Overview of Project Developments and Applications in Data Science

Yang, Chia-Shuo

Table of contents

a. Projects development and application

b. Projects from Work and Internship Experience

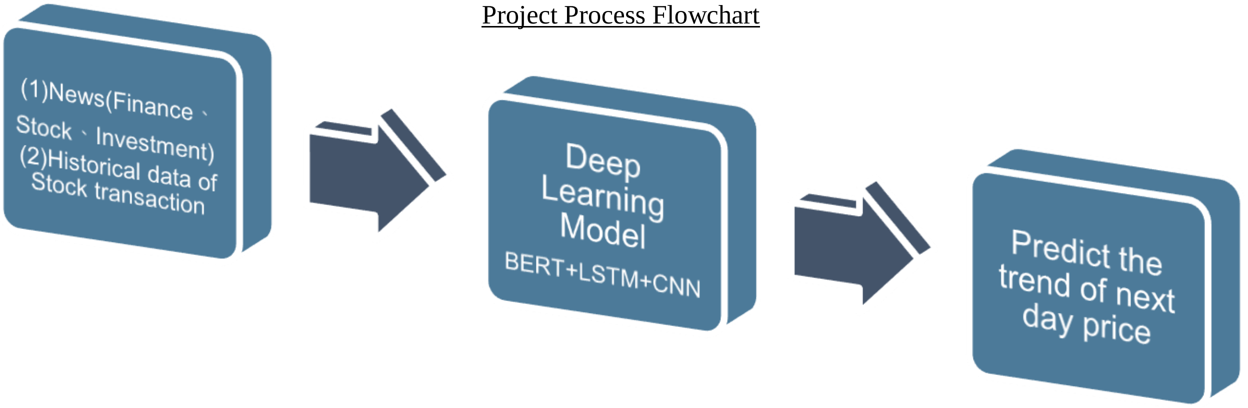
Project development and application - Cultivating programming skills through problem discovery and solving.

- **Project development and application - Using news to judge the trend of the stock market**

During my time working in PwC, I not only developed financial programs to optimize workflows but also accumulated extensive financial expertise. I also had exposure to various industries and participated in IPO (initial public offering) cases for multiple companies. After gaining experience at the firm, I participated in an industry trial program at the Institute for Information Industry to further study model building using machine learning, deep learning, and natural language analysis techniques, which were then applied to actual projects.

International events, stock trading information, futures trading information, and financial reports are important indicators when the firm conducts financial outlooks and predicts company performance. During the program, I used deep learning models to predict stock market trends to see if it could be effective. After continuous experimentation and effort, the model prediction achieved 87%, surpassing the initial benchmark of 70% from the reference literature. However, there is still room for improvement in the program's internal structure and language usage, so I hope to further my studies in graduate school to continue increasing my programming logic and abilities and gain more systematic and comprehensive programming skills from structured teaching in school.

Project Process Flowchart


(1)News(Finance、Stock、Investment)
(2)Historical data of Stock transaction
Deep Learning Model
BERT+LSTM+CNN
Predict the trend of next day price

Project Results Presentation

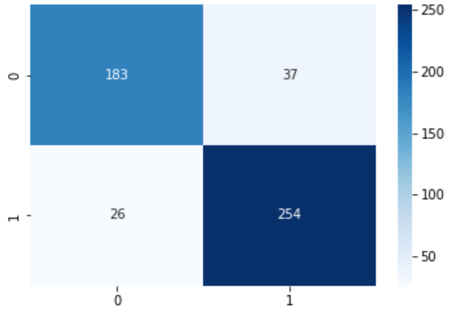
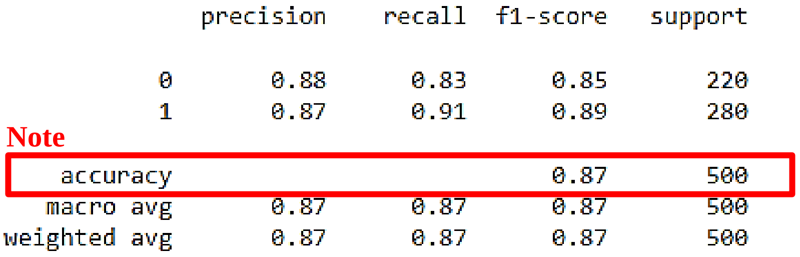


Figure 1-1 Confusion Matrix

Explanation for Figure 1-1:
0 represents a decrease in the stock market, while 1 represents an increase in the stock market.



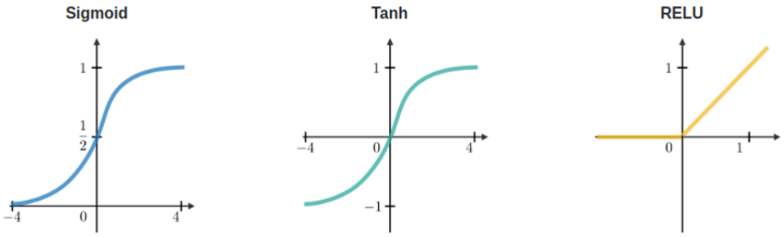
```
              precision    recall  f1-score   support

           0       0.88      0.83      0.85       220
           1       0.87      0.91      0.89       280
Note
    accuracy                           0.87       500
   macro avg       0.87      0.87      0.87       500
weighted avg       0.87      0.87      0.87       500
```

Figure 1-2 Classification

Explanation for Figure 1-2:
Based on the notes in the figure, it can be seen that the default accuracy rate is 87%, and both precision and recall also have good performance.

Activation Function



Important findings:
Activation function is a critical component in the model as it determines the "magnitude and sign" of the output value and has a crucial impact on model training. Adjustments need to be made for different datasets.

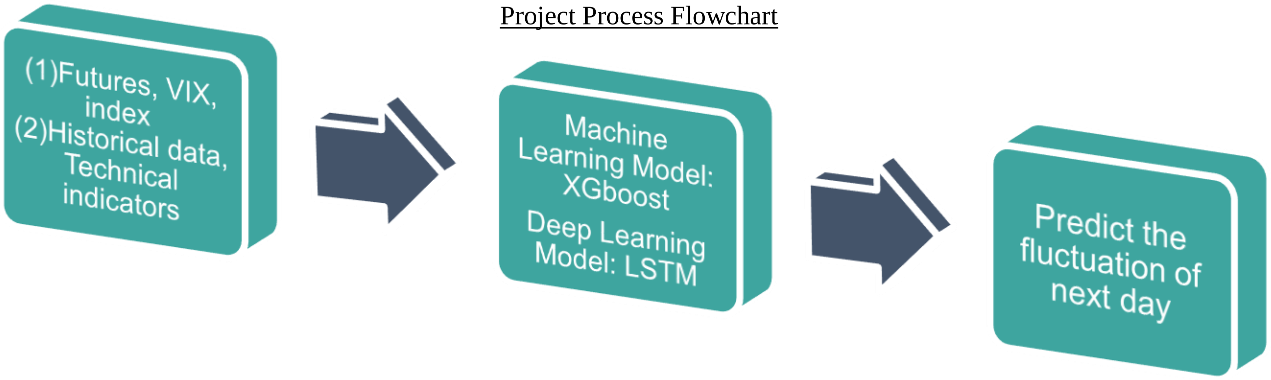
- **Project Development and Application - Predicting Stock Index Changes Through Futures Trading Information**

stock market trading history data and futures trading information are used to predict the magnitude of stock market fluctuations. Futures can be considered as a pioneer indicator of the stock market, as market trends can be inferred from the trend of futures, the price difference and deviation between futures and spot prices. Therefore, machine learning models and deep learning models are used to explore the relationship between futures and the stock market, to see if they can effectively predict stock market volatility.

