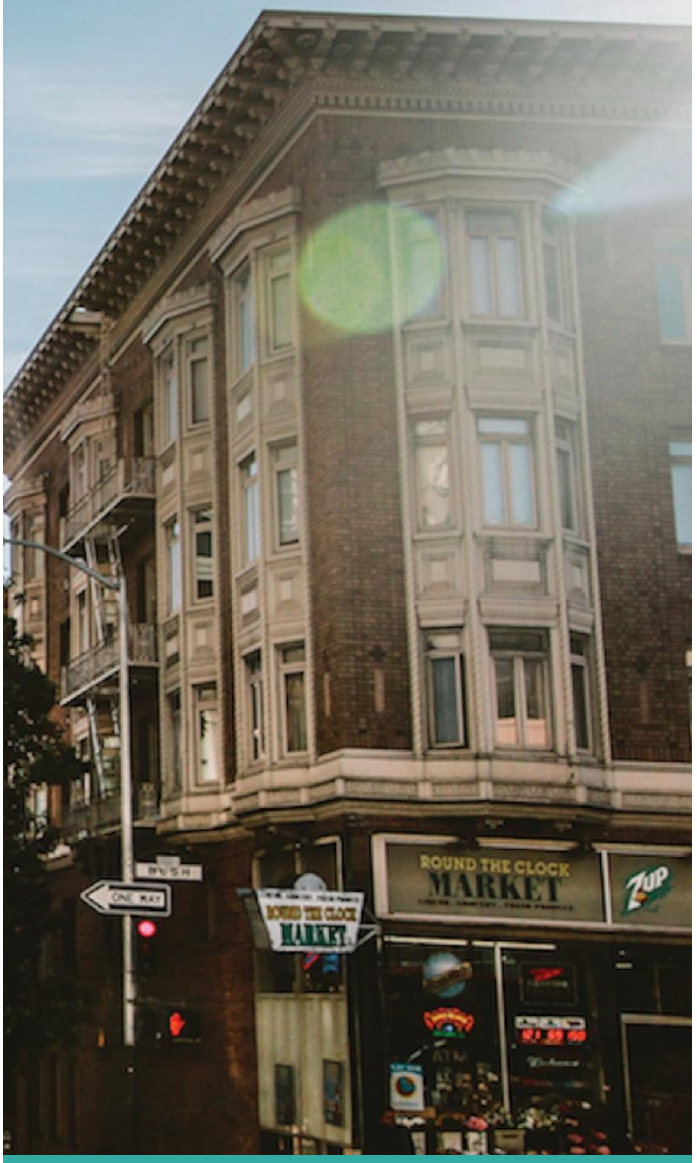
LAB 5: DECISION TREE



OCTOBER 22

BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. Decision Tree

We will continue to use our financial dataset which comprises 26,154 companies from fiscal year 2000-2024. We are dropping the categorical variables that have missing values as we reviewed, they are not crucial for the analysis. The resulting dataframe contains 235,996 records and 301 fields. The target variable will be “ggroup” which is the Global Industry Classification Group for the companies. The explanatory variables will be “at” Total Assets, “ni” Net Income, “revt” Total Revenue, “ceq” Total Common/Ordinary Equity, “epspx” Earnings Per Share (Basic) Excluding Extraordinary Items, “capx” Capital Expenditure, “oibdp” Operating Income Before Depreciation, “wcap” Working Capital, “dltt” Long-Term Debt and “xsga” Selling General & Administrative Expenses.
We define X as a matrix with the chosen explanatory variables and y as the field “ggroup”. Then we map the respective industry names and add them into our dt_data dataframe with the rest of the explanatory variables. The resulting dataframe is shown below.

```
First five rows of the dataset:
        at      ni     revt      ceq  epspx    capx   oibdp     wcap     dltt  \
0  701.854  18.531  874.255  340.212   0.69  13.134  64.367  360.464  179.987
1  710.199 -58.939  638.721  310.235  -2.08  12.112  27.207  286.192  217.699
2  686.621 -12.410  606.337  294.988  -0.39   9.930  30.745  192.837  164.658
3  709.292   3.504  651.958  301.684   0.11  10.286  47.491  300.943  248.666
4  732.230  15.453  747.848  314.744   0.58  13.033  61.774  314.517  227.159

     xsga       Industry
0  96.077  Capital Goods
1  85.037  Capital Goods
2  78.845  Capital Goods
3  81.165  Capital Goods
4  87.902  Capital Goods
```

After defining dt_data we split the X and y variables into training and test data using a 70/30 split, using a random state = 42 and stratification. We employ the 'gini' criterion for splitting, gini impurity is used to measure how often a randomly chosen element from the set would be incorrectly labeled. The max_depth=5 for the decision tree, the depth of the decision tree prevents overfitting. A deeper tree captures more complexity but may lead to overfitting, while a shallower tree generalizes better.

