Equity Classification Using Financial Ratios

Winter in Data Science 2025

RISHAB JAIN

23B2527

1. Introduction

Equity classification is a fundamental problem in finance, where investors attempt to determine whether a company represents a good investment opportunity based on its financial performance. Traditionally, this process relies on manual analysis of financial statements and evaluation of financial ratios using predefined thresholds. While effective to some extent, such approaches are often subjective, time-consuming, and limited in their ability to capture complex relationships among financial variables.

With the availability of large-scale financial data and advances in machine learning, it is now possible to approach equity classification as a supervised learning problem. Machine learning models can learn patterns from historical financial data and evaluate multiple financial indicators simultaneously, allowing for more consistent and data-driven decisions.

This project focuses on building a machine learning–based system that classifies companies into **Good** or **Bad** investment categories using financial ratios derived from company financial statements.

---

2. Objectives

The objectives of this project are:

1. To automate the computation of financial ratios from raw financial data
2. To engineer ratio-based features suitable for machine learning models
3. To train and compare multiple machine learning algorithms for equity classification
4. To evaluate model performance using investment-relevant metrics
5. To analyze which financial ratios contribute most to investment quality predictions

---

3. Data Requirements

The dataset used in this project consists of structured financial data for multiple companies across different time periods.

3.1 Income Statement Data

- Revenue
- Cost of Goods Sold (COGS)
- Operating Expenses
- EBIT
- Net Income

3.2 Balance Sheet Data

- Current Assets
- Fixed Assets
- Total Assets
- Current Liabilities
- Long-Term Debt
- Total Liabilities
- Shareholders’ Equity

3.3 Cash Flow Statement Data

- Operating Cash Flow
- Capital Expenditure
- Free Cash Flow

3.4 Market Data

- Stock Price
- Earnings Per Share (EPS)

- Market Capitalization

3.5 Target Variable

- **Investment Quality**: Binary label indicating whether a company is classified as *Good* or *Bad* based on historical performance.

---

4. Evaluation Metrics

In equity classification, incorrect predictions have different financial consequences. Predicting a bad company as a good investment can result in capital loss, while predicting a good company as bad only leads to a missed opportunity.

To account for this, the following metrics were used:

- **Accuracy** – Overall correctness of predictions
- **Precision** – Reliability of “Good” investment predictions
- **Recall** – Ability to correctly identify good investments
- **F1-Score** – Balance between precision and recall
- **ROC-AUC** – Ability of the model to distinguish between good and bad investments

Greater emphasis was placed on **precision**, as minimizing false positives is critical in investment decisions.

---

5. Data Preprocessing

5.1 Handling Missing Values

Financial datasets often contain missing values due to incomplete reporting. Median imputation was applied to numerical features, as it is robust to extreme values and preserves the central tendency of financial distributions.

5.2 Outlier Treatment

Financial ratios can sometimes take extreme values that distort model learning. To mitigate this, ratio values were capped at the 1st and 99th percentiles, reducing the influence of outliers while preserving overall trends.

5.3 Safe Division for Financial Ratios

Many financial ratios involve division operations that may result in division by zero or infinite values. To ensure numerical stability, a safe division function was implemented.

```
    result = numerator / denominator

    result = result.replace([np.inf, -np.inf], np.nan)

    return result.fillna(default)
```

This function ensures reliable ratio computation across all companies.

---

6. Machine Learning Models

6.1 Feature Engineering Using Financial Ratios

More than 30 financial ratios were computed across profitability, liquidity, leverage, efficiency, valuation, cash flow, and growth categories. These ratios normalize financial data and enable fair comparison across companies.

```
ratios = pd.DataFrame()


# Profitability ratios

ratios['Net_Profit_Margin'] = safe_divide(

    df['Net_Income'], df['Revenue']

) * 100
```

```
ratios['ROA'] = safe_divide(
    df['Net_Income'], df['Total_Assets']
) * 100

ratios['ROE'] = safe_divide(
    df['Net_Income'], df['Shareholders_Equity']
) * 100

# Liquidity and leverage ratios
ratios['Current_Ratio'] = safe_divide(
    df['Current_Assets'], df['Current_Liabilities']
)

ratios['Debt_to_Equity'] = safe_divide(
    df['Total_Liabilities'], df['Shareholders_Equity']
)
```

Highly correlated ratios were removed using correlation analysis, resulting in a final feature set of approximately 20–25 ratios.

6.2 Models Implemented

The following models were trained and evaluated:

1. Conventional rule-based baseline

2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Gradient Boosting
6. Support Vector Machine (SVM)
7. Neural Network (MLP)

Feature scaling was applied where required, and stratified train-test splits ensured balanced class distribution.

6.3 Model Training and Evaluation Pipeline

All models were trained and evaluated using a common pipeline to ensure fair comparison.

```
results = []


for name, model in models.items():
    model.fit(X_train_scaled, y_train)


    y_pred = model.predict(X_test_scaled)
    y_proba = model.predict_proba(X_test_scaled)[:, 1]


    results.append({
        'Model': name,
        'Accuracy': accuracy_score(y_test, y_pred),
        'Precision': precision_score(y_test, y_pred),
```

```
        'F1-Score': f1_score(y_test, y_pred),
        'ROC-AUC': roc_auc_score(y_test, y_proba)
    })
```

---

7. Results and Analysis

7.1 Model Performance

Typical performance ranges observed:

Model	Accuracy
Gradient Boosting	87–92%
Random Forest	85–90%
Neural Network	83–89%
SVM	80–87%
Logistic Regression	76–84%
Rule-Based Baseline	65–75%

Machine learning models consistently outperformed the traditional approach by a margin of **15–25%**.

7.2 Feature Importance Analysis

Feature importance analysis from tree-based models revealed that the most influential ratios were:

1. Return on Equity (ROE)
2. Net Profit Margin
3. Operating Cash Flow Margin
4. Debt-to-Equity Ratio
5. Current Ratio

This highlights the importance of **profitability and cash flow strength** in determining investment quality.

7.3 Error Analysis

False positive predictions were found to be the most financially damaging. Ensemble models such as Random Forest and Gradient Boosting achieved lower false positive rates, making them more suitable for real-world investment screening.

---

8. Conclusion

This project demonstrates that machine learning techniques can significantly improve equity classification using financial ratios. By learning complex relationships across multiple financial indicators, ML models outperform traditional rule-based investing approaches in accuracy and reliability.

While the system does not replace comprehensive fundamental analysis, it serves as an effective screening tool that enhances efficiency and decision consistency. Ensemble methods emerged as the most reliable models for this task.

Future improvements may include sector-specific models, time-series features, explainable AI techniques, and portfolio-level evaluation.