This project aims to predict stock market fluctuation. Technical indicators such as Simple Moving Average (SMA), Relative Strength Index (RSI), and Moving Average Convergence Divergence (MACD) are used as features for training. In addition to predicting stock market volatility, F score is used to calculate the importance of each feature, and to determine the ranking and importance of each feature in predicting the target.

Although this project has shown positive results in predicting stock price fluctuation and trend (MAE decreased from 7.578 to 1.008), there are still many issues that need to be further studied, such as how to find the rules of volatility, how to incorporate effective features, and how to construct better models and practical applications. These are the goals of further research and development.

Project Process Flowchart



(1)Futures, VIX, index
(2)Historical data, Technical indicators

Machine Learning Model: XGboost
Deep Learning Model: LSTM

Predict the fluctuation of next day

Project Result Presentation (Experimental Group)

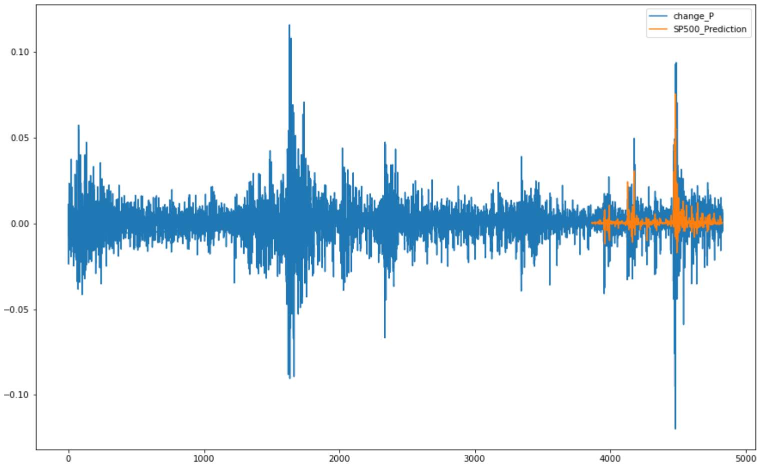


Figure 2-1: Comparison of **actual stock price volatility ratio (blue)** with predicted volatility ratio (orange)

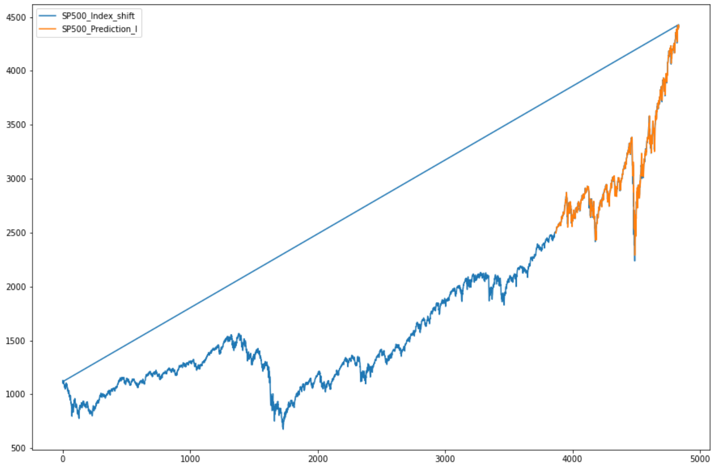


Figure 2-2: Comparison of **actual stock price (blue)** with predicted stock price (orange) converted from predicted volatility ratio

Experimental Group Result Explanation:

(1) Figures 2-1 and 2-2 are predictions made for the S&P 500 index. From Figure 2-1, we can see that the volatility is kept within the actual range, although the performance of profit limit for upward trends is limited, it also reduces the risk of loss during downward trends.

(2) Figure 2-2 shows that after converting to stock prices, the model can effectively capture the trend and trend of stock prices.

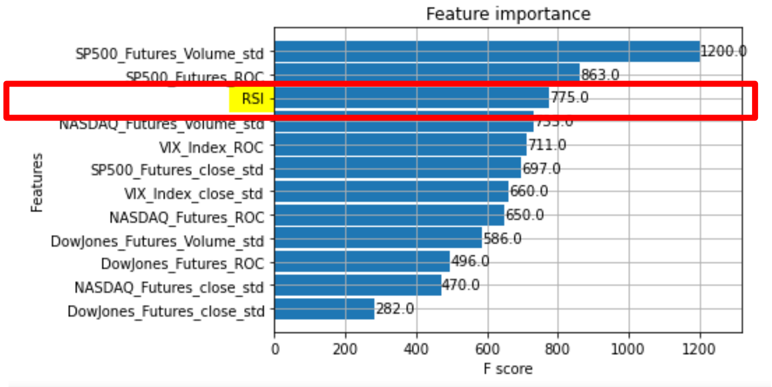


Figure 2-3: **Feature Importance**

Important findings:

After trial and adjustment, it was found that RSI can reduce MAE, and its effect is better than the other two technical indicators. At the same time, the F-score of each feature was calculated, and it was found that RSI is indeed an important indicator in this project prediction.