Next, we train the data using the Decision Tree Classifier and use it to make predictions on the test data. The resulting confusion matrix has a shape of (27, 27). The classification report provided evaluates the performance of the decision tree classifier on a multi-class classification task. The key metrics analyzed are precision, recall, f1-score, and support across multiple industry categories. The resulting classification report is shown below.

Classification Report:

	precision	recall	f1-score	support
Capital Goods	0.38	0.14	0.20	3702
Transportation	0.15	0.24	0.19	3325
Technology Hardware & Equipment	0.16	0.26	0.20	3961
Commercial & Professional Services	1.00	0.00	0.00	1866
Health Care Equipment & Services	1.00	0.00	0.00	1064
Pharmaceuticals, Biotechnology & Life Sciences	0.00	0.00	0.00	674
Media & Entertainment	1.00	0.00	0.00	1663
Energy	0.15	0.20	0.17	1994
Diversified Financials	1.00	0.00	0.00	619
Utilities	1.00	0.00	0.00	1819
Media	1.00	0.00	0.00	435
Insurance	1.00	0.00	0.00	1331
Retailing	1.00	0.00	0.00	628
Consumer Durables & Apparel	1.00	0.00	0.00	3606
Software & Services	0.71	0.90	0.80	19651
Semiconductors & Semiconductor Equipment	0.43	0.94	0.59	5248
Telecommunication Services	0.36	0.32	0.34	2687
Materials	1.00	0.00	0.00	1238
Household & Personal Products	1.00	0.00	0.00	437
Food & Staples Retailing	0.15	0.43	0.22	4312
Equity Real Estate Investment Trusts (REITs)	1.00	0.00	0.00	3050
Food, Beverage & Tobacco	1.00	0.00	0.00	1320
Automobiles & Components	1.00	0.00	0.00	1111
Consumer Services	1.00	0.00	0.00	1340
Real Estate Management & Development	0.51	0.77	0.62	1995
Banks	0.35	0.27	0.30	1337
Real Estate	1.00	0.00	0.00	386
accuracy			0.42	70799
macro avg	0.72	0.17	0.13	70799
weighted avg	0.62	0.42	0.35	70799

Accuracy: 42.32%

Overall Model Performance

The classification report provides insight into the performance of a Decision Tree Classifier used to predict the Global Industry Classification Group (ggroup) based on 10 financial indicators. The model's overall accuracy is 42.32%, which is higher than Naïve Bayes, indicating moderate predictive power, but the per-category performance varies significantly.

