FedCSIS 2024 Data Science Challenge: Predicting Stock Trends

Statistical Learning and Data Mining Final Assignment

Team Name: Unicas15

- Mahdi Islam
- Musarrat Tabassum

Contents

- Data Overview
- Evaluation Metrics
- Initial Insights
- Detailed Analysis
- Data Preprocessing
- Model Selection
- Hyperparameter Tuning
- Model Prediction & Result Analysis

Data Overview

- The goal of the challenge is the prediction of stock trends.
- Dataset contains key financial indicators for 300 companies chosen from 11 different sectors from 10 years.
- Each company is described by values of 58 indicators that are derived from its financial statements.
- Dataset also contains information on 1 year change for each indicator indicating a trend in the considered values.

Evaluation Metrics

The evaluation metric for this challenge is the average error cost measure with the error cost matrix given below:

	-1	0	1
-1	0	1	2
0	1	0	1
1	2	1	0

Table 1: Error cost matrix

- The error value is computed as:

err = (confusion_matrix(preds, gt) * cost_matrix)/length(gt));

where the multiplication is done element-wise.
Here, gt =Ground Truth, preds = Model Prediction.

Class 1 : Decision to buy.
Class 0 : Decision to not trade at all.
Class -1: Decision to sell.

Initial Insights

Initial Insights

- Number of samples in train and test set are distributed across 11 sectors.
- Figure 1 below shows the similar distribution of both train and test set.

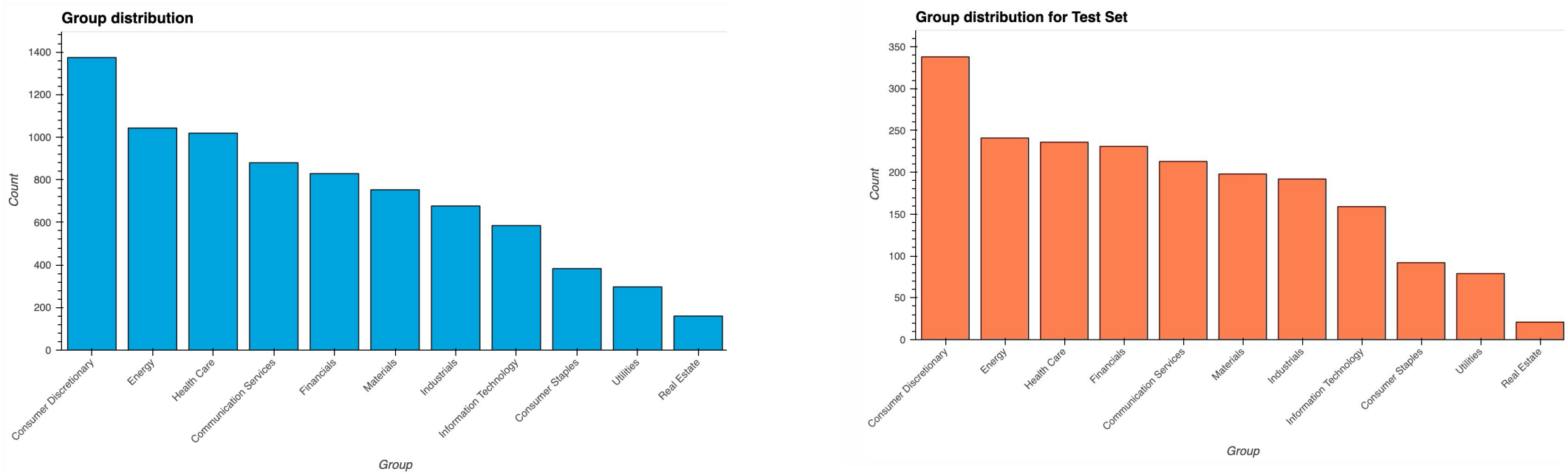

Figure 1: Distribution of 11 sectors

Initial Insights

- Total missing values in train set: 28933. Among them 26174 values are not available (NA) and the rest are empty strings regarded as not applicable.
- Total missing values in test set: 6784. Among them 6094 values are not available (NA) and the rest are empty strings regarded as not applicable.