Mean Absolute Error = **1.007666**

Project Results Presentation (Control Group)

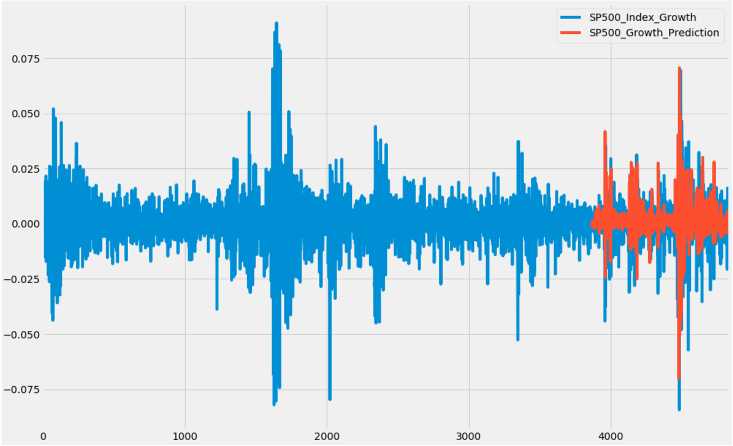


Figure 2-4: Comparison of **actual price volatility ratio (blue)** and predicted price volatility ratio (red)

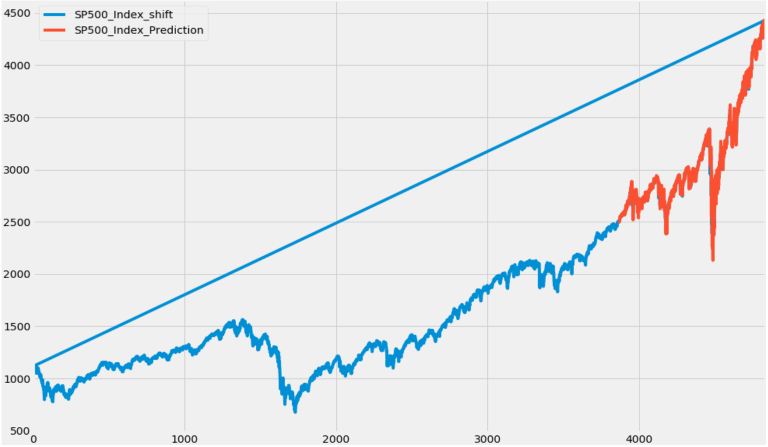


Figure 2-5: Comparison of **actual stock price (blue)** and predicted stock price (red) after conversion of volatility ratio

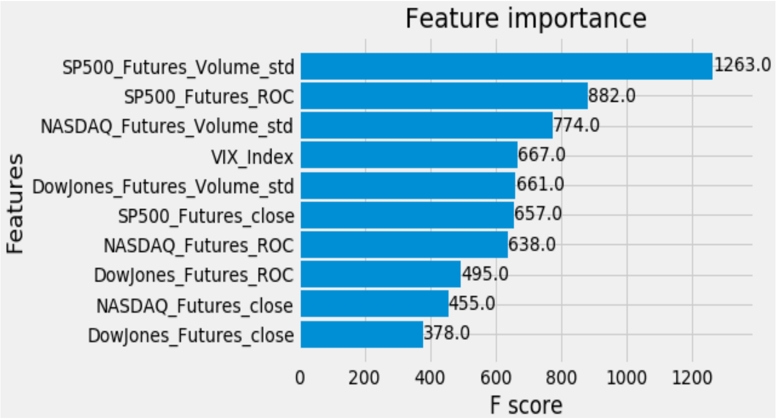


Figure 2-6: **Feature Importance**

Explanation of control group results:

The prediction results of Figures 2-4 and 2-5 show that although the trend of volatility is captured, the predicted fluctuation range of price movement will exceed the actual volatility, resulting in corresponding risks and larger losses.

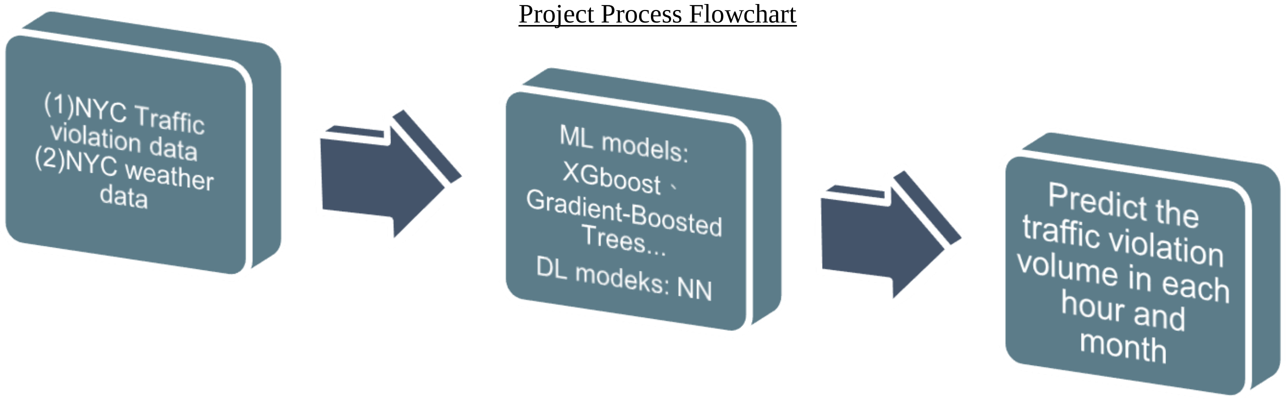
Mean Absolute Error = 7.57784

- **Project Development and Application - New York City Traffic Violation Data Analysis and Prediction**

In addition to the required coursework for my graduate program, I also enhanced my skills through programming courses offered by the Computer Science department. As I am passionate about data science, I am interested in exploring the analysis and application of big data. Therefore, I chose to take Professor Vincent Tseng's course on Big Data Analysis Techniques and Applications to gain further knowledge about big data and to learn Apache Spark and big data processing techniques through hands-on experience.

For this project, we focused on predicting and analyzing traffic violations in order to assist in city traffic planning, police deployment, and predicting violation occurrences under different conditions. Our dataset, which match 3V (Volume, Variety, Velocity) of big data. Our data include traffic information, violation records, geographical information, weather data, numeric types, and text. It is approximately 20 GB. In addition, we used an experimental and control group approach to examine whether weather factors could effectively improve prediction accuracy.

Project Process Flowchart


(1)NYC Traffic violation data
(2)NYC weather data
ML models:
XGboost、
Gradient-Boosted Trees...
DL modeks: NN
Predict the traffic violation volume in each hour and month

Project Result Presentation (Experimental Group)

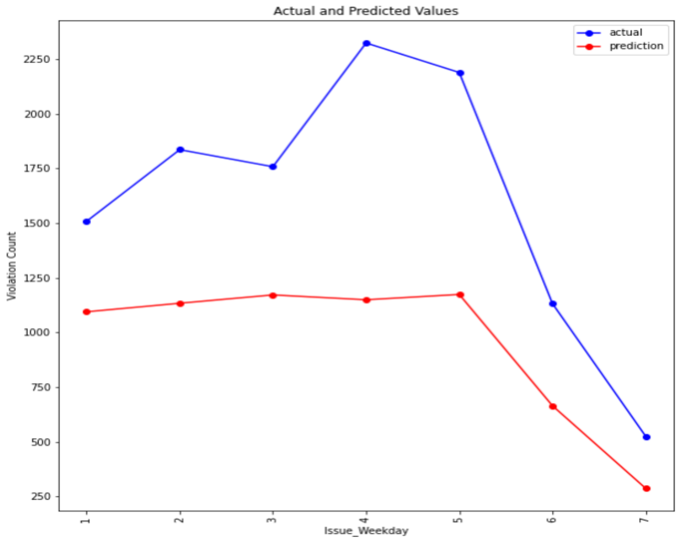


Figure 3-1: Comparison of **actual (blue)** and predicted (red) violation occurrences during the week

