

LAMARR

Institute for Machine Learning
and Artificial Intelligence



CENTER FOR TRUSTWORTHY
DATA SCIENCE AND SECURITY
RESEARCH ALLIANCE

tu technische universität
dortmund

Case Study: Automated Data Analytics Using LLMs

Simon Klüttermann, Emmanuel Müller

17.07.2024

TU Dortmund University - Department of Computer Science

Every Winter Semester: Data Mining Cup
Solve Data Mining Challenges together
& compete against other groups



FedCSIS 2024 Data Science

Challenge: Predicting Stock Trends



1 month, 2 weeks ago

FedCSIS 2024 Data Science Challenge: Predicting Stock Trends is the 10th data science challenge organized under the auspices of the Conference on Computer Science and Intelligence Systems (<https://fedcsis.org/>). In this anniversary edition, the task is related to financial data - participants are asked to predict the performance of investments in selected stocks from several industry sectors. The competition is sponsored by Yettel.Bank (Former Mobi Banka) (<https://www.yettelbank.rs/en/>) and the Conference on Computer Science and Intelligence Systems series.

Overview



The topic of this year's data science competition is the prediction of stock trends. The dataset contains key financial indicators for 300 companies chosen from 11 different sectors of the S&P 500 index, from 10 years. Each company is described by values of 58 indicators that are derived from its

"My Students" Code

Last Week I was talking to a thesis student

And he wanted to change a plot a bit

To do this, he copied all his code into Gemini (Google Chatgpt) and let it change it

```
1 import json
2 import matplotlib.pyplot as plt
3
4
5 with open('results_per_algorithm.json', 'r') as f:
6     data = json.load(f)
7
8
9 datasets = list(data['results'].keys())
10 percent_corr_values = [data['results'][dataset]['percent_corr'] for dataset in datasets]
11
12
13 plt.figure(figsize=(10, 6))
14 plt.scatter(datasets, percent_corr_values, color='blue', marker='o')
15
16
17 plt.xlabel('Dataset - Removed Algorithm')
18 plt.ylabel('Percentage of Combined AUC > 0.5')
19 plt.title('Scatter Plot of Percentage Values for Each Dataset and Removed Algorithm')
20 plt.xticks(rotation=90)
21 plt.grid(True)
22 plt.tight_layout()
23
24
25 plt.savefig('plots/scatter_plot.png')
26
27
28 plt.show()
```

So, do I even need the student?

Or how much of his work can we automatize here?

Let's find out together!

In this Case Study: Try to solve Data Mining Challenges as automatically as possible

Searching for about 12 Students

Separate the task into subtasks

- Data handling

- Exploratory data analysis

- Feature Engineering

- ...

See how much of this a LLM can do

Possibility to write (and present) a research paper

Any Questions?

Write me: Simon.Kluettermann@cs.tu-dortmund.de