- Number of samples in train set: 8000
- Number of samples in test set: 2000
- Number of column in train set: 119
- Number of columns in test set: 117

MAIA

Detailed Analysis

Detailed Analysis

In both train and test set, same indicators have similar percentage of missing values where some of the columns have around 20% missing values as shown in Figure 2 and Figure 3.

Indicator Names	*NA Count*	NA Percentage
1-year Absolute Change of Inventory Turnover, TTM	367	18.35
1-year Absolute Change of Quick Ratio	367	18.35
1-year Absolute Change of Inventories Percentage of Total Assets	367	18.35
Inventory Turnover, TTM	367	18.35
Inventories Percentage of Total Assets	361	18.05
Quick Ratio	361	18.05
1-year Absolute Change of Accounts Receivable Turnover, TTM	169	8.45
Accounts Receivable Turnover, TTM	166	8.3
1-year Absolute Change of Payables Turnover, TTM	155	7.75
Payables Turnover, TTM	155	7.75

Figure 2

#	Indicator	Indicator Names	NA Count	NA Percentage
0	I21	Inventory Turnover, TTM	1,549	19.3625
1	dI21	1-year Absolute Change of Inventory Turnover, TTM	1,549	19.3625
2	dI50	1-year Absolute Change of Quick Ratio	1,549	19.3625
3	dI48	1-year Absolute Change of Inventories Percentage of Total Assets	1,549	19.3625
4	I50	Quick Ratio	1,520	19.0
5	I48	Inventories Percentage of Total Assets	1,520	19.0
6	dI24	1-year Absolute Change of Accounts Receivable Turnover, TTM	713	8.9125
7	I24	Accounts Receivable Turnover, TTM	701	8.7625
8	dI14	1-year Absolute Change of Payables Turnover, TTM	671	8.3875
9	I14	Payables Turnover, TTM	671	8.3875
10	dI26	1-year Absolute Change of Average Net Trade Cycle Days, TTM	535	6.6875

Figure 3

MAIA

Detailed Analysis

Communication Services	Consumer Discretionary	Energy	Health Care	Materials	Real Estate	Utilities	Indicators	Indicator Names
880	142	132	60	54	160	121	I21	Inventory Turnover, TTM
880	142	132	60	54	160	92	I50	Quick Ratio
880	142	132	60	54	160	121	dI50	1-year Absolute Change of Quick Ratio
880	142	132	60	54	160	121	dI21	1-year Absolute Change of Inventory Turnover, TTM
880	142	132	60	54	160	121	dI48	1-year Absolute Change of Inventories Percenta...
880	142	132	60	54	160	92	I48	Inventories Percentage of Total Assets
637	0	0	34	0	0	0	I14	Payables Turnover, TTM
637	0	0	34	0	0	0	dI14	1-year Absolute Change of Payables Turnover, TTM
489	0	0	34	0	0	0	I49	Current Ratio
489	0	0	34	0	0	0	I51	Working Capital to Total Assets
489	0	0	34	0	0	0	I52	Cash Ratio
489	0	0	34	0	0	0	dI45	1-year Absolute Change of Total Current Assets...
489	0	0	34	0	0	0	dI4	1-year Absolute Change of EBITDA Percentage of...
489	0	0	34	0	0	0	dI27	1-year Absolute Change of Current Asset Turnov...
489	0	0	34	0	0	0	I27	Current Asset Turnover, TTM
489	0	0	34	0	0	0	I26	Average Net Trade Cycle Days, TTM
489	0	0	34	0	0	12	dI26	1-year Absolute Change of Average Net Trade Cy...
489	47	0	34	78	0	0	I24	Accounts Receivable Turnover, TTM
489	0	0	34	0	0	0	I45	Total Current Assets Percentage of Total Assets
489	0	0	34	0	0	0	dI46	1-year Absolute Change of Total Current Liabil...
489	0	0	34	0	0	0	dI15	1-year Absolute Change of Cash Flow from Opera...
489	0	0	34	0	0	0	dI49	1-year Absolute Change of Current Ratio
489	0	0	34	0	0	0	I12	Working Capital Percentage of Total Revenue, Y...
489	0	0	34	0	0	0	dI12	1-year Absolute Change of Working Capital Perc...

Communication services sector has the highest number of missing values.

