Georgia Institute of Technology

H. Milton Stewart School of Industrial and Systems Engineering

ISyE 6414-B: Regression Analysis

Final Project Report

Forecasting Stock Returns in the Chinese Market

Team 9

Authors
Lincoln(Po Lin) Lin
Chang Woo Choi
Huilin Fang
Xuesong Yang

December 2, 2024

Contents

1 Introduction **1**
1.1 Background . . . 1
1.2 Objectives . . . 1
1.3 Significance . . . 1

2 Problem Statement **1**
2.1 Possible Challenges . . . 2

3 Data Description **2**
3.1 Dependent Variable . . . 3
3.2 Independent Variable: Part I . . . 3
3.3 Independent Variable: Part II . . . 4

4 Analyses **4**
4.1 Preliminary Data Analysis . . . 5
4.2 Correlation . . . 6
4.3 Modeling . . . 7
4.3.1 Step 1: Preliminary model . . . 7
4.3.2 Step 2: Delete variables without significant contribution . . . 7
4.3.3 Step 3: Transformation based on scatter plot . . . 7
4.3.4 Step 4: Transformation based on outside knowledge . . . 7
4.3.5 Step 5: Seasonality . . . 7
4.3.6 Step 6: Multicollinearity . . . 7
4.3.7 Step 7: Outliers . . . 8
4.3.8 Step 8: Autocorrelation . . . 8
4.3.9 Step 9: Residual Analysis . . . 8
4.3.10 Step 10: Final Model . . . 8

5 Conclusions and Recommendations **8**
5.1 Results . . . 8
5.1.1 Model Fit . . . 8
5.1.2 Predictors . . . 9
5.1.3 Validation . . . 9
5.2 Discussion and Limitation . . . 9
5.3 Conclusion . . . 10

6 Appendix **11**

1 Introduction

1.1 Background

Over the past decades, the Chinese stock market has undergone developments driven by globalization, technological innovation, and increased participation from investors from various industries. These advancements have changed the market to some extent, fostering possibilities for financial modeling and predictive analysis. However, despite its progress, the market still exhibits traits typical of economies, such as regulatory unpredictability, policy-driven market shifts, and heightened sensitivity to investor sentiment.

These unique characteristics of uncertainty complicate forecasting efforts and call for specialized methodologies. Traditional financial theories alone often fall short in addressing the intricate dynamics of the Chinese stock market, necessitating the integration of advanced statistical models and machine learning techniques. This project seeks to leverage these tools to identify key drivers of stock returns and provide accurate predictions of the market. By bridging theoretical and practical perspectives, it tries to answer how macroeconomic factors, market structure, and behavioral influences interact to shape stock performance.

1.2 Objectives

As stated in the background above, this project is designed to fulfill two primary objectives:

1. **Develop an Acceptable Prediction Model:** Utilize advanced statistical methods to create a Multiple Linear Regression model capable of forecasting stock returns with acceptable precision, addressing the unique demands of the market.
2. **Identify Actionable Trends:** Analyze and extract key trends and patterns that drive stock price movements, offering insights to support strategic investment decisions.

1.3 Significance

Forecasting stock returns in the market is crucial for both investors and financial analysts. The market's mix of volatility, regulatory changes, and large number of features makes the task challenging, but the reward is extensive. Reliable predictive models can help investors and analysts make informed decisions for both short-term gains or long-term goals.

The project's goal is to provide tools and insights that help investors navigate the Chinese stock market's complexities, enabling data-driven decisions in a dynamic financial environment.

2 Problem Statement

The primary problem addressed in this project is the development of a predictive model capable of forecasting stock returns in the Chinese stock market. Similar to stock markets in other countries, the Chinese market is characterized by significant uncertainty and structural complexity. Successfully identifying key underlying factors and building a predictive regression model for stock returns requires in-depth knowledge of finance and understanding of the multifaceted variables influencing stock performance. This project will utilize historical stock data from 2021 to construct a regression model that predicts returns.