Key Observations

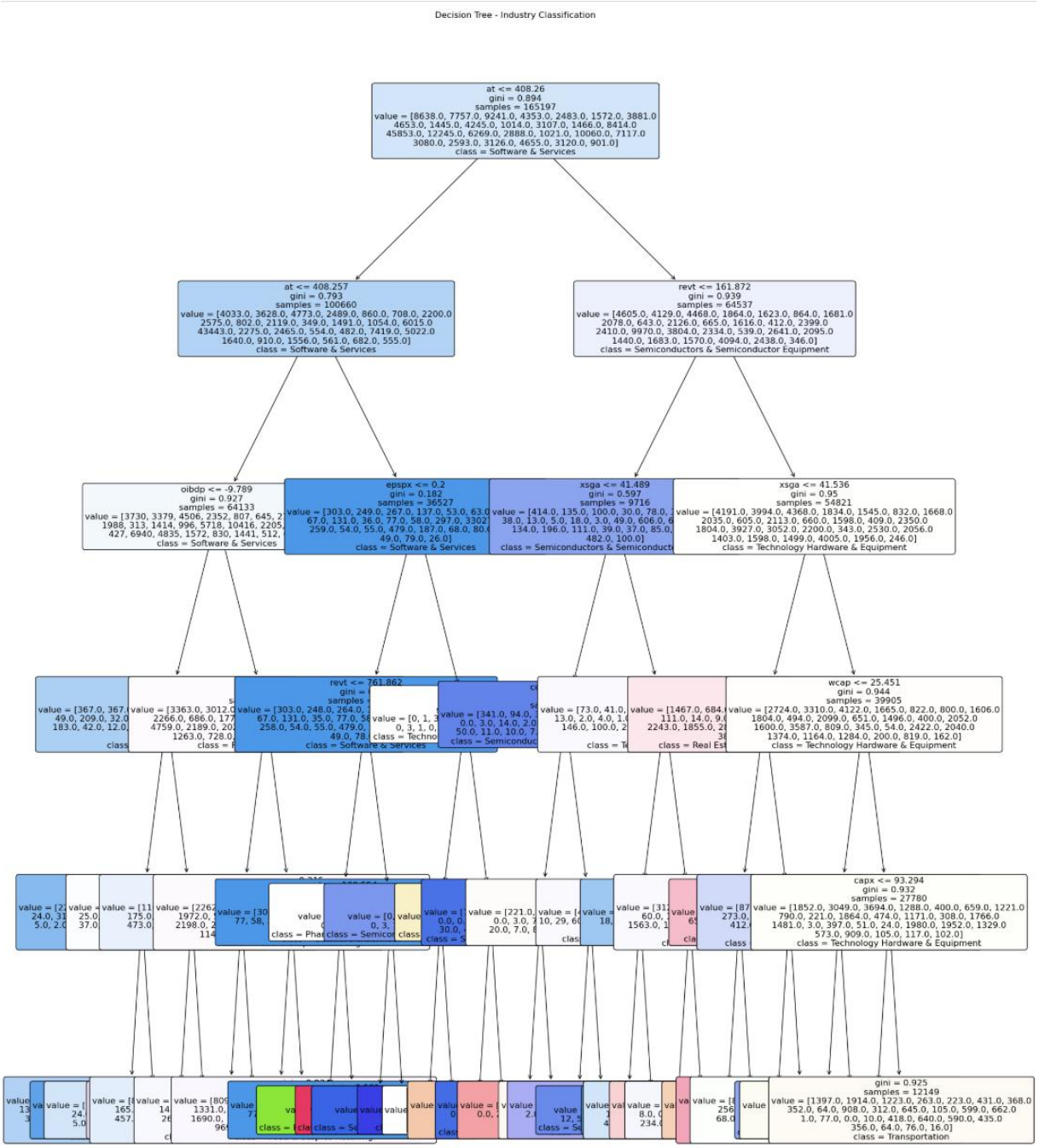
Disproportionate Class Performance: Some industry groups such as "Semiconductors & Semiconductor Equipment" and "Software & Services" have relatively strong recall scores (0.94 and 0.90, respectively), indicating that the model is good at identifying these categories. Categories like "Semiconductors & Semiconductor Equipment" (0.43 precision, 0.94 recall) indicate that the model frequently predicts this class but does so with lower precision.

However, many categories have recall scores close to 0, meaning that the model struggles to classify them correctly.

Overfitting to Certain Categories: The macro-average recall (0.17) is significantly lower than the macro-average precision (0.72), indicating that the model might be biased towards a few dominant classes while failing to generalize across all categories.