Detailed Analysis

Total Missing Values by Sector

#	*Total Missing Values*	Percentage	Sector Names
0	19,784	75.58646	Communication Services
5	1,445	5.520746	Health Care
3	1,134	4.332544	Energy
9	960	3.667762	Real Estate
1	946	3.614274	Consumer Discretionary
8	944	3.606633	Materials
10	749	2.861618	Utilities
4	212	0.809964	Financials
2	0	0.0	Consumer Staples
6	0	0.0	Industrials
7	0	0.0	Information Technology

Total Non-Applicable Values

#	Total Non-Applicable Values	Percentage	Sector Names
0	398	14.37342	Communication Services
1	265	9.570242	Consumer Discretionary
2	400	14.445648	Consumer Staples
3	391	14.120621	Energy
4	193	6.970025	Financials
5	200	7.222824	Health Care
6	237	8.559047	Industrials
7	171	6.175515	Information Technology
8	234	8.450704	Materials
9	35	1.263994	Real Estate
10	235	8.486818	Utilities

75% of the total missing values belong to the Communication Services sector.

Percentage of Non-Applicable values are evenly distributed among the sectors.

MAIA

Detailed Analysis

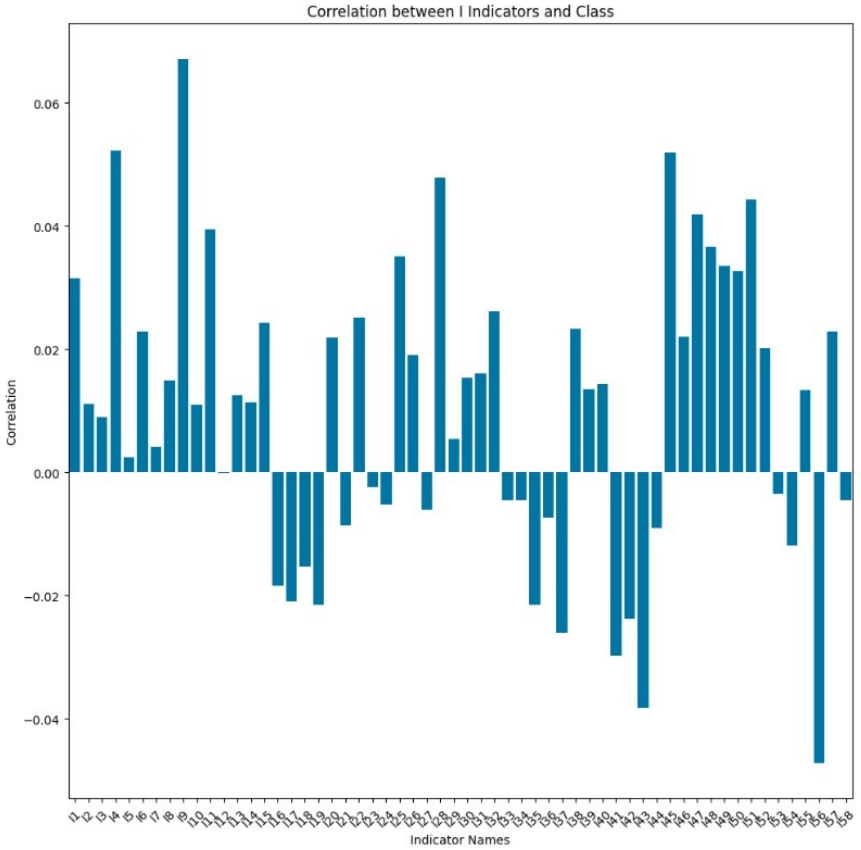

Figure-4

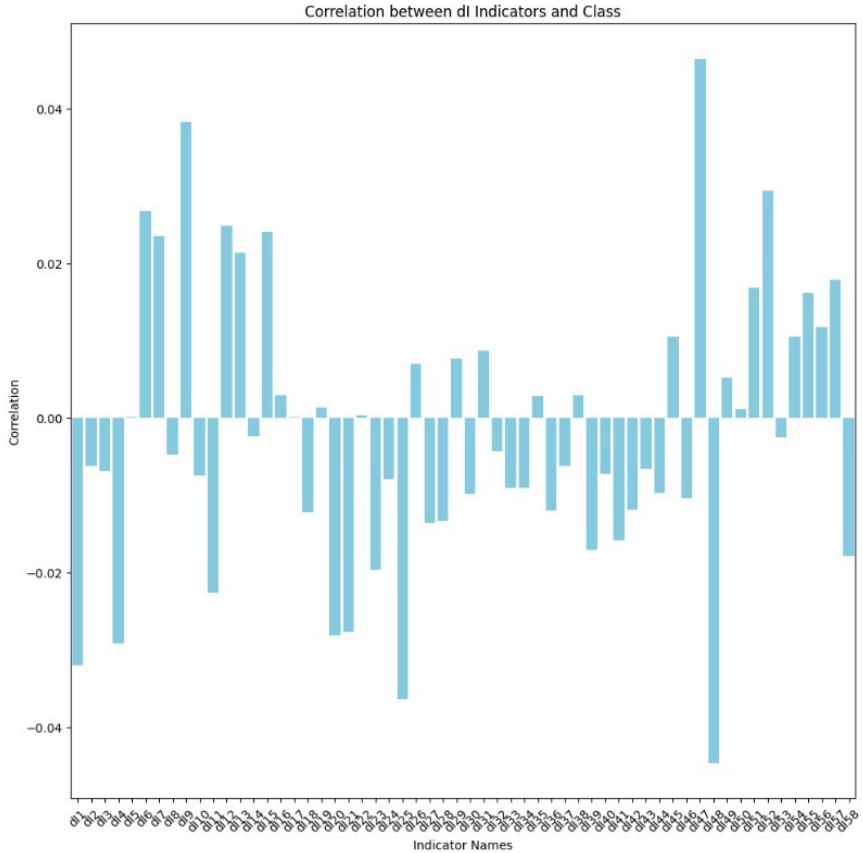

Figure-5

Correlation between 58 Indicators (Figure-4) and their derivatives (Figure-5) with respect to the target column ‘Class’ can be seen from the bar plots.

Data Preprocessing

Data Preprocessing

Handling Missing Values

- Non-Applicable/empty missing values were replaced with 0.
- Not available (NA) missing values were handled in two different ways.
  a. Imputed by median of each column.
  b. Imputed using advanced imputing technique ‘missForest’ which uses columns with non-missing value of the same row to predict missing value using iterative imputing and ‘Random Forest’ machine learning algorithm.

Data Preprocessing

Outlier Detection

- The dataset contained severe outliers and most of the columns were skewed. The outliers are visualized in a lower dimension in Fig-7 using PCA.

