Project Report: Predicting Going Concern Risk Using Financial Ratios

1. Objective This project aimed to develop a machine learning model to predict whether a company is at risk of receiving a going concern (GC) opinion based solely on financial ratios. The use case simulates the role of data-driven risk flagging systems relevant in audit, assurance, and credit ratings contexts, such as those employed at Moody's.

2. Data Overview The dataset consists of 38,252 firm-year observations and 70 financial variables, believed to originate from WRDS and Audit Analytics. The target variable is binary:

- `1`: Company received a GC opinion
- `0`: Company received a clean audit opinion

3. Methodology

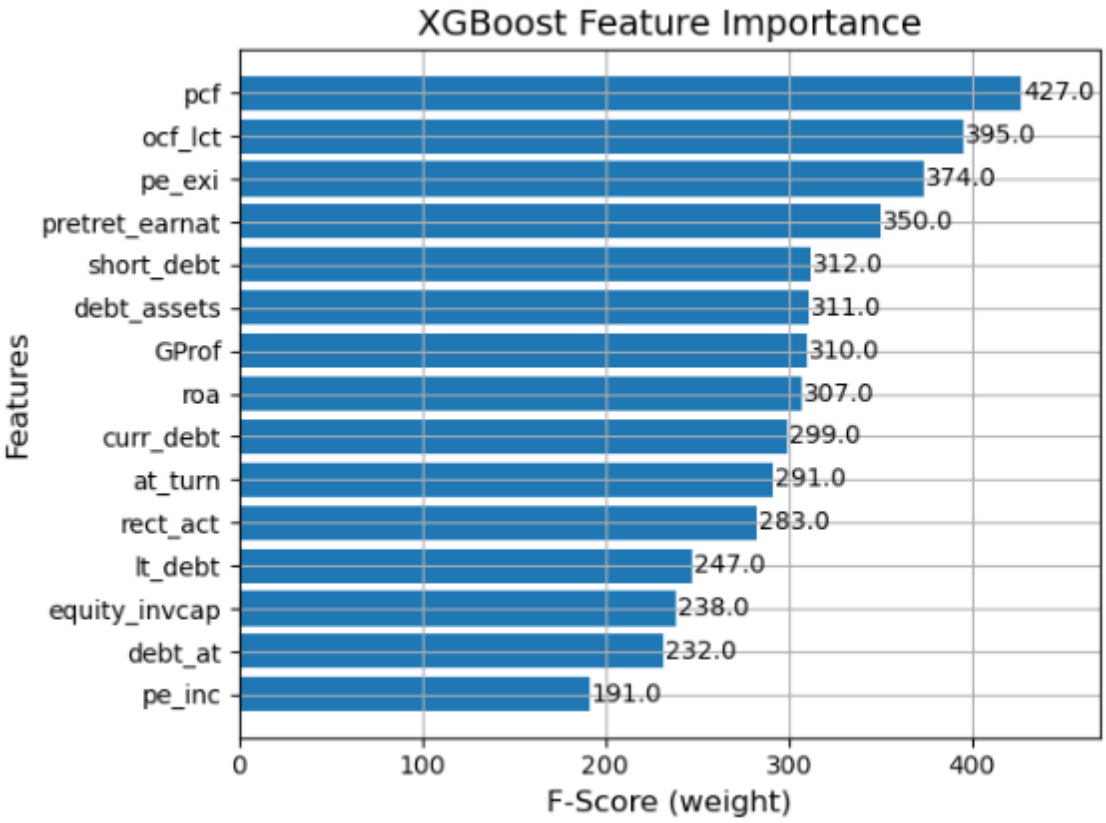
- **Data Preprocessing:**
  - Dropped identifier variables: `cid`, `fyear`, `datadate`
  - Applied median imputation to missing values
  - Selected the top 15 financial ratios using ANOVA F-test (`SelectKBest`)
- **Modeling Approaches:**
  - Logistic Regression with class weight balancing
  - XGBoost with SMOTE (Synthetic Minority Oversampling)
  - XGBoost with threshold tuning to optimize recall and F1 score
- **Class Imbalance Handling:**
  - Applied SMOTE to oversample minority class (GC cases)
  - Tuned `scale_pos_weight` in XGBoost for better balance

4. Model Performance

Model	Precision (GC)	Recall (GC)	F1 Score (GC)	Accuracy
Logistic Regression	0.24	0.84	0.37	0.86
XGBoost + SMOTE	0.34	0.64	0.45	0.93
XGBoost (Max Recall Threshold)	0.17	0.96	0.29	0.77
XGBoost (Best F1 / Default)	0.34	0.64	0.45	0.93

The XGBoost model with SMOTE and default threshold delivered the best balance of performance, achieving 93% accuracy and an F1 score of 0.45. Threshold tuning enabled high recall (96%) where required.

5. Feature Importance Top financial ratios based on XGBoost importance scores included:



- **pcf**: Price-to-Cash Flow Ratio (used 427 times by the model)
    - Indicates how much investors are willing to pay for each dollar of operating cash flow, which can reflect perceived earnings quality and cash liquidity.
- **ocf_lct**: Operating Cash Flow / Current Liabilities (used 395 times)
    - Measures short-term liquidity, showing how well operating cash flow can cover short-term obligations.
- **pe_exi**: Price/Earnings excluding extraordinary items (used 374 times)
    - Focuses on normalized profitability, filtering out one-time anomalies.
- **pretret_earnat**, **short_debt**, **debt_assets**, **roa**, **GProf**, **curr_debt**, and others
    - These capture solvency, profitability, and leverage — key signals of financial distress that auditors and rating agencies consider when evaluating GC risk.

These values reflect how frequently each feature was used in decision trees within the XGBoost model. The more frequently a feature is used, the more important it is considered. This frequency-based measure ("F-score") helps identify which variables are most influential in predicting GC risk.

6. Interpretation & Practical Use

- **Precision (GC):** Likelihood that a flagged company is truly distressed
- **Recall (GC):** Coverage of actual GC cases identified
- **F1 Score (GC):** Balanced measure of precision and recall
- **Threshold Tuning:** Enables flexible trade-offs between catching more GC firms vs reducing false positives

- **Feature Usage:** XGBoost uses tree-based logic to repeatedly split the data. Features that help make more useful splits get ranked higher in importance.

7. Limitations

- **Limited Feature Scope:** Only financial ratios were used; qualitative factors, audit committee disclosures, and macroeconomic indicators were excluded, which may impact real-world GC judgments.
- **Data Source Confidentiality:** Due to privacy constraints, the dataset origin (WRDS/Audit Analytics) and firm identifiers were not verifiable, potentially limiting transparency.
- **Synthetic Sampling Risk:** SMOTE introduces synthetic observations which, while useful, may not fully capture the underlying distribution of truly distressed firms.
- **Imbalanced Label Bias:** Despite rebalancing, models may still overfit to the majority class without careful tuning and monitoring.
- **Feature Importance Limitation:** Frequency-based importance scores (F-score) don't tell us how impactful the feature was—just how often it was used. Alternative metrics like "gain" could provide deeper insight.
- **Threshold Tuning Tradeoff:** High recall scenarios (e.g. 96%) often come at the cost of reduced precision (17%), leading to a high number of false positives.