2.1 Possible Challenges

Building an effective forecasting model for the Chinese stock market comes with inherent challenges, as outlined below:

1. **High Volatility:** Stock prices in the Chinese market are highly volatile, often influenced by sudden regulatory interventions, economic disruptions, or global market conditions. Furthermore, historical data patterns, while useful, may not always reliably predict future movements due to the **unpredictable nature** of the market.
2. **Market-Specific Characteristics:** The Chinese market operates under a unique blend of developed and emerging market features, requiring models to address these hybrid dynamics effectively.

To address these constraints, the model will incorporate mechanisms to adjust for extreme conditions, including policy changes and global economic shocks. Additionally, it will be designed to handle the irregularities and noise inherent in stock market data, ensuring more robust and reliable predictions.

The forecasting model is built upon the following key assumptions:

1. **Cross-Stock Pattern Similarity:** The model assumes that similar behavioral patterns exist across different stocks within the Chinese market, though these patterns may not hold for stock markets outside of China.
2. **Uncertainty in Forecasting:** Acknowledging the inherently unpredictable nature of stock markets, the model focuses on identifying probabilities and trends rather than providing absolute predictions.

These assumptions define the model's scope and limitations, ensuring a realistic and well-grounded approach to forecasting.

3 Data Description

The dataset is sourced from Wind, a stock market data platform used for market analysis in China. Wind aggregates historical data from various sources, making it a credible resource for analysis. The dataset covers daily data from January 1, 2021, to December 31, 2021.

Initially, the dataset consisted of over 1 million rows, but for this project, it was down-sampled to focus on the stocks in the Chinese Securities Index 300 (CSI 300), resulting in 3572 rows. Each row represents a specific stock's performance on a given trading day, including timestamped indicators.

The dataset includes **1 target variable (dependent variable), 14 key features (independent quantitative variables) and 6 independent qualitative variables (5 from feature engineering)**, providing a basis for modeling.

3.1 Dependent Variable

Stock Return (ret): The primary target variable represents the percentage change in a stock's price over a **20-day period**. It is calculated using the formula:

$$\text{Stock Return} = \frac{\text{Price at end of period} - \text{Price at start of period}}{\text{Price at start of period}} \times 100 \tag{3.1}$$

This variable is continuous and quantitative, serving as the output for the predictive model.

3.2 Independent Variable: Part I

14 quantitative variables are defined and calculated as follows:

- **CleverMoneyV1**: Measures the flow of institutional "smart money" into the stock, indicating high-volume trading by professional investors.
- **CORAV1**: A correlation factor that tracks the relationship between stock returns and a reference index over a rolling window.
- **InDayVol**: Total trading volume recorded during the trading day, indicating liquidity.
- **EndAmountPortion**: The percentage of the total trade amount accumulated by the end of the trading day.
- **ACMA**: Adjusted cost metric, reflecting the estimated average acquisition cost while accounting for market anomalies.
- **VSPlow**: A volatility skew metric that evaluates the difference between implied volatilities of lower strike prices, indicating pricing imbalances.
- **IntradayExtremeReturn**: The highest percentage return achieved within a single trading day.
- **IndayRsi**: Relative Strength Index calculated intraday, a momentum indicator that highlights overbought or oversold conditions.
- **IndayHighLow**: Difference between the highest and lowest prices of the stock within a trading day, providing insights into volatility.
- **EarnVolumeRatio**: Ratio of earnings-related trading volume to the total trading volume, reflecting the impact of earnings announcements.
- **VolRebalancePriceVl**: A factor based on volatility rebalancing that adjusts stock prices for different volatility levels.
- **VSPAPP**: Pricing pressure derived from volume-skewed pricing in the stock's trading activity.
- **CDPPV1**: A capital distribution pattern volume metric, tracking long-term distribution trends.
- **HVRealizedSkew**: Historical volatility skewness, measuring asymmetry in return distribution over a given period.

