**Northeastern University**

**DS5500 33766 Capstone: Appl in Data Science SEC 01**

**Spring 2024**

**Team or Individual Work:**

* Who are the members of your team? (If applicable)

Haozhou Xia

Amari Parris

Yaseen Ellison

* What roles and responsibilities have you assigned to each member?

This project has several stages: Data collection and transformations, Model implementations and comparisons, Stress testing and model validations, Data visualizations, and result analysis.

We conduct the data transformation together, and each individual of us is responsible for implementing different classification models and reviewing the model performance. For the last stage, we have to analyze the performances to select the most accurate model for the dataset and then visualize the results.

**Objective for the Semester:**

* What specific goals do you aim to achieve by the end of this semester in relation to your capstone project?

The objective is to apply various machine learning models to deepen our understanding of the field. By performing data transformations and cleanings, I plan to transform raw, real-world data into a suitable dataset for self-developed probability of default models. Furthermore, I intend to enhance the accuracy of these probabilities of default models through careful refinement.

**Topic Selection:**

* What topic have you chosen for your capstone project?

Building probability of default models using commercial bank credit risk management

* Why is this topic of interest to you, and how is it relevant to your field of study?

Considering the recent events in the Economics World, such as the collapse of Silicon Valley Bank and the failure of Credit Suisse, the importance of robust model risk has been paid more attention to. Team member Haozhou Xia once interned in the model risk management division but was mainly responsible for replicating the data extraction process and documentation review related to risk management. For this capstone project, we aim to explore different models from a modeler's perspective, seeking ways to enhance the predictive accuracy of these models.

**Problem Specification:**

* What specific problem or challenge does your project aim to address or solve within your chosen topic?

Our research indicates that most commercial banks predominantly use logistic regression models. However, other classification models might be more adept at detecting default risk. We plan to explore options like gradient boosting, random forest, XGBoost, and Light GBM to enhance the accuracy of models.

* How does this problem relate to the current state of research or industry needs in your field?

We believe that successfully completing this project will enrich our expertise and significantly contribute to risk management in the banking industry. Our team, with its background in finance and banking, is eager to apply data science knowledge to real-world financial challenges. Additionally, as digital transformation and big data applications become increasingly prevalent across various industries, including banking, achieving enhanced accuracy in predicting loan default probabilities could be a notable innovation in the sector.

**Significance of the Problem:**

* Why is addressing this problem important? What impact do you anticipate your project will have on the field, industry, or society?

**Feasibility and Scope:**

* Considering your resources, skills, and timeline, how feasible is it to address this problem within the scope of your capstone project?

We have a clear picture of the different steps in this capstone:

1. Data collection and transformations,
2. Model implementations and comparisons,
3. Stress testing and model validations,
4. Data visualizations and result analysis.

We have access to a comprehensive dataset free from concerns regarding data privacy or ethical issues. In the first 1 to 1.5 months, our aim is to thoroughly understand this dataset and conduct the necessary data transformations. Over the following two months, we plan to implement various classification models and compare their accuracies. We aim to enhance model accuracy, which we belie. However, if the performances of these models do not show significant differences, we will also undertake a detailed analysis to understand the reasons behind this outcome.

**Objective Setting:**

* What are the primary objectives of your research?
* How do these objectives contribute to the field or topic you are studying?

**Initial Literature Review:**

* List at least 8 research papers you have reviewed.
* How does each pape`r contribute to your understanding of your chosen topic?

**Dataset Selection:**

* What type of dataset are you planning to use?
* Why have you chosen this specific dataset, and how is it relevant to your research objectives?

**Research Methodology:**

* What research methodology (qualitative, quantitative, or mixed methods) have you chosen?
* Explain your rationale for choosing this methodology.

**Implementation Tools:**

* What programming language(s) will you be using for your project?
* What libraries or tools do you plan to utilize, and why are they necessary for your project?