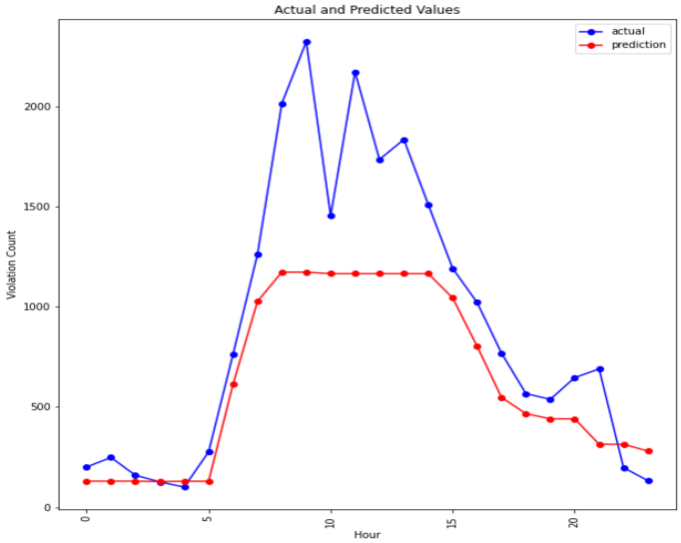


Figure 3-2: Comparison of **actual (blue)** and predicted (red) violation occurrences during the day

Explanation of experimental results:

(1) Figures 3-1 and 3-2 indicate that weather data is an important factor in predicting traffic violations, as the inclusion of weather data effectively captures patterns and trends.
(2) See the next page for a comparison of feature importance between the experimental and control groups.

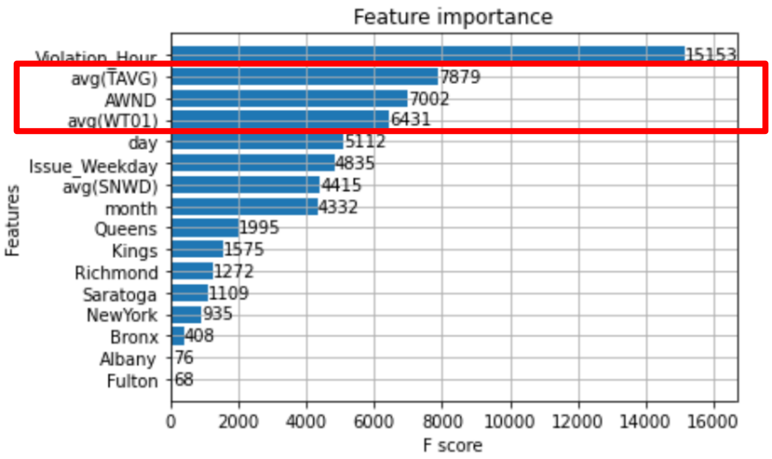


Figure 3-3: **Feature Importance**

Important findings:

After experimenting and calculating feature importance, it is evident from the red box that **weather data is a key factor** in predicting the number of violations, and **contributes significantly to improving prediction accuracy**.

R2 score = **0.7422**

Project Results Presentation (Control Group)

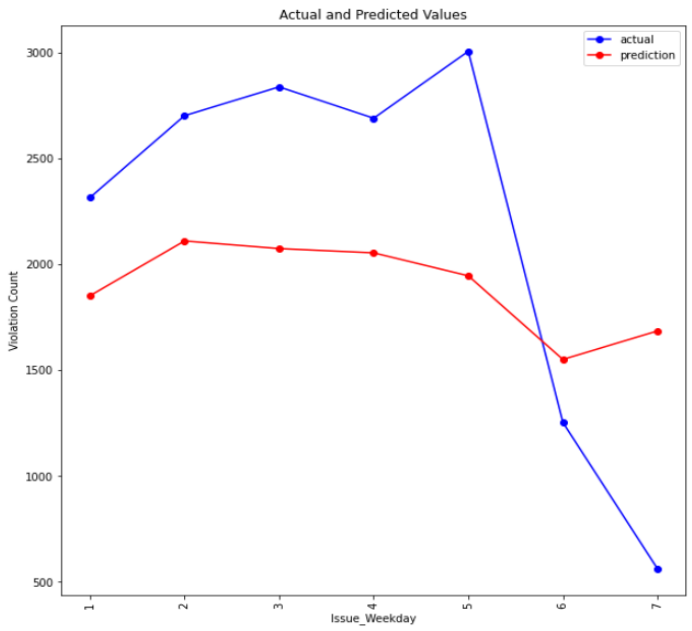


Figure 3-4: Comparison of **actual (blue)** and predicted (red) violation occurrences during the week

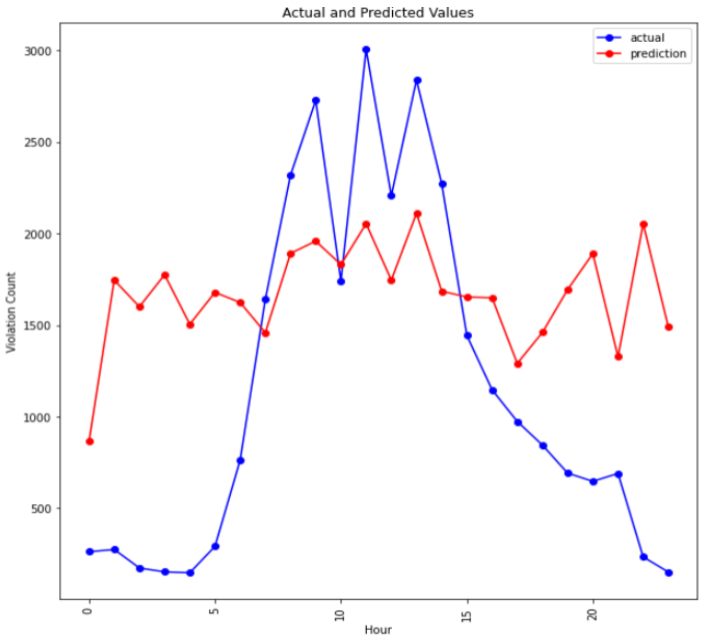


Figure 3-5: Comparison of **actual (blue)** and predicted (red) violation occurrences during the day

