**Data Science Capstone Topic Approval Form**

**Student Name:** Kevin Sandoval

**Student ID:** 012278847

**Capstone Project Name:** Predicting Loan Defaulters with Machine Learning

**Project Topic**: Using Machine Learning and Random Forests to predict loan defaulters

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

**Research Question:** What factors contribute towards default rate and can the important factors help predict whether a customer will default?

**Hypothesis**: The factors statistically significantly affect default rate.

**Null Hypothesis**-. The factors do not statistically significantly affect default rate. **Alternate Hypothesis**-. The factors statistically significantly affect default rate.

**Context:** The research question would benefit from a data analysis as finding contributing factors to the default rate could help the company prepare for a specific customer to default or use those to prefer some customers over others. Then using those contributing factors to create a model to predict whether a customer will default will help save the company time, money, and effort.

**Data:** The data I am using is provided by WGU.

The data set includes 34 columns and 148,670 rows of customer data. The data includes an ID for each customer, a “Status” column which indicates whether the customer has defaulted on their loan, as well as many other defining customer characteristics, both quantitative and qualitative.

WGU owns the data and I am allowed to use the data as it is supplied to me specifically for this final project.

**Data Gathering:** As I am using the given WGU dataset, I will not need to gather any data.

**Data Analytics Tools and Techniques**: I will perform appropriate data wrangling, data cleaning, and other data preparation techniques in order to prepare the data for analysis. I will then create a model to help predict if a customer will default on their loan. To do this, I will use the Random Forest Classifier that I used in D603 Task 1. I will then use a confusion matrix and classification report to analyze the performance of the model.

**Justification of Tools/Techniques:** It is appropriate to use the Random Forest Classifier as I am focused on the binary outcome of if a customer defaults or not, and so the Classifier works well for predictions.

**Project Outcomes**: The key anticipated project outcomes and deliverables are to find a model that helps the company predict whether a customer will default. The factors used will be the ones most closely linked to the default rate and the model will use those to assist the company in finding customers who will default.

**Projected Project End Date**: 5/16/2025

**Sources**: Only WGU official resources will be used.

**Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Instructor's Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.