2023
Data Collection and Preparation	2 weeks	December 25, 2023	January 5, 2024
Model Development and Training	2 weeks	January 8, 2024	January 19, 2024
Model Evaluation and Testing, Integration	2 weeks	January 22, 2024	February 2, 2024
Application Delivery and Deployment	1 week	February 5, 2023	February 9, 2024

Evaluation Plan

Throughout the phases of development, various verification methods will be applied to ensure the deliverable meets the specified requirements. Data profiling will be used to validate the consistency of the collected data and its alignment with the objectives. Insights gained from Exploratory Data Analysis (EDA) will be verified using visualizations and unit tests will be performed during data preprocessing, ensuring proper encoding and handling of missing values. In training the model, cross-validation and performance metric evaluation will take place to verify the effectiveness of the machine learning model.

Upon completion of the project, user acceptance testing (UAT) will be utilized to ensure that the application is meeting end-users' expectations. End-to-End testing will be used to verify the seamless workflow of the loan eligibility application, ensuring accurate and reliable results.

Resources and Costs

Resource	Description	Cost
Hardware	Workstations are already owned. Network is already provisioned.	$0
Software	Jupyter Notebook, Python 3.11, and all libraries necessary for this project are free and open source.	$0
Cloud Infrastructure	Necessary cloud infrastructure for model deployment and hosting, yearly cost.	$9,000
Data Scientist	Engineer to perform data analysis and processing. $63/hr x 80 hours	$5,040
Machine Learning Engineer	Engineer to develop and fine-tune the machine learning model. $60/hr x 80 hours	$4,800
Software Engineer	Engineer to perform model integration and testing. $71/hr x 80 hours	$5,680
User Training	No specific user training is required.	$0
	Total	$24,520