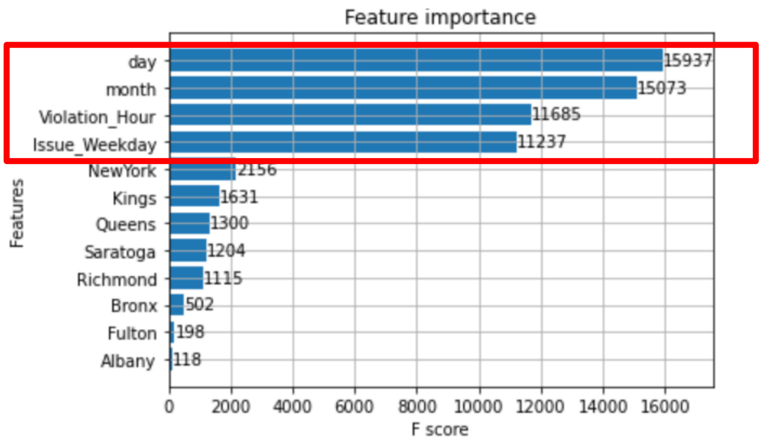


Figure 3-6: **Feature Importance**

Control Group Results Explanation:

The prediction results in Figures 3-4 and 3-5 show that without including weather data, **it is difficult to capture patterns effectively**, and the accuracy of the prediction decreases significantly.

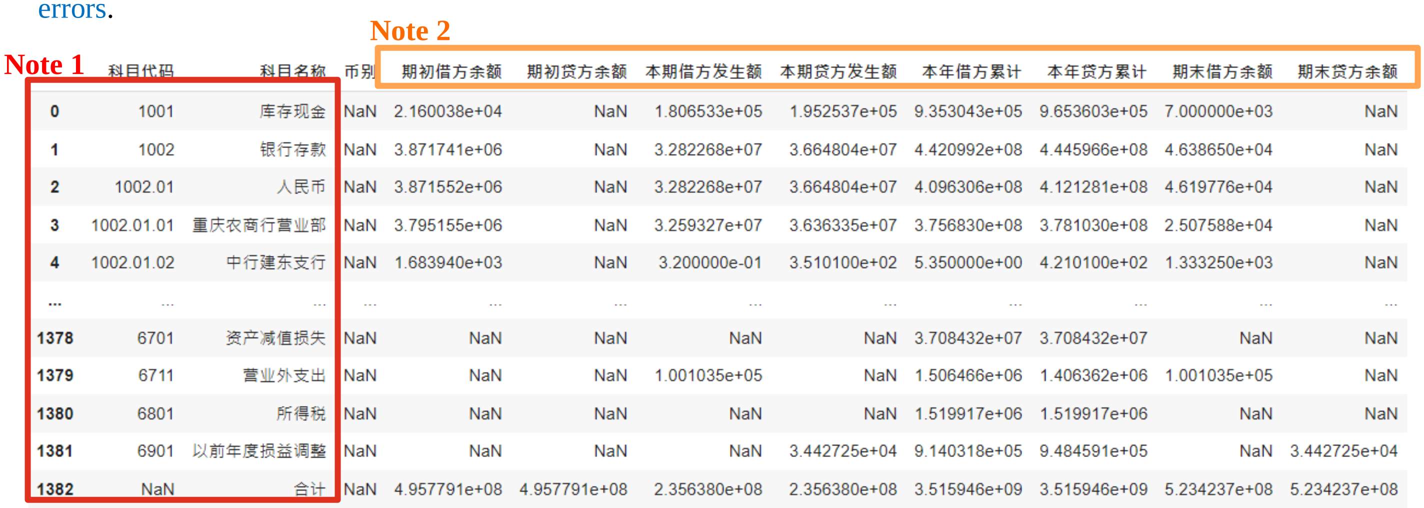
R2 score = **0.5744**

Projects from Work and Internship Experience:

- **Financial Tool Development - Filtering and Verification of Important Financial Accounting Items and Financial Statement Amounts**

During my work at the PwC, I had to deal with large amounts of financial data from different companies. Due to differences in accounting methods used by various institutions and units, it was necessary to standardize the financial data before analyzing and comparing it. In order to effectively shorten the time needed to standardize the accounting data, I decided to develop a financial tool to address this pain point.

To optimize the workflow, I recorded the accounting processing methods, logic, and related accounting details one by one, and used Python to establish programs for each process. This enabled the standardization of accounting information from different companies, and allowed for quick organization and analysis of financial data, while meeting accounting standards. As a result of this program, the workload that originally took two to three days was reduced to less than a morning to organize and verify the data, greatly improving work efficiency and reducing human errors.



	科目代码	科目名称	币别	期初借方余额	期初贷方余额	本期借方发生额	本期贷方发生额	本年借方累计	本年贷方累计	期末借方余额	期末贷方余额
0	1001	库存现金	NaN	2.160038e+04	NaN	1.806533e+05	1.952537e+05	9.353043e+05	9.653603e+05	7.000000e+03	NaN
1	1002	银行存款	NaN	3.871741e+06	NaN	3.282268e+07	3.664804e+07	4.420992e+08	4.445966e+08	4.638650e+04	NaN
2	1002.01	人民币	NaN	3.871552e+06	NaN	3.282268e+07	3.664804e+07	4.096306e+08	4.121281e+08	4.619776e+04	NaN
3	1002.01.01	重庆农商行营业部	NaN	3.795155e+06	NaN	3.259327e+07	3.636335e+07	3.756830e+08	3.781030e+08	2.507588e+04	NaN
4	1002.01.02	中行建东支行	NaN	1.683940e+03	NaN	3.200000e-01	3.510100e+02	5.350000e+00	4.210100e+02	1.333250e+03	NaN
...	...	...	...	...	...	...	...	...	...	...	...
1378	6701	资产减值损失	NaN	NaN	NaN	NaN	NaN	3.708432e+07	3.708432e+07	NaN	NaN
1379	6711	营业外支出	NaN	NaN	NaN	1.001035e+05	NaN	1.506466e+06	1.406362e+06	1.001035e+05	NaN
1380	6801	所得税	NaN	NaN	NaN	NaN	NaN	1.519917e+06	1.519917e+06	NaN	NaN
1381	6901	以前年度损益调整	NaN	NaN	NaN	NaN	3.442725e+04	9.140318e+05	9.484591e+05	NaN	3.442725e+04
1382	NaN	合计	NaN	4.957791e+08	4.957791e+08	2.356380e+08	2.356380e+08	3.515946e+09	3.515946e+09	5.234237e+08	5.234237e+08

