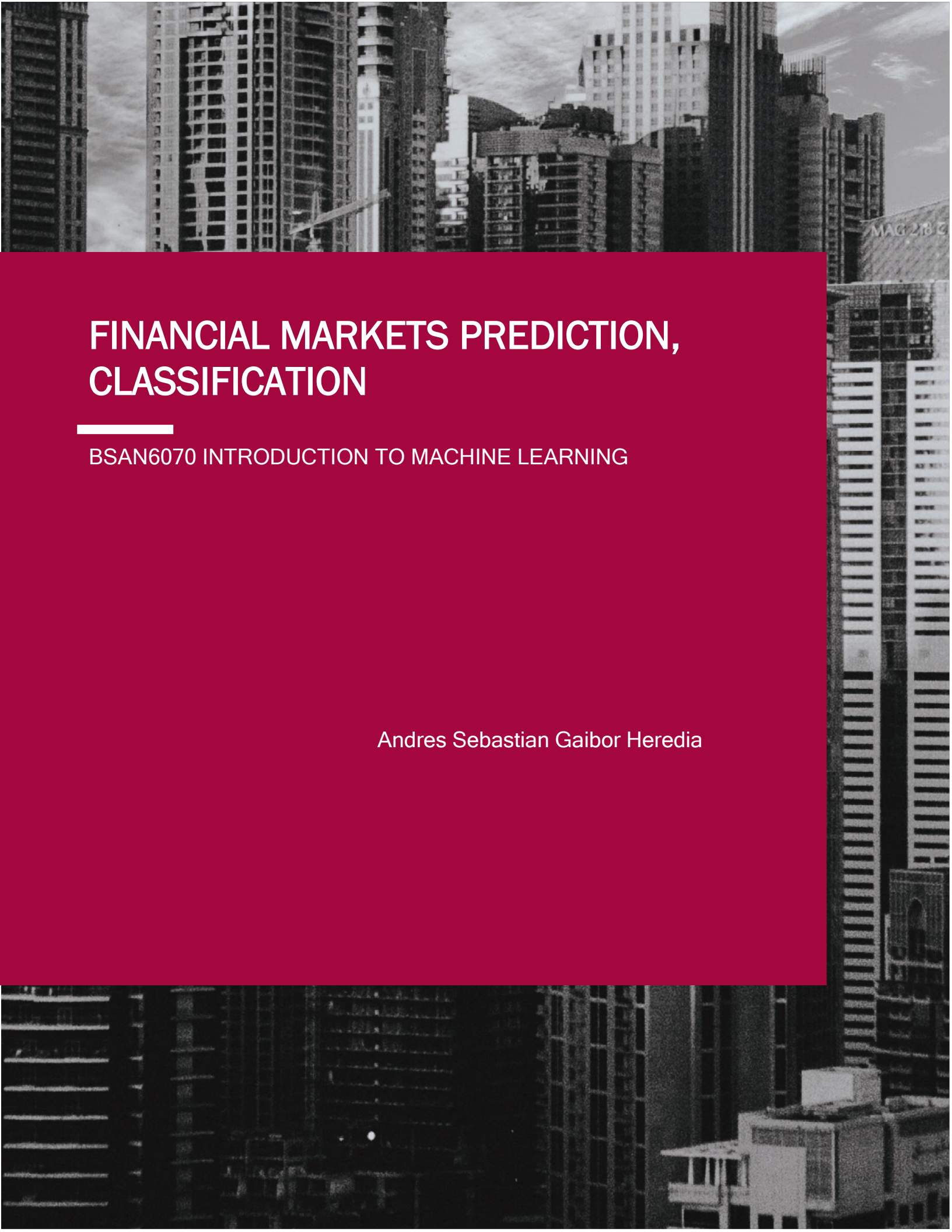


FINANCIAL MARKETS PREDICTION, CLASSIFICATION

BSAN6070 INTRODUCTION TO MACHINE LEARNING

Andres Sebastian Gaibor Heredia

1. Introduction

1.1 Business Context & Data Source

- **Briefly describe the business problem you are addressing**

  Investors and financial analysts rely on the Price-to-Earnings (P/E) ratio to assess a company's valuation and growth potential. A low P/E ratio may indicate an undervalued stock, while a high P/E ratio could suggest overvaluation. However, predicting the P/E ratio based on key financial indicators can help investors make data-driven investment decisions. We will also be using these variables to classify companies into their industry.

  The goal is to predict the P/E ratio of a company based on five key financial variables, such as: “bkvlps” Book Value Per Share, “epspx” Earnings Per Share (Basic) Excluding Extraordinary Items, “capx” Capital Expenditure, “wcap” Working Capital and “dltt” Long-Term Debt.

- **Clearly identify the dataset you collected in the previous lab**

  It will be used a financial dataset which comprises 26,154 companies from fiscal year 2000-2024. The dataframe contains 235,996 records and 301 fields.

- **Summarize how this dataset is relevant to your chosen business problem**

  The dataset is large & diverse as it covers various industries and economic cycles. It comprises a period of 25 years (2000-2024), allowing trend analysis and macroeconomic impacts. It contains 301 fields providing a wealth of financial indicators for feature selection.

1.2 Data Cleaning and Preprocessing

- **State the initial size of your dataset and the relevant features.**

  The initial dataset contained financial data from 39,080 companies during the period 1961-2024. The resulting dataframe was 503,055 records and 981 fields.