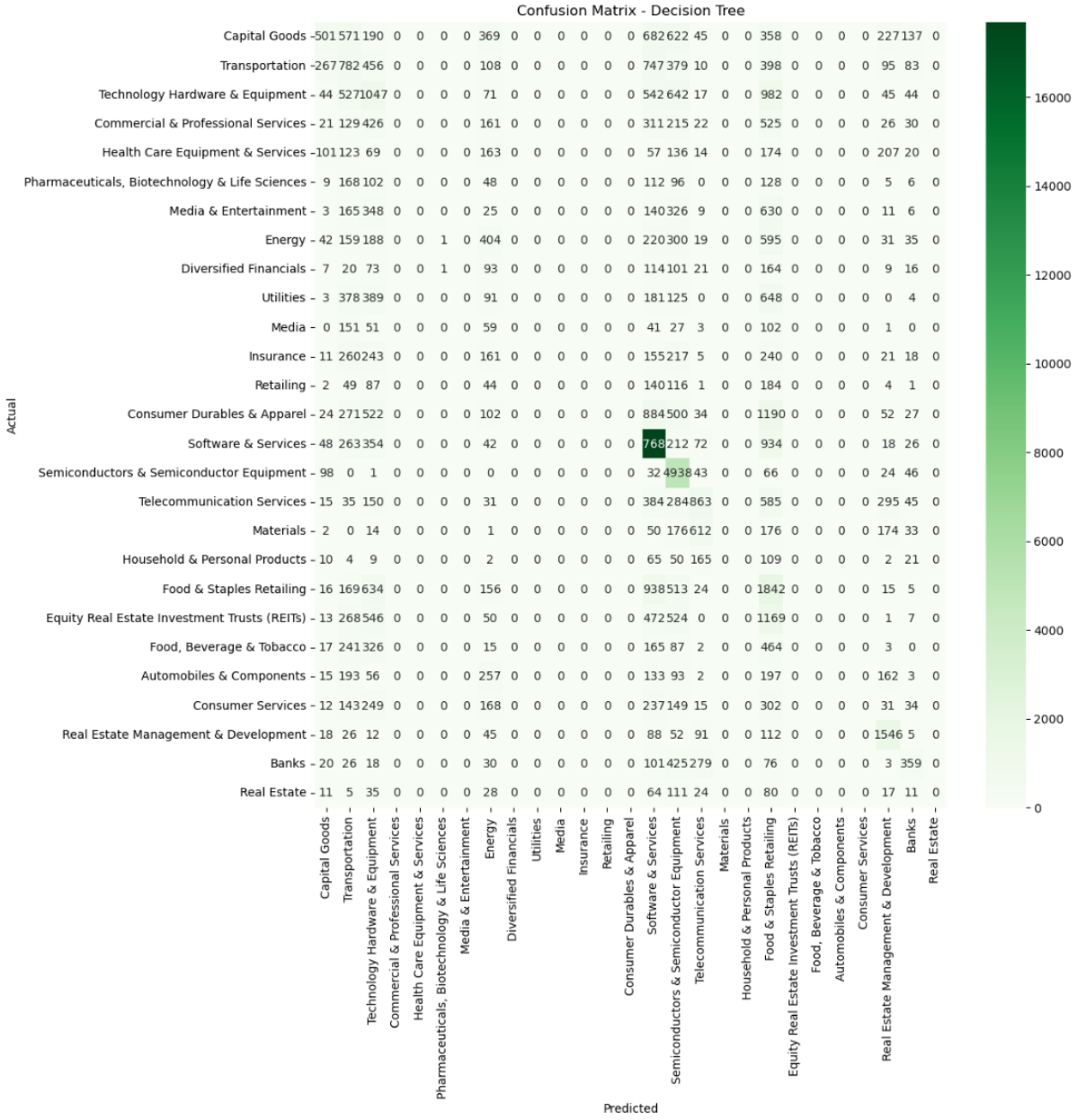
Potential Feature Influence: The model struggles with categories where financial metrics overlap (e.g., "Commercial & Professional Services" vs. "Consumer Services"), possibly due to insufficient differentiating features.

The resulting decision tree from the analysis is shown below,



Confusion Matrix

The confusion matrix provides insights into the performance of the Decision Tree Classifier in classifying companies into their Global Industry Classification Group (ggroup) based on financial features. The confusion matrix can be seen below.



Key Observations

Strong Performance in Some Categories: The classifier performs relatively well in recognizing certain industries, as seen by large diagonal values for:

- Software & Services (17,682 correct classifications)
- Semiconductors & Semiconductor Equipment (4,938 correct classifications)
- Food & Staples Retailing (1,842 correct classifications)

These industries likely have distinct financial characteristics that help differentiate them from others.

Significant Misclassifications: Several industries suffer from high misclassification rates, particularly: Capital Goods, Transportation, and Energy have many misclassified instances. Diversified Financials, Media & Entertainment, and Retailing have very few correct classifications, meaning the model struggles to distinguish them. Some industries like Pharmaceuticals, Biotechnology & Life Sciences, and Real Estate appear to be almost entirely misclassified.

Key Takeaways

- **Feature Engineering Improvements:** Consider normalizing financial variables (e.g., dividing by total assets or revenue) to capture relative financial performance rather than absolute values.
- **Introduce sector-specific financial ratios** (e.g., return on assets, operating margin) to help distinguish industries with similar financial structures.
- **Data Preprocessing:** Address class imbalance by applying sampling techniques (oversampling underrepresented classes or undersampling dominant ones).