Figure 4-1: Original Financial Data

Explanation of Figure 4-1: Original Financial Data

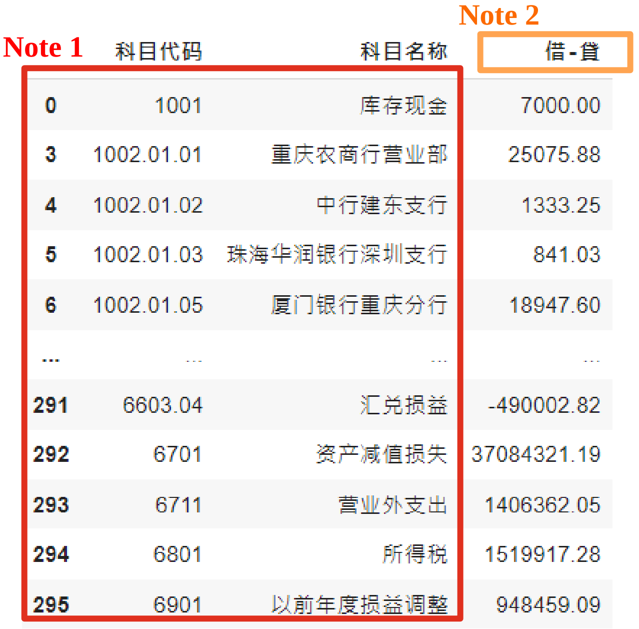
1. Note 1 in Figure 1 indicates that the institution had many duplicate and excessive items that would cause problems such as duplicate calculations or missing numbers during calculation and analysis, resulting in significant financial calculation errors.
2. Note 2 shows that there were missing values in the accounting numbers that needed to be reorganized and recalculated.

Result Explanation:

Using a program to organize and select the correct accounting items improved efficiency and reduced human error rates.

Note 1: There were a total of 1,382 original accounting items, but after reorganization, only 295 actual accounting items were used.

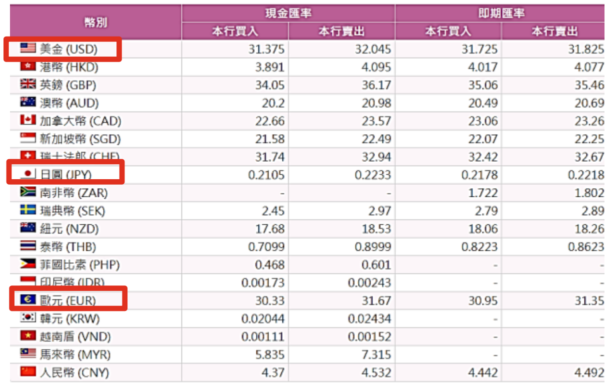
Note 2: The original financial data was also calculated, leaving only the final result.



	科目代码	科目名称	借-貸
0	1001	库存现金	7000.00
3	1002.01.01	重庆农商行营业部	25075.88
4	1002.01.02	中行建东支行	1333.25
5	1002.01.03	珠海华润银行深圳支行	841.03
6	1002.01.05	厦门银行重庆分行	18947.60
...	...	...	...
291	6603.04	汇兑损益	-490002.82
292	6701	资产减值损失	37084321.19
293	6711	营业外支出	1406362.05
294	6801	所得税	1519917.28
295	6901	以前年度损益调整	948459.09

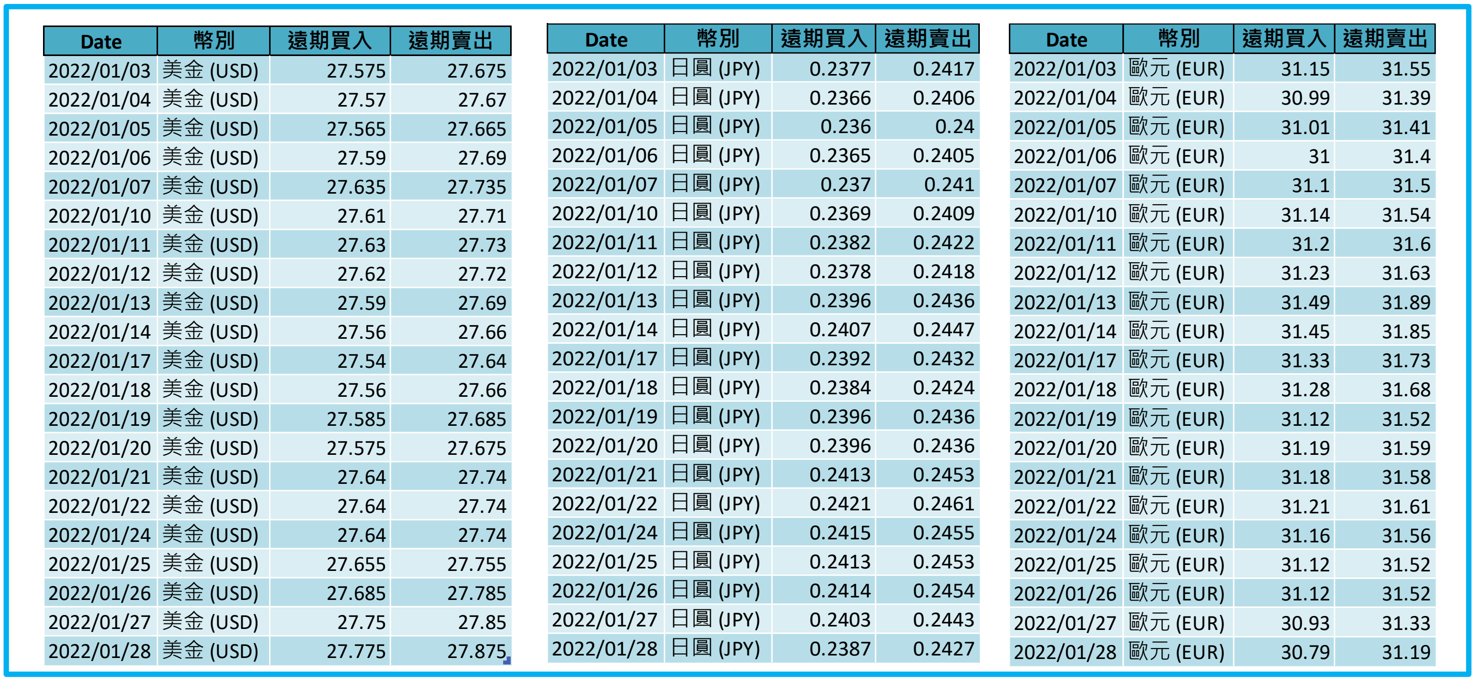
Figure 4-2: Data after Reorganization and Verification

- **Financial Tool Development - Data Acquisition, and Exchange Rate Analysis in Various Countries**
  - Additionally, as multiple companies were multinational enterprises and required currency conversion, a program was developed for updating exchange rates. Data acquisition and update were performed based on the required period and currency type. This not only improved efficiency and data accuracy but also significantly reduced the time required for manual searching of different currencies and the probability of human error.