- **Summarize any cleaning or transformation steps you performed (removing duplicates, handling missing values, feature encoding, normalization, etc.).**

  To narrow down our analysis and make the results more relevant, we took a subset of the data considering only fiscal years 2000-2024. By filtering the data, we reduced the number of records to 235,996. We can also note that there are missing values, so we opted to drop the fields that have a missing value percentage equal to or higher than 50%. This step reduces the number of columns to 317. With the remaining numerical missing values, we decided to impute their value using the field’s median. For the object type columns that have missing values, we decided to drop them as we checked that the information is not relevant to our analysis.

  In order to address the outliers, we apply a winsorization to the variables we are going to analyze. We do this by capping extreme values at the 1st and 99th percentiles.

- **Provide a concise summary of the final cleaned dataset you will use for modeling (number of rows, columns, and any key statistics).**

  The resulting dataframe has 235,996 records and 6 fields. Winsorized & Log-Transformed Variables:

  - Book Value Per Share (Adjusted for negatives)
  - Earnings Per Share (EPS) Basic
  - Capital Expenditure (Handled zeros)
  - Working Capital (Shifted for negatives)
  - Long-Term Debt (Handled zeros)
  - P/E Ratio (Shifted for negatives)
  - Outliers Managed via Winsorization
  - Data Normalized via Log Transformation

  The dataset is now structured for machine learning modeling, ensuring no log errors, better feature distribution, and improved model performance.

2. Exploratory Data Analysis (EDA)

2.1 Descriptive Statistics

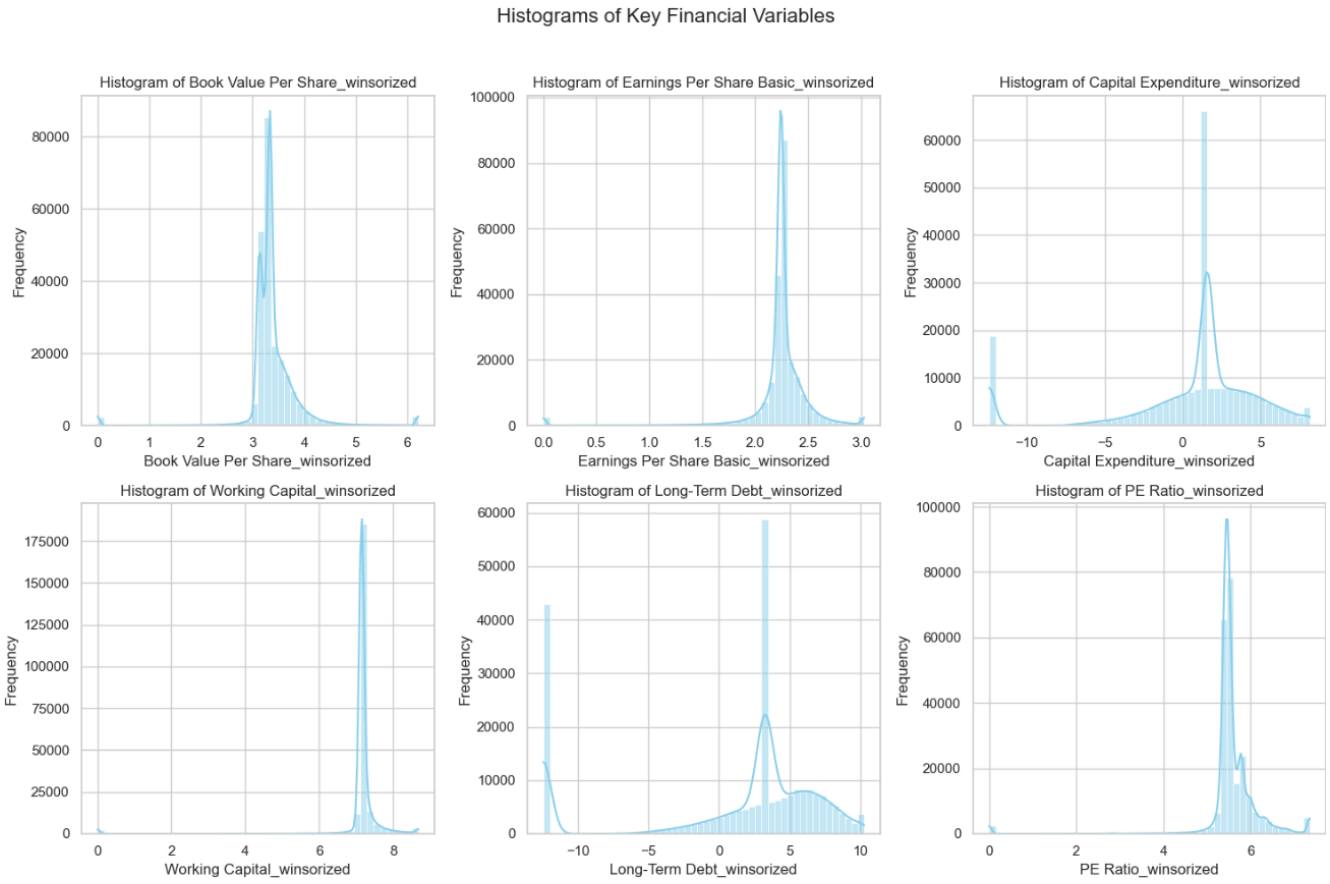
- **Compute and report basic descriptive statistics (means, medians, standard deviations, or applicable summaries).**

