Data Science Capstone Topic Approval Form

Student Name: Bernard Connelly

Student ID: 012300379

Capstone Project Name: Detecting Anomalous SEC Filings to Identify Financial Risk

Project Topic: This project uses unsupervised machine learning to analyze public financial statement data filed with the Securities and Exchange Commission (SEC). The objective is to detect filings that deviate significantly from industry norms or historical behavior, which may signal elevated financial risk or reporting irregularities. The Isolation Forest algorithm will be applied to core financial metrics such as assets, liabilities, revenues, and net income from 10-K filings between 2023 and 2025.

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

Research Question: Can unsupervised machine learning identify anomalous patterns in standardized financial filings that may indicate potential financial irregularity, risk exposure, or reporting inconsistencies?

Hypothesis:

Null Hypothesis- There is no statistically detectable anomaly pattern in financial filings that differentiates high-risk or irregular companies from their peers.
Alternate Hypothesis- Statistically detectable anomalies exist in certain financial filings, indicating outliers in financial reporting that may reflect financial risk or irregularities.

Context: The financial collapse of firms often stems from deteriorating fundamental details that may go unnoticed in time-series data, but appear clearly when compared to peer benchmarks. With the increasing availability of structured financial data from the SEC, it's now possible to apply machine learning techniques to detect outliers across balance sheet and income statement metrics. This project will use Isolation Forests to flag anomalous SEC filings that warrant deeper scrutiny. These anomalies could represent aggressive accounting, earnings manipulation, financial instability, or sector outliers — insights useful for regulators, investors, and analysts.

Data: The data comes from the SEC's Financial Statement Data Sets, which contain quarterly and annual structured XBRL filings from all publicly traded companies. Key files used include sub.txt (submission metadata) and num.txt (financial values). Focus will be on 10-K (annual) filings from 2023–2025, filtered to include commonly reported financial tags such as Assets, Liabilities, NetIncomeLoss, Revenues, and StockholdersEquity.

The data is owned and maintained by the U.S. Securities and Exchange Commission. It is publicly available for unrestricted use under open government data policies, making it fully eligible for use in this academic project.

Data Gathering: Data was manually downloaded from the SEC's website across nine quarterly reporting periods from Q1 2023 through Q1 2025, then unzipped and ingested into Python using pandas. Two key files (sub.txt and num.txt) were merged and reshaped into a single DataFrame where each row represents a company filing and columns represent key financial metrics.

Data Analytics Tools and Techniques: As noted above, unsupervised anomaly detection using Isolation Forests will be utilized primarily for this project. The data will be preprocessed using feature scaling and will include calculated financial ratios. A variety of python libraries will be used to achieve these tasks, including pandas, scikit-learn and matplotlib to visualize the data.

Justification of Tools/Techniques: Isolation Forest is well-suited for high-dimensional numeric data where the goal is to isolate rare or unexpected patterns. It does not require labeled data, making it ideal for financial filings where "fraud" or "failure" is not explicitly tagged. The model is also easily interpretable and lightweight, making it practical for large-scale datasets like SEC filings.

Project Outcomes:

- A cleaned, structured dataset of company filings (2023–2025) with standardized financial features
- A trained Isolation Forest model to flag potential anomalies
- A ranked list of anomalous filings for further investigation
- Visualizations showing how flagged companies differ from peers
- A capstone report explaining methodology, model design, findings, and limitations
-

Projected Project End Date: June 30, 2025.

Sources:

U.S. Securities and Exchange Commission. “Financial Statement Data Sets.”
https://www.sec.gov/data/financial-statement-data-sets

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. June 2, 2025

Reviewed by: Daniel J. Smith, PhD, MBA

Comments: