LAB 5: SVM

OCTOBER 22

BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. SVM Regression

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. From there we select 5,000 records to be able to find the hyperparameters using GridSearchCV due to computing power limitations. We select the 10 explanatory variables: “at” Total Assets, “ni” Net Income, “revt” Total Revenue, “ceq” Contributed Equity, “epspx” Earnings per Share Excluding Extraordinary Items, “capx” Capital Expenditure, “oibdp” Operating Income Before Depreciation, “wcap” Working Capital, “dltt” Long-Term Debt, “xsga” Selling & General Administrative Expenses. We define X_reg as a matrix containing 5,000 records of the explanatory variables and y_reg as a dataframe containing the industry codes. We split the data into 80/20 training and test sets, using random state = 42. Then we standardized the features using StandardScaler() for better clustering performance. Then we find the hyperparameters using GridSearchCV where “c” = 0.1, “epsilon” = 0.1 and “kernel”: linear.

Evaluating the SVR Model

Mean Squared Error (MSE) = 1,933,950.19

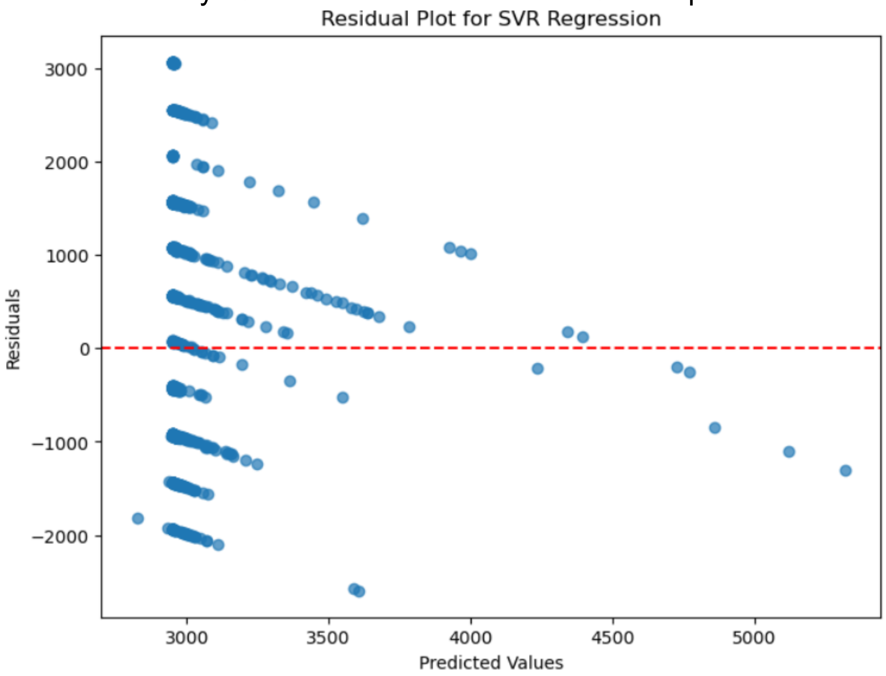
MSE measures the average squared difference between the predicted and actual values. A high MSE (like 1,933,950) suggests that the model's predictions are far from the actual values. This indicates that the SVR model is not performing well.

R^2 Score = -0.0034

R^2 (coefficient of determination) measures how well the model explains the variance in the target variable. Since R^2 is negative (-0.0034), this means SVR is performing worse than a baseline model that predicts the mean of the target variable for all instances.

Residual Plot

A residual plot visualizes the difference between the actual and predicted values. Ideally, residuals should be randomly scattered around zero with no clear pattern.



Residual Pattern Indicates a Problem: The residuals are not randomly distributed; instead, they form a structured pattern. There is an evident trend where residuals become more negative as predicted values increase, suggesting heteroscedasticity (variance of residuals changes with predictions).
This indicates that SVR is systematically underestimating or overestimating predictions, which is a sign of poor model fit.