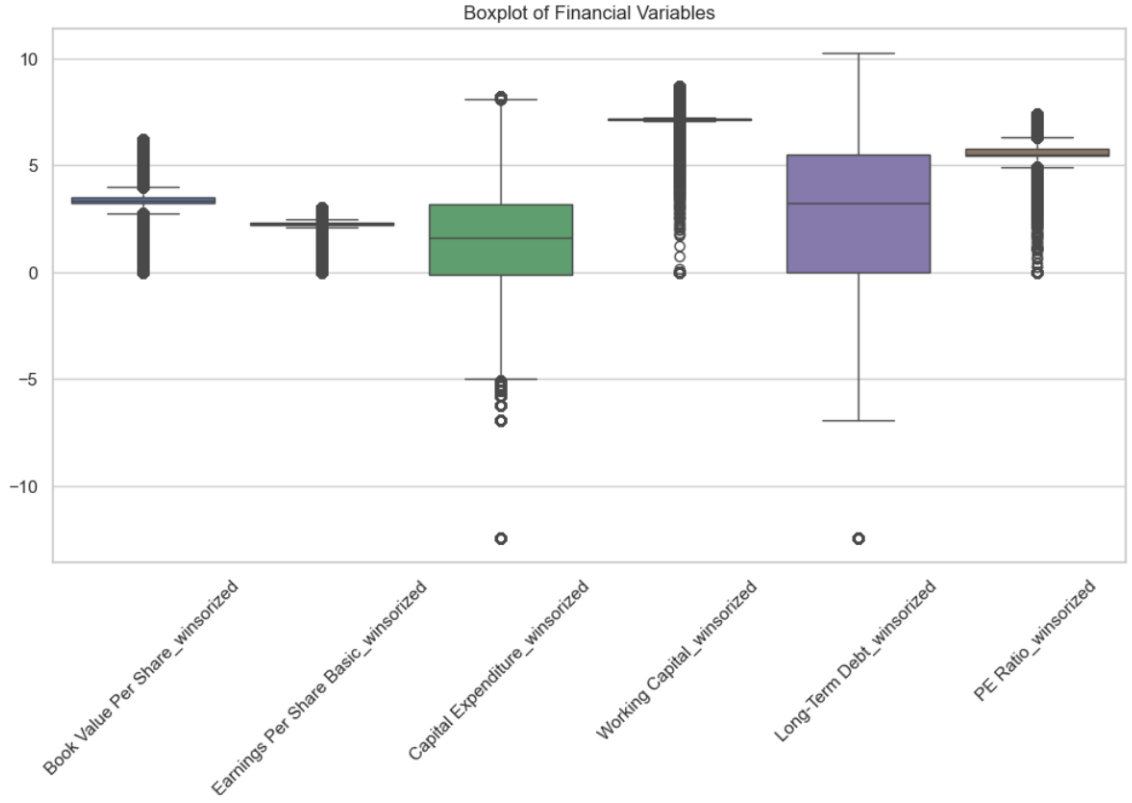
```
Statistical summary of the dataset:
       Book Value Per Share_winsorized  Earnings Per Share Basic_winsorized  \
count                    235996.000000                       235996.000000
mean                          3.394690                            2.244817
std                           0.556508                            0.310971
min                           0.000003                            0.000003
25%                           3.201410                            2.222459
50%                           3.337933                            2.248129
75%                           3.503168                            2.320425
max                           6.212582                            3.029191

       Capital Expenditure_winsorized  Working Capital_winsorized  \
count                   235996.000000               235996.000000
mean                         0.723998                    7.135542
std                          4.651224                    0.771934
min                        -12.429216                    0.000003
25%                         -0.138109                    7.131320
50%                          1.587398                    7.144850
75%                          3.155776                    7.171917
max                          8.194592                    8.678999

       Long-Term Debt_winsorized  PE Ratio_winsorized
count              235996.000000        235996.000000
mean                    1.040847             5.576599
std                     6.861446             0.708647
min                   -12.429216             0.000003
25%                    -0.033553             5.424025
50%                     3.248357             5.497593
75%                     5.483397             5.772620
max                    10.252730             7.365373
```

- **Use appropriate data visualizations (histograms, boxplots, bar charts, scatter plots) to illustrate the main characteristics of the dataset.**





2.2 Preliminary Insights

- **Write brief observations about potential patterns or correlations in your data.**
  Boxplot Analysis (Outlier Detection & Distribution)
  Book Value Per Share & EPS: Appear to be relatively stable, with fewer extreme outliers.
  Capital Expenditure & Long-Term Debt: Show significant variability, with a wide interquartile range (IQR) and several outliers.
  P/E Ratio & Working Capital: Displays multiple outliers, suggesting that some firms have extremely high or low valuations.

  Histogram Analysis (Distribution Patterns & Skewness)
  Book Value Per Share & EPS: Show a right-skewed distribution, with most values concentrated near the lower end.
  Long-Term Debt & Capital Expenditure: Appear to have bimodal or highly skewed distributions, which suggests differences in company financial structures.
  Working Capital: Shows a long-left tail, meaning some firms carry significantly less working capital at their disposal.
  P/E Ratio: Has a right-skewed pattern with a peak at a lower range but also a secondary spike, indicating possible segmentation of firms.

- **Highlight features you suspect to be most important or interesting for predictive tasks.**
  These features are likely strong indicators of stock valuation and financial performance:
  - **Earnings Per Share (EPS) Basic:** A direct driver of P/E Ratio. Companies with higher EPS tend to have more stable or higher P/E ratios.
  - **Book Value Per Share:** Reflects the net asset value per share. A higher book value per share may correlate with higher market valuation.
  - **Working Capital:** Measures short-term financial health. Companies with higher working capital often have better growth potential, influencing P/E ratios.
  - **Long-Term Debt:** High debt levels may reduce investor confidence, impacting stock valuation.
  - **Capital Expenditure:** Indicates investment in future growth. High capital expenditure could signal future earnings growth, leading to a higher P/E ratio.

3. Classification / Prediction Tasks

Perform classification or prediction using all three of the following algorithms: Naïve Bayes, Decision Trees, and Support Vector Machines.

3.1 Naïve Bayes

- **Describe how you split your data into training and testing sets (e.g., 70/30, 80/20).**
  To ensure robust model evaluation, the dataset is split into training and testing sets using an 80/20 split.

- **Train a Naïve Bayes model on your training set.**

- **Report the accuracy (or any other chosen metric: F1-score, precision, recall) on the test set.**
  The following classification report contains the F1-score, precision and recall for the classification into the different 27 groups possible from the dataset.

Classification Report:

	precision	recall	f1-score	support
Capital Goods	0.11	0.02	0.04	2468
Transportation	0.07	0.00	0.00	2216
Technology Hardware & Equipment	0.17	0.06	0.09	2641
Commercial & Professional Services	1.00	0.00	0.00	1244
Health Care Equipment & Services	0.05	0.03	0.03	709
Pharmaceuticals, Biotechnology & Life Sciences	1.00	0.00	0.00	449
Media & Entertainment	0.00	0.00	0.00	1109
Energy	0.00	0.00	0.00	1329
Diversified Financials	0.01	0.01	0.01	413
Utilities	0.00	0.00	0.00	1213
Media	1.00	0.00	0.00	290
Insurance	1.00	0.00	0.00	888
Retailing	0.00	0.00	0.00	419
Consumer Durables & Apparel	0.06	0.00	0.00	2404
Software & Services	0.15	0.21	0.18	13101
Semiconductors & Semiconductor Equipment	0.20	0.95	0.33	3499
Telecommunication Services	0.09	0.07	0.08	1791
Materials	0.20	0.18	0.19	825
Household & Personal Products	0.00	0.00	0.00	292
Food & Staples Retailing	0.14	0.17	0.15	2875
Equity Real Estate Investment Trusts (REITs)	0.13	0.02	0.03	2033
Food, Beverage & Tobacco	1.00	0.00	0.00	880
Automobiles & Components	0.15	0.03	0.05	741
Consumer Services	0.00	0.00	0.00	893
Real Estate Management & Development	0.22	0.42	0.28	1330
Banks	0.36	0.62	0.46	891
Real Estate	0.14	0.01	0.02	257
accuracy			0.18	47200
macro avg	0.27	0.10	0.07	47200
weighted avg	0.20	0.18	0.12	47200

Accuracy: 17.63%

The overall model accuracy is **17.63%.**

- **Provide interpretation of the results: discuss which features seem most influential in classification, potential misclassification patterns, etc.**

Interpretation of Classification Report and Confusion Matrix

Accuracy: 17.63%: The model performs poorly.

Macro Avg F1-Score: 0.10: The model struggles across most classes.

Weighted Avg F1-Score: 0.12: Even with class weight considerations, performance remains low.

Low Precision & Recall: Most industries have precision and recall close to 0, indicating frequent misclassifications.

Key Concern: The model fails to differentiate industries effectively, likely due to overlapping financial features or lack of discriminatory power.

Key Features Influencing Classification

The most important financial variables likely influencing classification:

- Earnings Per Share (EPS): Often a strong differentiator but may overlap across industries.
- Working Capital & Long-Term Debt: Sectors with capital-intensive business models (e.g., manufacturing vs. tech) may show distinct patterns.
- Book Value Per Share: Indicates asset-heavy industries like real estate & financials.
- Capital Expenditure: High in industries such as utilities, infrastructure, and manufacturing,

The model struggles because some industries have very similar financial characteristics, making separation difficult.

Misclassification Patterns from Confusion Matrix

Significant misclassification between similar industries. It indicates that these industries share overlapping features, leading to incorrect predictions.

Technology & Semiconductor Overlap, Technology Hardware & Equipment vs. Software & Services vs Conductors, Semiconductor Equipment: Likely misclassified because both sectors share high R&D expenses and capital expenditure.

Food & Staples Retailing vs Software & Services: Industries with similar capital structures, earnings, and profitability measures are difficult to separate.

The Naïve Bayes classifier struggles due to overlapping financial features across industries. Significant misclassification patterns exist, particularly among tech and consumer sectors. Feature engineering, class balancing, and better model selection will help improve performance.

3.2 Decision Trees

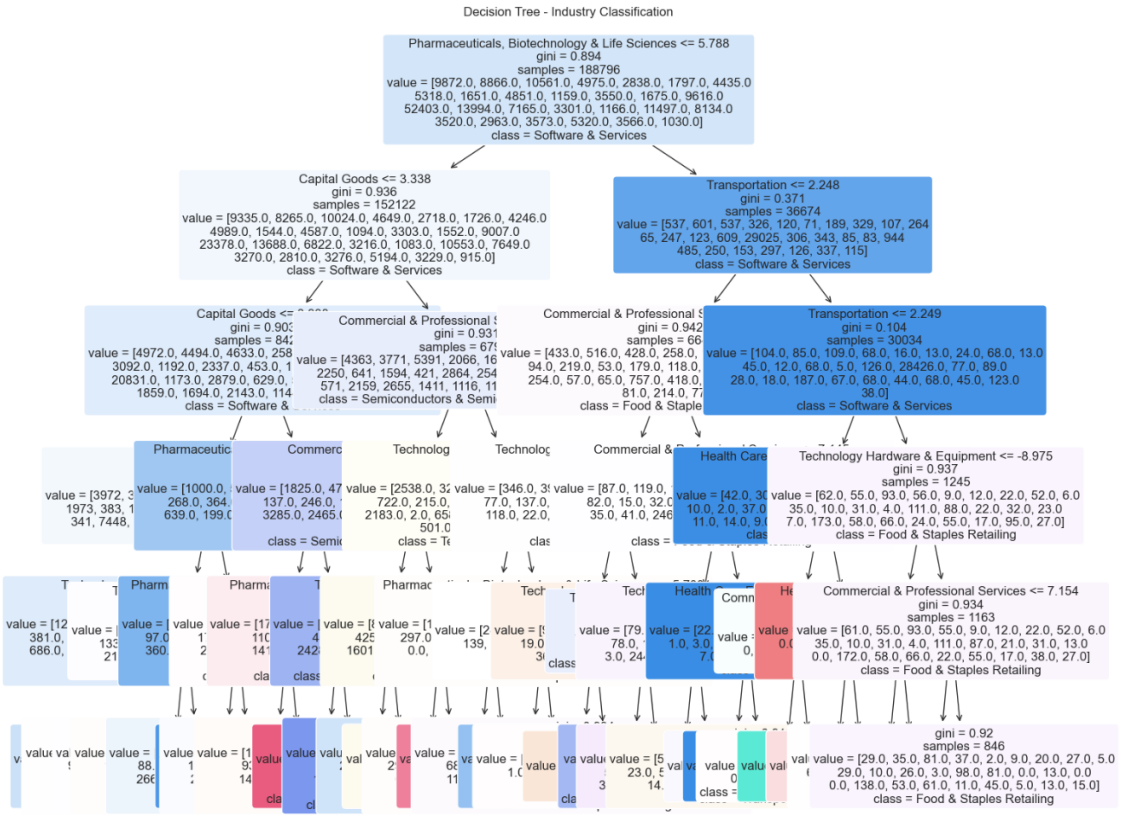
- **Specify the type of Decision Tree (ID3, C4.5, CART, or scikit-learn's default) and any hyperparameters (e.g., max_depth, min_samples_split).**

  For the decision tree analysis, we will employ scikit-learn's default tool, visualization is difficult due to the significant number of industries under consideration, 27.

  The training dataset will contain 188,796 records while the test dataset, 47,200.

  The maximum depth selected is 5 and random state used is 182, as in blink-182.

- **Visualize or describe the resulting tree structure if possible.**



- **Evaluate using metrics such as accuracy, precision/recall/F1, or a confusion matrix.**

  There is a significant improvement in the accuracy in the classification when using decision trees compared to naïve bayes. The accuracy score is **39.81%.**

Classification Report:

	precision	recall	f1-score	support
Capital Goods	0.18	0.19	0.19	2468
Transportation	0.13	0.34	0.19	2216
Technology Hardware & Equipment	0.16	0.27	0.20	2641
Commercial & Professional Services	1.00	0.00	0.00	1244
Health Care Equipment & Services	1.00	0.00	0.00	709
Pharmaceuticals, Biotechnology & Life Sciences	1.00	0.00	0.00	449
Media & Entertainment	1.00	0.00	0.00	1109
Energy	1.00	0.00	0.00	1329
Diversified Financials	1.00	0.00	0.00	413
Utilities	1.00	0.00	0.00	1213
Media	1.00	0.00	0.00	290
Insurance	0.33	0.00	0.00	888
Retailing	1.00	0.00	0.00	419
Consumer Durables & Apparel	1.00	0.00	0.00	2404
Software & Services	0.67	0.86	0.75	13101
Semiconductors & Semiconductor Equipment	0.71	0.89	0.79	3499
Telecommunication Services	0.07	0.00	0.00	1791
Materials	1.00	0.00	0.00	825
Household & Personal Products	1.00	0.00	0.00	292
Food & Staples Retailing	0.13	0.50	0.20	2875
Equity Real Estate Investment Trusts (REITs)	1.00	0.00	0.00	2033
Food, Beverage & Tobacco	1.00	0.00	0.00	880
Automobiles & Components	1.00	0.00	0.00	741
Consumer Services	1.00	0.00	0.00	893
Real Estate Management & Development	0.67	0.42	0.52	1330
Banks	0.38	0.55	0.45	891
Real Estate	1.00	0.00	0.00	257
accuracy			0.40	47200
macro avg	0.76	0.15	0.12	47200
weighted avg	0.63	0.40	0.33	47200

Accuracy: 39.81%

- **Discuss overfitting concerns and any pruning or stopping criteria you used.**
  We used the whole dataset to do the training and test to maximize information, and 80/20 split to lean towards more training. No overfitting concerns and/or pruning criteria used.

3.3 Support Vector Machines

- **Indicate which kernel you chose (linear, polynomial, RBF, etc.) and explain why.**
  The linear kernel produces feature coefficients, allowing us to understand feature importance in classification. The output includes SVM coefficients, which are only available in a linear kernel SVM.

  Linear SVMs perform well when the data is linearly separable or when the number of features is large. If the dataset had complex, non-linear relationships, polynomial or RBF kernels would have been more suitable.