3.3 Independent Variable: Part II

6 independent qualitative variables, 5 derived from feature engineering as follows:

MarketCapIndicator:

- **Purpose:** Classifies stocks into large-cap, mid-cap, and small-cap categories.
- **Definition:** A categorical variable differentiating stocks based on market capitalization: $x > 0.7 \rightarrow \mathbf{3}$, $0.3 < x \leq 0.7 \rightarrow \mathbf{2}$, $x \leq 0.3 \rightarrow \mathbf{1}$.

JumpIndicator:

- **Purpose:** Flags stocks with unusual price movements.
- **Definition:** A binary variable indicating significant price jumps: $x > 0.7 \rightarrow \mathbf{1}$, $x \leq 0.7 \rightarrow \mathbf{0}$.

MorningAfternoonVolatilityIndicator:

- **Purpose:** Highlights volatility differences between trading periods.
- **Definition:** Captures relative volatility between morning and afternoon trading: $x > 0.7 \rightarrow \mathbf{1}$, $x \leq 0.7 \rightarrow \mathbf{0}$.

LiquidityIndicator:

- **Purpose:** Describes how easily a stock can be bought or sold in the market.
- **Definition:** Captures the stock's daily turnover rate: $x > 0.7 \rightarrow \mathbf{1}$, $x \leq 0.7 \rightarrow \mathbf{0}$.

ConsecutiveLimitIndicator:

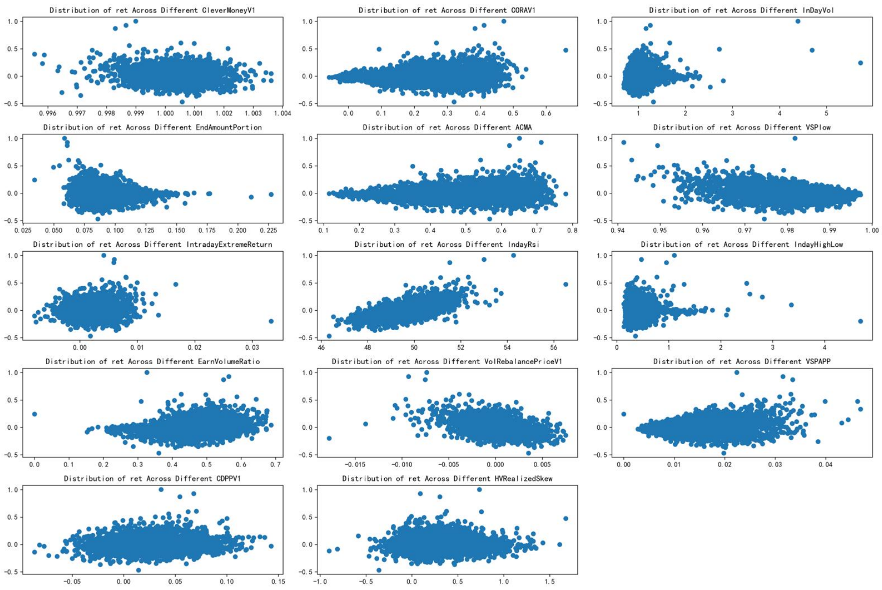
- **Purpose:** Signals extreme price movements under regulatory constraints.
- **Definition:** Indicates consecutive limit prices in one month: $x > 0 \rightarrow \mathbf{1}$, $x \leq 0 \rightarrow \mathbf{0}$.