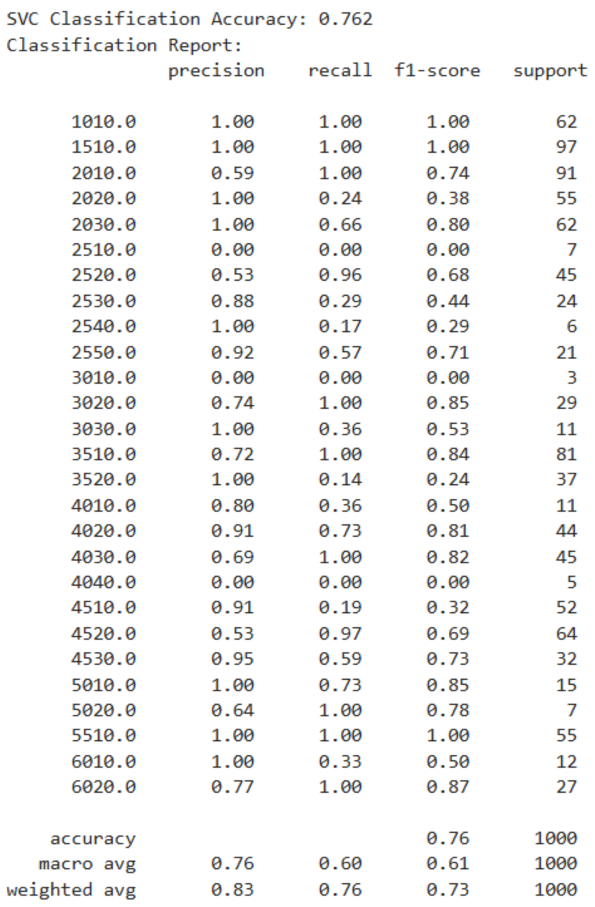
Residuals Are Not Centered Around Zero: In a well-fitted model, residuals should be evenly distributed above and below the zero line. Here, we see more residuals above zero for lower predicted values and more negative residuals for higher predicted values. This suggests that SVR struggles with capturing the relationship in the data.

2. SVC Classification

This time around we will use SVC to classify companies into their corresponding industries. 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. From there we select 5,000 records to be able to find the hyperparameters using GridSearchCV due to computing power limitations. We select the 10 explanatory variables: “at” Total Assets, “ni” Net Income, “revt” Total Revenue, “ceq” Contributed Equity, “epspx” Earnings per Share Excluding Extraordinary Items, “capx” Capital Expenditure, “oibdp” Operating Income Before Depreciation, “wcap” Working Capital, “dltt” Long-Term Debt, “xsga” Selling & General Administrative Expenses. We define X_reg as a matrix containing 5,000 records of the explanatory variables and y_reg as a dataframe containing the industry codes. We split the data into 80/20 training and test sets, using random state = 42. Then we standardized the features using StandardScaler() for better clustering performance. Then we find the hyperparameters using GridSearchCV where “c” = 10, “gamma” = scale and “kernel”: linear.

Evaluating the SVC model

The classification report presents the precision, recall, and F1-score for each industry code when using Support Vector Classification (SVC).



```
SVC Classification Accuracy: 0.762
Classification Report:
              precision    recall  f1-score   support

      1010.0       1.00      1.00      1.00        62
      1510.0       1.00      1.00      1.00        97
      2010.0       0.59      1.00      0.74        91
      2020.0       1.00      0.24      0.38        55
      2030.0       1.00      0.66      0.80        62
      2510.0       0.00      0.00      0.00         7
      2520.0       0.53      0.96      0.68        45
      2530.0       0.88      0.29      0.44        24
      2540.0       1.00      0.17      0.29         6
      2550.0       0.92      0.57      0.71        21
      3010.0       0.00      0.00      0.00         3
      3020.0       0.74      1.00      0.85        29
      3030.0       1.00      0.36      0.53        11
      3510.0       0.72      1.00      0.84        81
      3520.0       1.00      0.14      0.24        37
      4010.0       0.80      0.36      0.50        11
      4020.0       0.91      0.73      0.81        44
      4030.0       0.69      1.00      0.82        45
      4040.0       0.00      0.00      0.00         5
      4510.0       0.91      0.19      0.32        52
      4520.0       0.53      0.97      0.69        64
      4530.0       0.95      0.59      0.73        32
      5010.0       1.00      0.73      0.85        15
      5020.0       0.64      1.00      0.78         7
      5510.0       1.00      1.00      1.00        55
      6010.0       1.00      0.33      0.50        12
      6020.0       0.77      1.00      0.87        27

    accuracy                           0.76      1000
   macro avg       0.76      0.60      0.61      1000
weighted avg       0.83      0.76      0.73      1000
```

Accuracy = 76.2%: The model correctly classifies 76.2% of the industry codes.
Macro Avg (F1-score) = 0.61: Indicates an imbalanced performance across classes.
Weighted Avg (F1-score) = 0.73: Suggests that the model is somewhat biased toward majority classes.