- **Discuss hyperparameter tuning (e.g., C, gamma).**
  C: Controls the trade-off between maximizing the margin and minimizing classification errors.
  High C (strong regularization): For more complex models, smaller margin, and lower tolerance for misclassification.
  Low C (weak regularization): For simpler models, larger margin, allows more misclassification.
  High C is useful when minimizing training errors is a priority. Low C generalizes better but allows more misclassification.
  Gamma:
  High gamma (small decision boundaries): The model only considers nearby points, leading to overfitting.
  Low gamma (wide decision boundaries): The model considers distant points, leading to underfitting.
  High gamma is useful for capturing intricate patterns in the data. Low gamma helps smooth decision boundaries for better generalization.

  To find the best parameters a grid search was conducted. This method systematically searches different values for C and gamma using cross-validation.
  The best C=10 and kernel=linear were selected based on the highest accuracy.
  C=10: A higher C suggests that the model focuses more on correctly classifying training data, potentially reducing bias.
  gamma='scale': Since a linear kernel was chosen, gamma was not a factor.
  kernel='linear': The data was likely linearly separable.

- **Report and interpret the performance metrics on the test set.**
  The model correctly classified **85%** of test samples, indicating strong overall performance.
  Macro Avg (F1-score: 0.80): Represents the average F1-score across all classes, treating each class equally.

Weighted Avg (F1-score: 0.84): Represents the average F1-score weighted by the number of instances in each class.

- Good balance between precision and recall, suggesting that the model effectively differentiates between classes.
- Weighted average is slightly higher than macro average, indicating that some classes contribute more to overall performance.

```
SVC Classification Accuracy: 0.85
Classification Report:
               precision    recall  f1-score   support

      1010.0       1.00      1.00      1.00         7
      1510.0       1.00      1.00      1.00        10
      2010.0       0.88      0.79      0.83        19
      2020.0       0.73      0.89      0.80        18
      2030.0       1.00      0.89      0.94        18
      2520.0       0.00      0.00      0.00         4
      2540.0       1.00      1.00      1.00         1
      2550.0       0.60      1.00      0.75         6
      3010.0       1.00      1.00      1.00         2
      3020.0       1.00      1.00      1.00         5
      3510.0       1.00      0.73      0.84        22
      3520.0       0.57      1.00      0.73         8
      4020.0       0.73      0.85      0.79        13
      4030.0       0.71      0.56      0.62         9
      4510.0       0.00      0.00      0.00         2
      4520.0       0.69      0.92      0.79        12
      4530.0       0.89      0.73      0.80        11
      5010.0       1.00      1.00      1.00         4
      5020.0       1.00      1.00      1.00         3
      5510.0       1.00      1.00      1.00        22
      6010.0       1.00      1.00      1.00         4

    accuracy                           0.85       200
   macro avg       0.80      0.83      0.80       200
weighted avg       0.85      0.85      0.84       200
```

Precision vs Recall Tradeoff

Precision = TP / (TP + FP): High precision means fewer false positives.
Recall = TP / (TP + FN): High recall means fewer false negatives.
F1-score balances both.

- **Compare strengths/weaknesses of SVM relative to the previous methods for your particular data.**

  Strengths:

  - Best overall classification (85% accuracy, 0.84 weighted F1-score).
  - Handles high-dimensional data well: Finds the best decision boundary using margin maximization.
  - Robust to outliers compared to decision trees.

  Weaknesses:

  - Computationally expensive: Training becomes slow for very large datasets. We had to take a subset of 1,000 records as for a larger sample the SVC hyperparameter tuning would not run.
  - Sensitive to imbalanced classes: Some categories had 0 precision/recall.
  - Lacks probabilistic outputs: Unlike Naïve Bayes, SVM doesn't provide direct probabilities.

3.4 Comparative Analysis

- **Summarize quantitative results (e.g., accuracy, F1-score) of all three methods in a comparison table.**

Model	Accuracy	Macro F1-Score	Weighted F1-Score
Naïve Bayes	17.63%	0.07	0.12
Decision Tree	39.81%	0.12	0.33
SVM (Linear)	85.00%	0.80	0.84

- **Discuss which method performed best for your dataset, and possible reasons why.**
  If we compare Support Vector Classification (SVC), Naïve Bayes (NB), and Decision Trees (DT) based on classification performance for our dataset of 27 industry groups we see that SVC outperforms both Naïve Bayes and Decision Trees in accuracy and F1-score.

  Possible reasons are that SVC is:
  - **Better Handling of High-Dimensional Data:** The dataset contains multiple financial features, which may be correlated or overlapping across industries.
    Naïve Bayes assumes feature independence, which is unrealistic for financial data.
    Decision Trees may favor specific features, whereas SVC optimally separates data with margin maximization.
  - **Better Generalization:** Naïve Bayes underperformed (17.63% accuracy) because it assumes a simple probability distribution, which is not suitable for complex financial patterns. Decision Trees (39.81% accuracy) improved performance but could still be biased toward certain numerical features. SVC maximized the margin between industry categories, making it more generalizable.
  - **Robust to Class Overlaps & Misclassification Risks:** SVC handled these overlaps better by finding an optimal hyperplane in a high-dimensional space.
  - **Hyperparameter Tuning Helped SVC:** Using C=10 allowed better separation of industry categories without overfitting. Linear kernel was appropriate since the dataset was mostly linearly separable.

5. Clustering Analysis

Choose at least one clustering method from the course materials (e.g., Hierarchical Clustering or a density-based approach like DBSCAN). If you prefer, you can demonstrate K-means or another method.