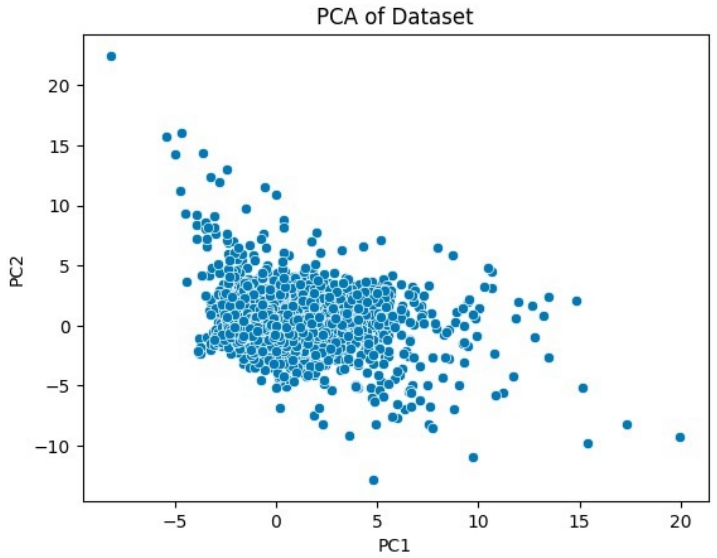

Figure-6

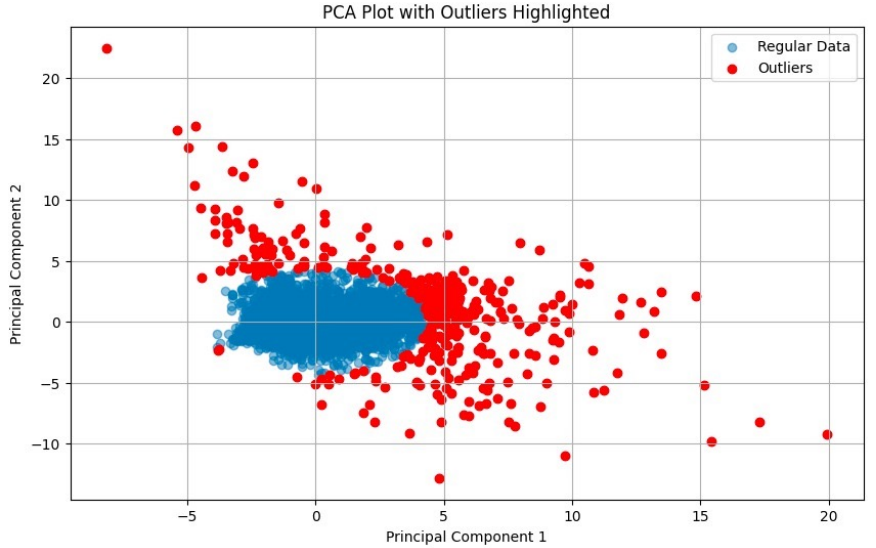

Figure-7

Data Preprocessing

Feature Engineering

```
#changing paramters in train set
df_train['Profatibility_comopiste'] = df_train[['I1','I2','I3','I4']].sum(axis = 1)/4
df_train['Liquidity_coposite'] = df_train[['I50','I51','I53']].sum(axis = 1)/3
df_train['Leverage_composite'] = df_train[['I54','I55','I56']].sum(axis = 1)/3
df_train['Operational_efficiency'] = df_train[['I22','I23','I24','I25','I26']].sum(axis = 1)/5
df_train['Validation_composite'] = df_train[['I39','I40','I41','I42','I43']].sum(axis = 1)/5

df_train['yr_Profatibility_comopiste'] = df_train[['dI1','dI2','dI3','dI4']].sum(axis = 1)/4
df_train['yr_Liquidity_coposite'] = df_train[['dI50','dI51','dI53']].sum(axis = 1)/3
df_train['yr_Leverage_composite'] = df_train[['dI54','dI55','dI56']].sum(axis = 1)/3
df_train['yr_Operational_efficiency'] = df_train[['dI22','dI23','dI24','dI25','dI26']].sum(axis = 1)/5
df_train['yr_Validation_composite'] = df_train[['dI39','dI40','dI41','dI42','dI43']].sum(axis = 1)/5
```

Due to similarity in the properties some of the financial indicators were combined to create single Indicators.

Data Preprocessing

Feature Selection and SMOTE

- For some trials columns containing more than 5% missing values were dropped.
- SMOTE algorithm was used in a few trials to tackle data imbalance of target column ‘Class’. It made the models overfit during training so later it was not used.

Model Selection

Model Selection

Trial 1

- **For the first trial, Random Forest was used. It was preferred because:**
    a) It is robust to outliers.
    b) Robust to unscaled data so doesn't requires scaling or normalizing.
    c) Ensemble of decision trees (bagging) so can handle complex data.
    d) Was combined with both median imputation and missForest imputation.

Model Selection

Trial 2

- **For the second trial, LightGBM was used. It was preferred because:**
    a) It's extremely fast, parallelizable.
    b) Robust to unscaled data so doesn't requires scaling or normalizing.
    c) Ensemble of decision trees (boosting) so can handle complex data.
    d) Was used with median imputation.

Model Selection

Trial 3

- **For the third trial, Histogram Gradient Boosting was used. It was preferred because:**
    a) It's extremely fast, parallelizable.
    b) Robust to unscaled data so doesn't requires scaling or normalizing.
    c) Ensemble of decision trees (boosting) so can handle complex data.
    d) It can also handle missing data, so no imputation was used.

Hyperparameter Tuning

- For hyperparameter tuning, Optuna was used to determine the best set of parameters.
- Optuna used the following function to minimize the loss.

*err = (confusion_matrix(preds, gt) * cost_matrix)/length(gt))*

Model Prediction & Result Analysis

- Random Forest with median imputing achieved best result: 0.7921 error score.
- Random Forest with missForest imputation achieved second best result: 0.7970 error score.
- Histogram Gradient Boosting also achieved a score of 0.7970.
- LightGBM with KNN imputation achieved a score of 0.8060.
- A voting classifier of weighted average of Random Forest, HGB and LightGBM got a score of 0.83.

Thank you