Strong Performing Classes**:** Industry codes 1010.0, 1510.0, 2020.0, 2030.0, 3030.0, 3520.0, 6010.0 all have high precision, recall, and F1-scores. This suggests that these industries have distinct features that the model easily differentiates.

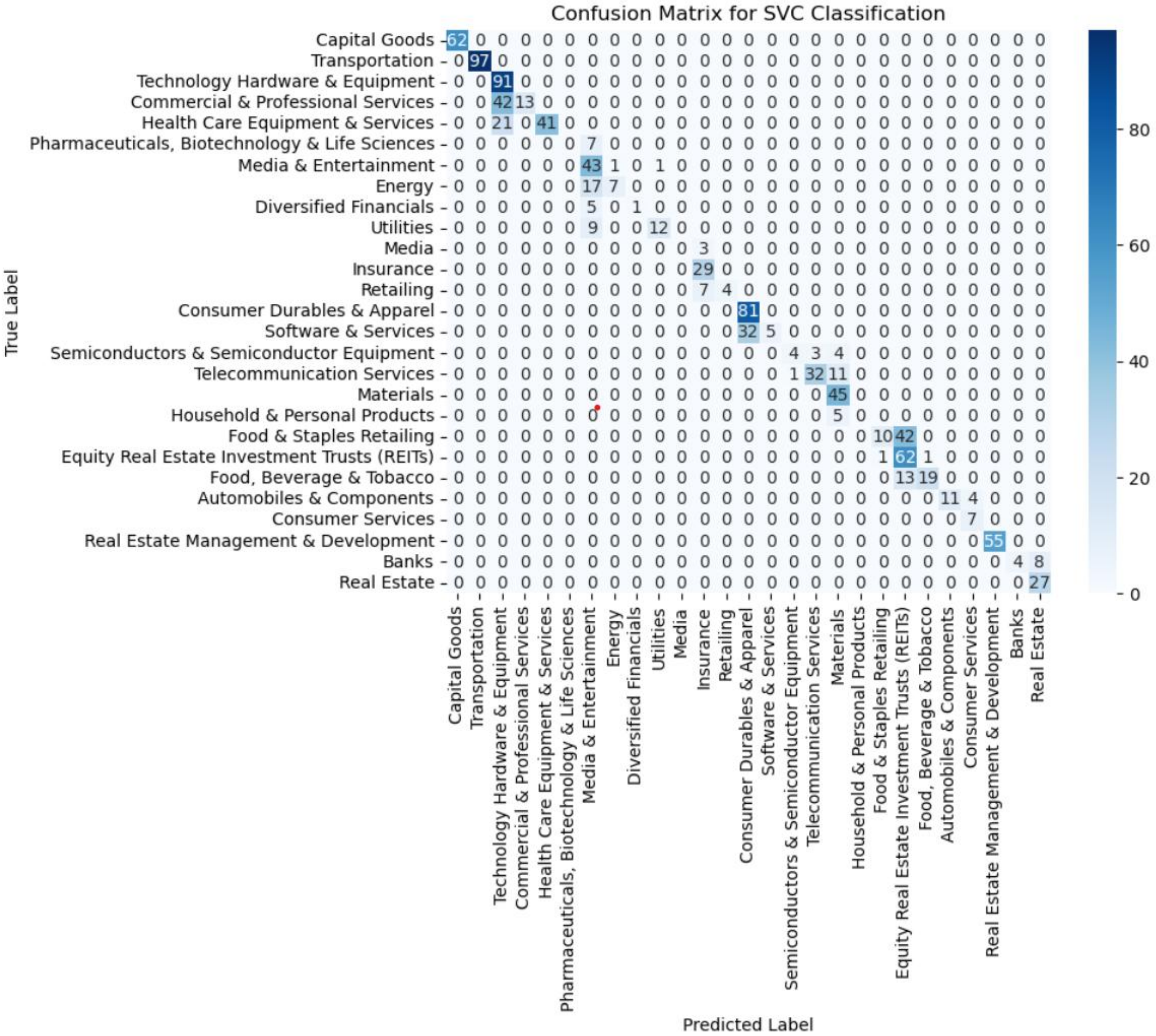
Poor Performing Classes: 2510.0, 2540.0, 3010.0, 3520.0, 4510.0 have low F1-scores (< 0.50), indicating misclassifications and low recall.
Some industry codes have very low support values (e.g., 2510.0, 3010.0, 5010.0). These small sample sizes make classification harder and lead to low recall.