MonthFactor:

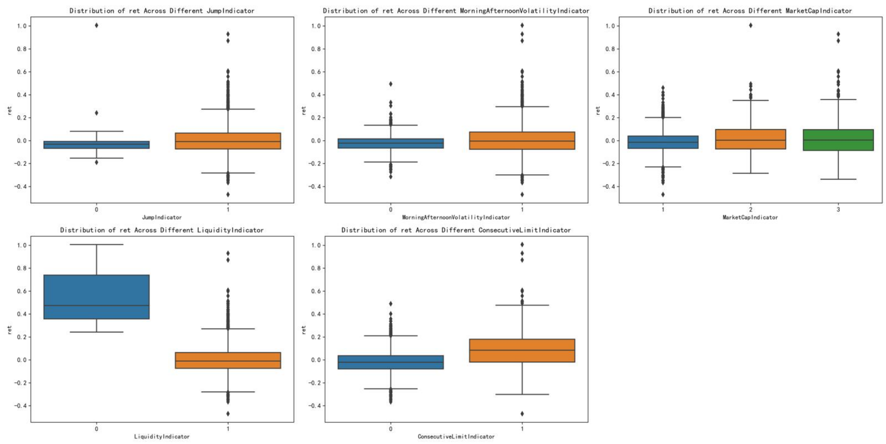
- **Purpose:** Highlights any seasonal trends in stock returns.
- **Definition:** A categorical variable indicating the calendar month of each row in the dataset, with January as the base level.

4 Analyses

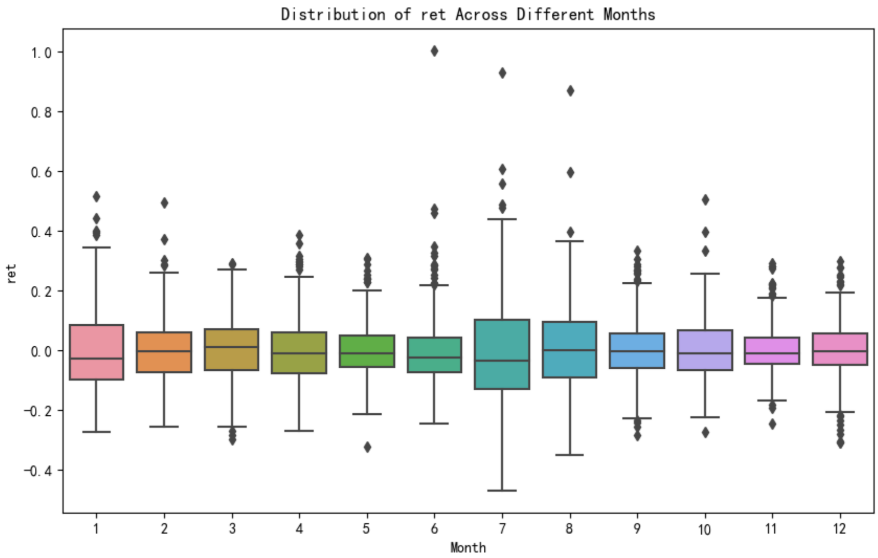
4.1 Preliminary Data Analysis



1. Scatter Plot of ret Across Different Features
Interpretation: Some features show a likely linear or logarithmic relationship with ret while others show a random scattering or no relationship with ret visible to the naked eye.