幣別	現金匯率		即期匯率	
	本行買入	本行賣出	本行買入	本行賣出
美金 (USD)	31.375	32.045	31.725	31.825
港幣 (HKD)	3.891	4.095	4.017	4.077
英鎊 (GBP)	34.05	36.17	35.06	35.46
澳幣 (AUD)	20.2	20.98	20.49	20.69
加拿大幣 (CAD)	22.66	23.57	23.06	23.26
新加坡幣 (SGD)	21.58	22.49	22.07	22.25
瑞士法郎 (CHF)	31.74	32.94	32.42	32.67
日圓 (JPY)	0.2105	0.2233	0.2178	0.2218
南非幣 (ZAR)	-	-	1.722	1.802
瑞典幣 (SEK)	2.45	2.97	2.79	2.89
紐元 (NZD)	17.68	18.53	18.06	18.26
泰幣 (THB)	0.7099	0.8999	0.8223	0.8623
菲國比索 (PHP)	0.468	0.601	-	-
印尼幣 (IDR)	0.00173	0.00243	-	-
歐元 (EUR)	30.33	31.67	30.95	31.35
韓元 (KRW)	0.02044	0.02434	-	-
越南盾 (VND)	0.00111	0.00152	-	-
馬來幣 (MYR)	5.835	7.315	-	-
人民幣 (CNY)	4.37	4.532	4.442	4.492

Figure 5-1: Taiwan Bank Exchange Rate Table (Daily)

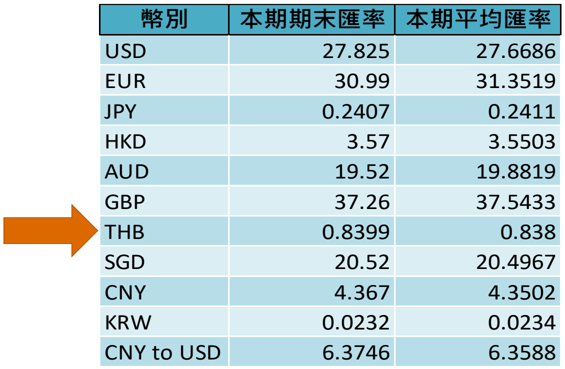


Date	幣別	遠期買入	遠期賣出
2022/01/03	美金 (USD)	27.575	27.675
2022/01/04	美金 (USD)	27.57	27.67
2022/01/05	美金 (USD)	27.565	27.665
2022/01/06	美金 (USD)	27.59	27.69
2022/01/07	美金 (USD)	27.635	27.735
2022/01/10	美金 (USD)	27.61	27.71
2022/01/11	美金 (USD)	27.63	27.73
2022/01/12	美金 (USD)	27.62	27.72
2022/01/13	美金 (USD)	27.59	27.69
2022/01/14	美金 (USD)	27.56	27.66
2022/01/17	美金 (USD)	27.54	27.64
2022/01/18	美金 (USD)	27.56	27.66
2022/01/19	美金 (USD)	27.585	27.685
2022/01/20	美金 (USD)	27.575	27.675
2022/01/21	美金 (USD)	27.64	27.74
2022/01/22	美金 (USD)	27.64	27.74
2022/01/24	美金 (USD)	27.64	27.74
2022/01/25	美金 (USD)	27.655	27.755
2022/01/26	美金 (USD)	27.685	27.785
2022/01/27	美金 (USD)	27.75	27.85
2022/01/28	美金 (USD)	27.775	27.875

Date	幣別	遠期買入	遠期賣出
2022/01/03	日圓 (JPY)	0.2377	0.2417
2022/01/04	日圓 (JPY)	0.2366	0.2406
2022/01/05	日圓 (JPY)	0.236	0.24
2022/01/06	日圓 (JPY)	0.2365	0.2405
2022/01/07	日圓 (JPY)	0.237	0.241
2022/01/10	日圓 (JPY)	0.2369	0.2409
2022/01/11	日圓 (JPY)	0.2382	0.2422
2022/01/12	日圓 (JPY)	0.2378	0.2418
2022/01/13	日圓 (JPY)	0.2396	0.2436
2022/01/14	日圓 (JPY)	0.2407	0.2447
2022/01/17	日圓 (JPY)	0.2392	0.2432
2022/01/18	日圓 (JPY)	0.2384	0.2424
2022/01/19	日圓 (JPY)	0.2396	0.2436
2022/01/20	日圓 (JPY)	0.2396	0.2436
2022/01/21	日圓 (JPY)	0.2413	0.2453
2022/01/22	日圓 (JPY)	0.2421	0.2461
2022/01/24	日圓 (JPY)	0.2415	0.2455
2022/01/25	日圓 (JPY)	0.2413	0.2453
2022/01/26	日圓 (JPY)	0.2414	0.2454
2022/01/27	日圓 (JPY)	0.2403	0.2443
2022/01/28	日圓 (JPY)	0.2387	0.2427

Date	幣別	遠期買入	遠期賣出
2022/01/03	歐元 (EUR)	31.15	31.55
2022/01/04	歐元 (EUR)	30.99	31.39
2022/01/05	歐元 (EUR)	31.01	31.41
2022/01/06	歐元 (EUR)	31	31.4
2022/01/07	歐元 (EUR)	31.1	31.5
2022/01/10	歐元 (EUR)	31.14	31.54
2022/01/11	歐元 (EUR)	31.2	31.6
2022/01/12	歐元 (EUR)	31.23	31.63
2022/01/13	歐元 (EUR)	31.49	31.89
2022/01/14	歐元 (EUR)	31.45	31.85
2022/01/17	歐元 (EUR)	31.33	31.73
2022/01/18	歐元 (EUR)	31.28	31.68
2022/01/19	歐元 (EUR)	31.12	31.52
2022/01/20	歐元 (EUR)	31.19	31.59
2022/01/21	歐元 (EUR)	31.18	31.58
2022/01/22	歐元 (EUR)	31.21	31.61
2022/01/24	歐元 (EUR)	31.16	31.56
2022/01/25	歐元 (EUR)	31.12	31.52
2022/01/26	歐元 (EUR)	31.12	31.52
2022/01/27	歐元 (EUR)	30.93	31.33
2022/01/28	歐元 (EUR)	30.79	31.19

Figure 5-2: Exchange Rate Data Extracted by Program



幣別	本期期末匯率	本期平均匯率
USD	27.825	27.6686
EUR	30.99	31.3519
JPY	0.2407	0.2411
HKD	3.57	3.5503
AUD	19.52	19.8819
GBP	37.26	37.5433
THB	0.8399	0.838
SGD	20.52	20.4967
CNY	4.367	4.3502
KRW	0.0232	0.0234
CNY to USD	6.3746	6.3588

Figure 5-3: Exchange Rate Fluctuation Analysis using the Program's Calculated Average and Period-End Exchange Rates.