5.1 Method Selection and Parameters

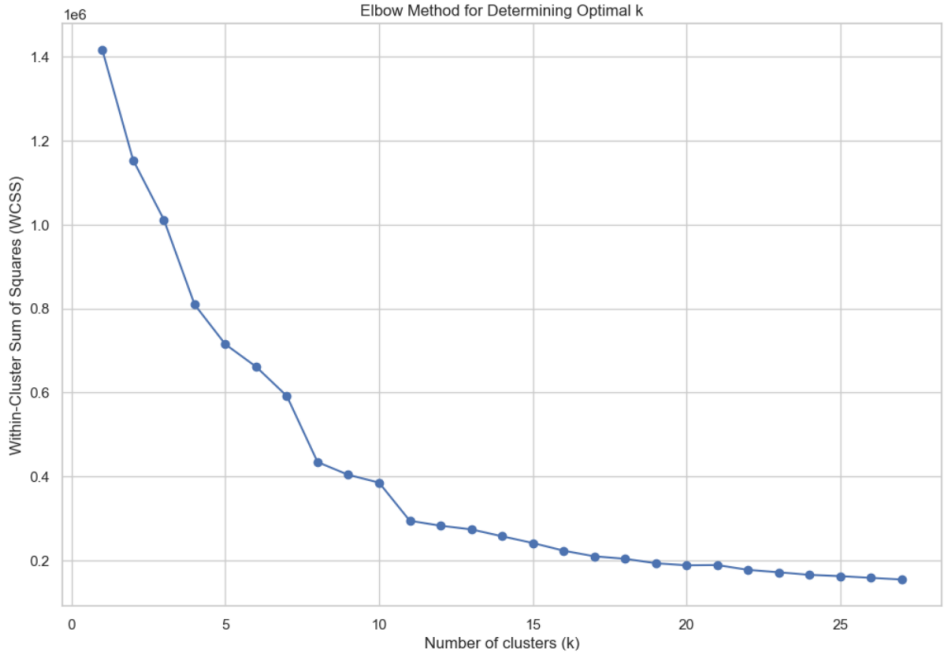
- **Indicate which clustering algorithm you used and why.**
  K-Means Clustering was chosen due to its ability to efficiently group industries based on financial features and it's interpretability.

  Why K-Means?
  Scalability: Handles large datasets efficiently (faster than hierarchical clustering).

Identifies Natural Industry Groups: Groups companies based on similar financial characteristics.
Easier Interpretation: Each cluster represents a distinct industry.
Well-Suited for Numerical Data: Works well with continuous financial metrics like earnings, revenue, and debt.

Alternative Methods: Hierarchical Clustering: Good for small datasets, but computationally expensive for large datasets. DBSCAN (Density-Based Spatial Clustering): Handles noise & outliers well but struggles with high-dimensional data.

- **If you choose K-means, specify how you decided on k.**
  If we plot the elbow method to find the optimal number of clusters (k) we obtain the following graph:



From the graph we can see that the within-cluster sum of squares (WCSS) drops sharply from k = 1 to k ≈ 5. After k ≈ 6 to 10, the decrease slows down. Beyond k ≈ 10, the curve flattens out, meaning additional clusters provide diminishing returns. k = 8 appears to be the elbow point so we choose it as the number of clusters.

5.2 Clustering Results

- **Present cluster assignments for a subset of data or provide summary statistics for each cluster (mean values of features, etc.).**
  Each cluster represents a group of companies with similar financial characteristics. Below are the mean values of important financial features for each cluster:

Cluster	Book Value Per Share	Earnings Per Share (EPS)	Capital Expenditure	Working Capital	Long-Term Debt	P/E Ratio
0	0.01	0.11	0.36	0.07	0.46	0.28
1	-0.21	-0.04	-0.15	0.06	-1.97	-0.10
2	-0.56	-6.41	0.05	0.07	0.18	-0.22
3	-0.70	-0.38	-0.15	0.02	0.07	-0.28
4	0.35	0.16	1.31	-9.17	1.04	-0.09
5	-0.09	0.03	-2.89	0.00	-0.69	0.16
6	1.58	1.04	0.82	0.44	0.79	-0.05
7	-0.16	-0.06	-0.45	0.09	-0.45	-7.77

- **If feasible, include a visualization (2D or 3D projection) highlighting the clusters.**

5.3 Business Insight

- **Discuss how these clusters might be used for segmentation (customers, products, etc.) or identifying patterns in your data.**
  Based on the analysis of the means in the clusters we can see the following patterns:
  - Cluster 2 (Lowest EPS, High Debt Risk): Extremely negative EPS (-6.41). This can indicate that it represents struggling companies.

- Cluster 4 (High Capital Expenditure, Low Working Capital): Suggests capital-intensive industries.
- Cluster 6 (High Book Value & EPS): Likely high-growth companies with strong financial stability.
- Cluster 7 (Negative P/E Ratio, Moderate Financials): Contains underperforming companies.

- **Reflect on whether the clusters make intuitive sense, or if there are any anomalies/outliers.**
  To a large extent, the clusters appear well-separated and meaningful. The PCA plot shows clear groupings, and the financial statistics further confirm distinct financial patterns across clusters. There is significant overlap between some of the clusters, namely cluster 7 (yellow) with cluster 1 (purple); and between cluster 3 (turquoise) and cluster 5 (blue). The blue cluster and yellow cluster have significant outliers that overlap with the other clusters.