Key Takeaways

- Model performs well on common industry codes (high support values).
- Rare industry codes suffer from low recall (model struggles with minority classes).
- Imbalanced dataset likely causing bias toward major industry codes.

Confusion Matrix

The confusion matrix helps us understand how well the Support Vector Classifier (SVC) is classifying different industries. The resulting matrix is shown below.



	Capital Goods	Transportation	Technology Hardware & Equipment	Commercial & Professional Services	Health Care Equipment & Services	Pharmaceuticals, Biotechnology & Life Sciences	Media & Entertainment	Energy	Diversified Financials	Utilities	Media	Insurance	Retailing	Consumer Durables & Apparel	Software & Services	Semiconductors & Semiconductor Equipment	Telecommunication Services	Materials	Household & Personal Products	Food & Staples Retailing	Equity Real Estate Investment Trusts (REITs)	Food, Beverage & Tobacco	Automobiles & Components	Consumer Services	Real Estate Management & Development	Banks	Real Estate
Capital Goods	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Transportation	0	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Technology Hardware & Equipment	0	0	91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Commercial & Professional Services	0	0	42	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Health Care Equipment & Services	0	0	21	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pharmaceuticals, Biotechnology & Life Sciences	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Media & Entertainment	0	0	0	0	0	0	43	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Energy	0	0	0	0	0	0	17	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Diversified Financials	0	0	0	0	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Utilities	0	0	0	0	0	0	9	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Media	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Insurance	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retailing	0	0	0	0	0	0	0	0	0	0	0	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Consumer Durables & Apparel	0	0	0	0	0	0	0	0	0	0	0	0	0	81	0	0	0	0	0	0	0	0	0	0	0	0	0
Software & Services	0	0	0	0	0	0	0	0	0	0	0	0	0	32	5	0	0	0	0	0	0	0	0	0	0	0	0
Semiconductors & Semiconductor Equipment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	4	0	0	0	0	0	0	0	0	0
Telecommunication Services	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	32	11	0	0	0	0	0	0	0	0	0
Materials	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0
Household & Personal Products	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
Food & Staples Retailing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	42	0	0	0	0	0	0
Equity Real Estate Investment Trusts (REITs)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	62	1	0	0	0	0	0
Food, Beverage & Tobacco	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	19	0	0	0	0	0
Automobiles & Components	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	4	0	0	0
Consumer Services	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
Real Estate Management & Development	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	0	0
Banks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	8
Real Estate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27

Overall Model Performance

Many industries have high diagonal values, indicating correct classifications. However, some misclassifications exist, especially for similar industries. Some classes have zero or very few predictions, suggesting poor performance for minority classes.

Strong Performing Industries (High Correct Predictions): Capital Goods (62), Transportation (97), Technology Hardware (91). These industries have high diagonal values, meaning the model classifies them correctly with high accuracy.

Industries with Misclassification Issues: Commercial & Professional Services (13 Correct, 42 Misclassified)
Health Care Equipment & Services (41 Correct, 21 Misclassified)
Retailing (4 Correct, 7 Misclassified)
Banks (4 Correct, 8 Misclassified)

Energy (7 Correct, 17 Misclassified)
These industries have noticeable misclassification rates, possibly due to similar feature distributions.

Industries with Poor Classification: Pharmaceuticals, Biotechnology & Life Sciences (7 Misclassified).
Household & Personal Products (5 Misclassified)
Media (3 Misclassified)
These industries are not classified well, likely due to small sample sizes or overlapping feature sets.

Possible Reasons for Misclassifications

- Feature Overlap: Some industries may have similar textual or numerical patterns, making them difficult to distinguish.
- Class Imbalance: Certain industries may have fewer training samples, causing the model to misclassify them.