Advanced AI/ML Topic Approval Form

The purpose of this document is to help you clearly explain your AI/ML topic, project scope, and timeline. Identify each of these areas so that you will have a complete and realistic overview of your project. Your instructor cannot sign off on your project topic without this information.

Note: You must fill out and submit this form. Space beneath each number will expand as needed.

Note: Any costs associated with developing the application will be the responsibility of the student.

INFORM INSTRUCTOR:

Potential use of proprietary company information: (Y/N)
N/A

PROBLEM AND GOAL:

1. Explain the business problem or organizational need that you identified.

   The business problem I have decided to pursue is customer churn. While this is popular topic, I'm sure, it is for good reason. In today's day in age most services are subscription based. Many organizations struggle to accurately predict customer churn, resulting in reactive rather than proactive retention strategies. High customer churn leads to increased acquisition costs, reduced revenue stability, and inefficient allocation of marketing resources. Businesses often have access to large volumes of customer data but lack predictive tools that transform this data into actionable insights. An AI/ML churn prediction system can help organizations identify at-risk customers early and apply targeted interventions to improve retention and customer satisfaction.

2. Identify one SMART goal of the proposed project using a SBI goal format.

   Situation: Customer churn negatively impacts revenue and long-term business sustainability.

   Behavior: Develop a machine learning model that predicts customer churn using historical customer behavior data (Classification: did they cancel subscription or no).

   Impact: Enable proactive retention strategies that reduce churn and improve customer lifetime value. An example of this may be targeted marketing strategies for those identified by the model as "at risk" of churning.

Smart Goal:

S: Design, train, and evaluate a random forest churn prediction model using my publicly available dataset.

M: The goal will be measurable by achieving at least 80% accuracy on my chosen evaluation metrics.

A: It is achievable as we will be using proven ML techniques on a detailed dataset.

R: Customer retention being a large part of business makes this project/goal relevant.

T: The project timeline is scheduled over five working days/sessions.

DATASET AND ANALYTICAL METHOD:

3. Briefly describe the dataset you will be working with by discussing the format, size, and features. Please also cite the source(s).

   I will be using the Telco Customer Churn Dataset from Kaggle.com – Kaggle. (2018). *Telco Customer Churn Dataset*.
   https://www.kaggle.com/blastchar/telco-customer-churn
   The dataset is a telecom/internet company that offers services like phone, internet, and media streaming.

   This dataset is formatted in CSV making it easy to work with upon download. It contains 7044 records with demographic features including gender, senior(y/n), partner, dependents. Account info features such as account, tenure, contract type, and payment method. It also includes a customer's service usage such as internet, streaming, and phone service. Lastly the dataset provides our target variable "Churn" (y/n) which is customers who have left within the last month.

   There is some correlation between some of the features in the data visible right away. A customer who does not opt into "InternetService" results in a null value for the features "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV" and "StreamingMovies". An obvious correlation is also service usage as it is related to the customer's monthly bill and total charges.

5. Identify the AI/ML analytical method you plan to use.

   Identifying the method is simple for churn, did the customer cancel services or not? Given the yes or no scenario on labeled and structured data we will be using supervised machine learning classification. Specifically Logistic regression as a baseline model and Random Forest Classifier as our primary tuned model. The reasoning for choosing random forest is its strong performance with tabular data and has built in feature importance which we will need to figure out which columns are

important in predicting churn. Will plan to pursue another model if results are not satisfactory.

Though it is known that the Random Forest Classifier (RFC) performs well on tabular data, there are some disadvantages. Based on my research it appears that the RFC seems to perform better in most cases when using larger data sets when the model has more data to produce more decision trees. This dataset, having 7,000 records, could be seen as a smaller dataset based on some of the ones I've seen on Kaggle.com. Another key disadvantage is its interpretability. It is clear to understand how the feature importance scores work, but individual decision paths of the trees are difficult to interpret potentially limiting transparency and could lead to some bias we are unaware of.

PROJECT MANAGEMENT AND CHARACTERISTICS

6. Outline major tasks and the schedule of the proposed project in a simple table. This will include anticipated timelines, due dates, costs, etc.

Task	Description	Timeline	Cost
Planning	Define Scope, data requirements	Day 1	0.00
Data Clean & Exploratory data analysis	Clean, process, and explore dataset	Day 2	0.00
Model Dev.	Split data, train and tune models.	Day 3	0.00
Evaluation	Evaluate chosen metrics. (accuracy, precision score, recall, confusion matrix)	Day 4	0.00
Documentation	Final report and documentation	Day 5	0.00

Identify the functional and nonfunctional requirements of your project.

Functional requirements:

- Load and preprocess data
- Train and evaluate model results
- Generate performance metrics
- Output predictions and feature importance

Nonfunctional requirements:

- Model runs efficiently on colab or Jupyter notebook locally.
- Eval metrics show Accuracy >= 80%, F1 >= 75%, ROC-AUC >= .80

- Code must be readable and reproducible.
- Preservation of data privacy.
- Execution time should not exceed a reasonable limit (<10minutes)

7. Describe any risks to the project and identify mitigation strategies.

   Risk 1: The dataset may contain missing or noisy data that could impact model accuracy and performance.

   Mitigation 1: Preprocessing techniques such as data cleaning, missing value, normalization, and handling any outliers will be applied before model development to prevent such issues.

   Risk 2: The RFC model may overfit the training data resulting in reduced performance.

   Mitigation 2: Cross-validating and parameter tuning will be used to improve performance and reduce overfitting.

   Risk 3: The data set will more than likely be imbalanced containing a lot more non-churn records which will create some bias.

   Mitigation 3: It will be important to not rely on Accuracy as a sole metric but including F1 score and ROC-AUC. One can also evaluate some metrics using class weighting and resampling techniques.

DESIGN AND DEVELOPMENT

1. Identify the programming/development language(s) you will use. Python
2. Describe the programming environment (integrated development environment) including both the software and hardware—you will use to create the application. Jupyter Notebook, Personal computer (64 gb ram, sufficient cpu and gpu. Common laptop or desktop can run these programs)
3. Identify the operating system(s) or platform(s) you will use to complete the project. Windows 11

This project does not involve human subject research and is exempt from WGU IRB review.

STUDENT'S SIGNATURE

Ryan Brunelle 12/31/25 9:44 AM EST dotloop verified XYKP-CDNM-AQWO-YQK1

12/31/2025

By signing and submitting this form, you acknowledge that you (the student) are responsible for any costs associated with developing and executing the application.

INSTRUCTOR'S SIGNATURE:

INSTRUCTOR APPROVAL DATE:

January 2, 2026