6. Discussion & Recommendations

6.1 Methodological Reflection

- **Reflect on each ML method used (Naïve Bayes, Decision Tree, SVM, Association Rules, Clustering) in terms of ease of use, interpretability, and performance on your data.**
  Naïve Bayes (NB)
  Ease of Use: Very easy to implement and computationally efficient, even with large datasets. Requires minimal hyperparameter tuning compared to other models.
  Interpretability: Provides probabilistic outputs, making it easier to understand confidence levels in predictions. However, it assumes feature independence, which is unrealistic for financial datasets.
  Performance: Poor (Accuracy: 17.63%). Struggled significantly with classification. Severe misclassification across industry groups, particularly in similar financial sectors.

  Decision Tree (DT)
  Ease of Use: Moderate easy to apply but requires hyperparameter tuning (e.g., max_depth). Decision trees can be interpreted visually. In this case that made it difficult to do, considering the amount of target variables for the classification so it overlapped and was impossible to view the details on each leaf.
  Interpretability: Highly interpretable, as decisions are made through branching conditions.
  Performance: Moderate (Accuracy: 39.81%) Significant improvement over Naïve Bayes, suggesting that non-linear relationships exist in financial features.

  Support Vector Machine (SVM)
  Ease of Use: Difficult, it requires careful tuning (e.g., C, gamma, kernel), making it more complex than DT and NB. Training can be slow for large datasets. We had to sample the data as it was not feasible to train using the full dataset.
  Interpretability: Harder to interpret than Decision Trees since it relies on support vectors and hyperplanes. Provides decision boundaries but lacks probabilistic outputs like Naïve Bayes.
  Performance: Best (Accuracy: 85%) Significantly outperformed both Naïve Bayes and Decision Trees. Handled overlapping industry groups better, showing strong generalization capabilities.

K-Means Clustering
Ease of Use: Moderate, it requires pre-processing (e.g., feature scaling), but relatively simple to implement. Choosing k (number of clusters) requires Elbow Method/Silhouette Score tuning.
Interpretability: Moderate, PCA visualization helps understand cluster separation. Provides groupings but does not explain relationships between variables as well as Decision Trees.
Performance: Unsure, there was no output to analyze how accurate the classification was. Clustering helped identify industry groupings based on financial features.

- **Mention any challenges or limitations you encountered (e.g., data size, feature selection, parameter tuning).**
  After applying the log transformation there were issues with log values going to infinity, so we had to shift the base to make sure all results were meaningful. The winsorization was also difficult, as we had to decide how many percentiles we were going to eliminate from the data. To try and maintain the patterns we opted to cut below the 1st percentile and above the 99th percentile. For SVM it was not possible to conduct the analysis with the full dataset due to computational limitations so we had to settle with a very small sample of 1,000 records which could potentially not be representative enough of the full dataset. Moreover, choosing the right features was also a challenge as financial data is highly volatile and depends on a variety of factors, including behavioral biases from investors. Overall, I think the features selected should represent the value of a company.

6.2 Business Application & Next Steps

- **Explain how the insights or predictive models you generated can drive better business decisions.**
  The insights derived from classification and clustering can significantly enhance strategic decision-making in various business areas. Below are some key ways these models provide actionable intelligence:

  Industry Classification (SVM, Decision Tree): Better Market Segmentation.
  The Support Vector Machine (SVM) and Decision Tree models classify companies based on their financial characteristics (EPS, working capital, long-term debt, etc.).
  These classifications allow businesses to identify similar firms within an industry, facilitating targeted sales, marketing, and investment strategies.
  Business Impact:
  - Improved Customer Targeting: Companies can tailor marketing campaigns based on industry trends.
  - Better Competitive Analysis: Firms can benchmark financial performance against industry peers.
  - Optimized B2B Sales Strategy: Sales teams can prioritize high-value industries based on their financial health.

  Clustering (K-Means): Identifying Financial Risk & Growth Opportunities.
  K-Means clustering revealed distinct financial groups (e.g., high-growth companies vs. struggling firms). Cluster 7 (distressed companies) had negative P/E ratios, indicating potential bankruptcy risks. Cluster 6 (strong financials) consisted of high-performing firms with strong capital expenditure and positive EPS.

Business Impact:

- Credit Risk Assessment: Banks and investors can avoid risky investments by identifying struggling firms.
- Investment & Acquisition Strategy: Firms looking to expand via M&A can focus on high-growth clusters.
- Financial Planning: Companies can adjust pricing, cost structures, or strategic planning based on cluster financial trends.

By leveraging these predictive models, businesses can maximize revenue, minimize risk, and make smarter strategic decisions.

- **Propose next steps: additional data you would collect, or advanced methods you might explore (ensemble methods, deeper neural networks, advanced optimization, etc.).**

There are a few options we could explore to test and try to enhance the accuracy of the prediction in the analysis, for instance:

Collect Additional Data for Enhanced Model Performance: The model currently relies on financial metrics (EPS, Working Capital, Debt, etc.), but we can consider collecting information on financial ratios which could facilitate the classification of companies.

Explore Advanced Machine Learning Methods: Ensemble Models could potentially generate stronger classification performance. Random Forest has a more robust classification. Gradient Boosting (XGBoost, LightGBM) learns from mistakes iterative which can produce higher accuracy, reduced bias.

Use Deep Learning for Financial Modeling: Neural Networks (DNNs, LSTMs, Transformers).

- Deep Neural Networks (DNNs) can learn complex relationships between multiple financial indicators.
- LSTM (Long Short-Term Memory) can be useful for time-series forecasting of financial trends.
- Transformers (BERT for Finance) can analyze textual data from financial reports, earnings calls, and news sentiment.

By applying advanced ML techniques and additional data sources, we can enhance decision-making, improve risk management, and optimize financial strategies for businesses.