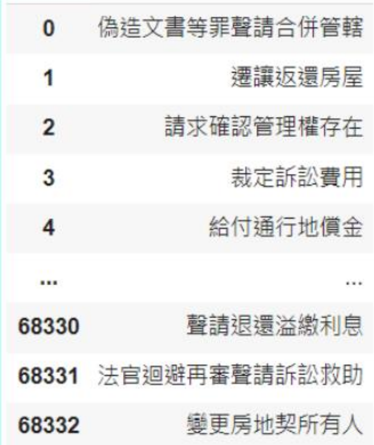
Explanation:

Figure 5-2 uses a program to extract exchange rate data for various currencies. The program then calculates the average and period-end exchange rates to analyze the fluctuations in exchange rates (as shown in Figure 5-3).

- **Application of Natural Language Processing Tools - Tools for Correspondence of Litigation Judgments and Laws, Legal Text Establishment, and Amount Extraction and Conversion**

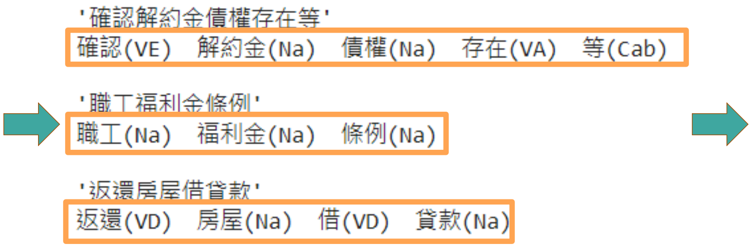
At the same time, during the summer internship, I was responsible for developing natural language analysis tools suitable for Traditional Chinese. The tools are used to extract violations of laws, important keywords, and compensation and fines from judicial judgments. The compensation amounts in Chinese characters are also converted into numerical format for ease of subsequent applications and classification. Furthermore, legal text is established for subsequent text mining.

Establishment Process Flowchart for Legal Text



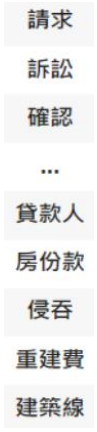
0	偽造文書等罪聲請合併管轄
1	遷讓返還房屋
2	請求確認管理權存在
3	裁定訴訟費用
4	給付通行地償金
...	...
68330	聲請退還溢繳利息
68331	法官迴避再審聲請訴訟救助
68332	變更房地契所有人

Figure 6-1: Original judgment document's ruling case



```
'確認解約金債權存在等'
確認(VE)  解約金(Na)  債權(Na)  存在(VA)  等(Cab)
'職工福利金條例'
職工(Na)  福利金(Na)  條例(Na)
'返還房屋借貸款'
返還(VD)  房屋(Na)  借(VD)  貸款(Na)
```

Figure 6-2: Part-of-speech tagging after sentence



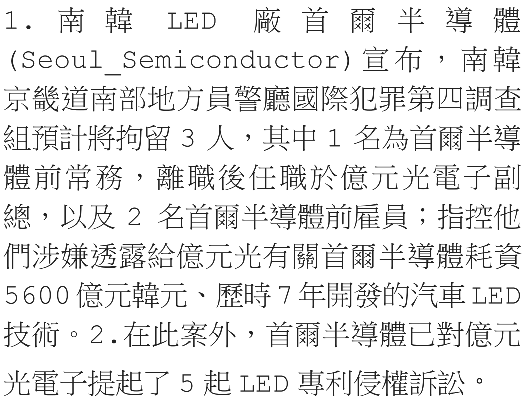
請求
訴訟
確認
...
貸款人
房份款
侵吞
重建費
建築線

Figure 6-3: Text establishment using different types of word segmentation

Explanation:

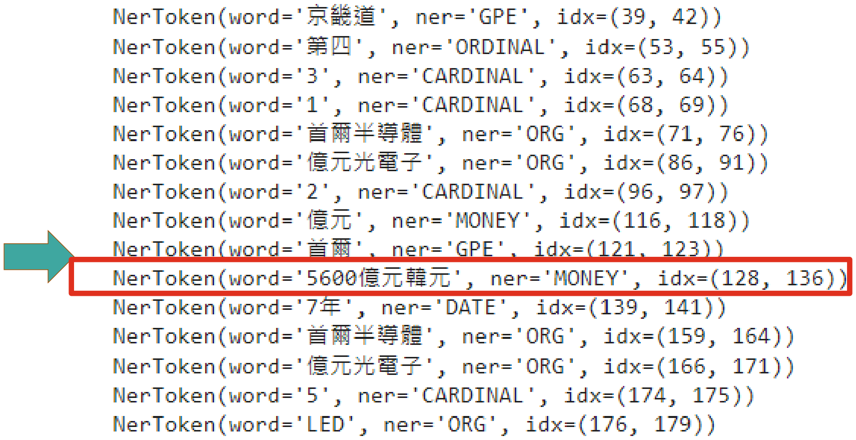
1. The ruling case of the Judicial Yuan is first subjected to text cleaning.
2. The case is segmented into words and tagged with parts of speech.
3. Different segments of words are collected to establish legal text.

Amount Extraction Process



1. 南韓 LED 廠首爾半導體(Seoul_Semiconductor)宣布，南韓京畿道南部地方員警廳國際犯罪第四調查組預計將拘留 3 人，其中 1 名為首爾半導體前常務，離職後任職於億元光電子副總，以及 2 名首爾半導體前雇員；指控他們涉嫌透露給億元光有關首爾半導體耗資 5600 億元韓元、歷時 7 年開發的汽車 LED 技術。2.在此案外，首爾半導體已對億元光電子提起了 5 起 LED 專利侵權訴訟。

Figure 6-4: Text of the original judgment document



```
NerToken(word='京畿道', ner='GPE', idx=(39, 42))
NerToken(word='第四', ner='ORDINAL', idx=(53, 55))
NerToken(word='3', ner='CARDINAL', idx=(63, 64))
NerToken(word='1', ner='CARDINAL', idx=(68, 69))
NerToken(word='首爾半導體', ner='ORG', idx=(71, 76))
NerToken(word='億元光電子', ner='ORG', idx=(86, 91))
NerToken(word='2', ner='CARDINAL', idx=(96, 97))
NerToken(word='億元', ner='MONEY', idx=(116, 118))
NerToken(word='首爾', ner='GPE', idx=(121, 123))
NerToken(word='5600億元韓元', ner='MONEY', idx=(128, 136))
NerToken(word='7年', ner='DATE', idx=(139, 141))
NerToken(word='首爾半導體', ner='ORG', idx=(159, 164))
NerToken(word='億元光電子', ner='ORG', idx=(166, 171))
NerToken(word='5', ner='CARDINAL', idx=(174, 175))
NerToken(word='LED', ner='ORG', idx=(176, 179))
```

Figure 6-5: Extracting the amount from the sentence

Explanation:

1. The text of the Judicial Yuan's ruling document is first subjected to text cleaning.
2. The text is segmented into words, and the amount is extracted and tagged.
3. The Chinese currency amounts are converted into numerical format for ease of subsequent operations.