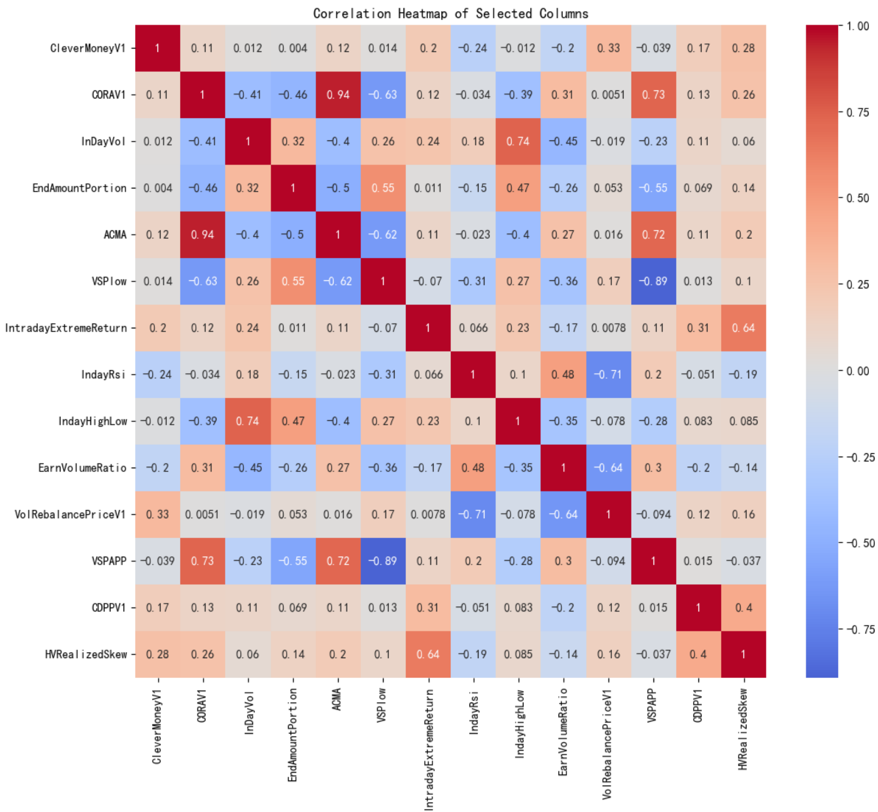


2. Distribution of ret Across Different Indicators
Interpretation: Some indicators show a significant difference between categories visible to the naked eye while others show no easily identifiable difference.



3. Distribution of ret Across Different Months Interpretation: With consideration into the scale for only the median values and interquartile ranges of ret, it is still possible a seasonality trend exists.

4.2 Correlation



Based on the correlation map of the features, there is a clear multicollinearity issue between some of the features. For example, the VSPAPP variable displays high correlation with multiple other features. However, for thorough analysis, all features were included in the preliminary model and the multicollinearity issue was addressed later using Variance

Inflation Factor (VIF).

4.3 Modeling

4.3.1 Step 1: Preliminary model

All predictor, with the exclusion of MonthFactor, were included in the initial model. MonthFactor was excluded so that it can be added later and the effect of seasonality, if any, could easily be identified. A strict stepwise regression approach (forward or backward) was not utilized in favor of incorporating confidential, financial knowledge of the Chinese stock market that was available to the authors.

4.3.2 Step 2: Delete variables without significant contribution

Similar to a backward regression approach, 6 predictors (ACMA, IntradayExtremeReturn, IndayHighLow, HVRealizedSkew, MorningAfternoonVolatilityIndicator, and MarketCapIndicator) with high p-values were removed. The result was only a reduction in R-squared by 0.003, an acceptable outcome.

4.3.3 Step 3: Transformation based on scatter plot

Upon review of the scatter plots of ret across different features, transformations of 3 features were obvious candidates. A logarithmic transformation of InDayVol and EndAmountPortion was completed and a squared transformation of IndayRsi was completed. The result was an increase in R-squared by 0.049.

4.3.4 Step 4: Transformation based on outside knowledge

Based on confidential, financial knowledge of the Chinese stock market that was available to the authors, a logarithmic transformation of ACMA and a squared transformation of VolRebalancePriceV1 were recommended to the authors and completed. However, the result was only an increase in R-squared by 0.002. The scatter plots of ret across those 2 features do not support such transformations, and thus, the outside knowledge may not be very accurate in this dataset.

4.3.5 Step 5: Seasonality

MonthFactor was added to assess seasonality. All months, except two, had statistically significant p-values (>0.05). The result was an increase in R-squared by 0.028.

4.3.6 Step 6: Multicollinearity

VIF was calculated for each predictor. ACMA_log and CORAV1 had a very high VIF (>10) and VSPAPP had a moderately high VIF (>5). Considering the VIF as well as the correlation map, ACMA_log and VSPAPP were selected to be removed from the model. VIF was calculated again for each predictor in the model after the removal, and no predictor had a VIF>5.

4.3.7 Step 7: Outliers

The presence of outliers was tested via measuring the Cook's Distance for each observation in the dataset and comparing to a threshold of 0.001 (roughly 4/n). Because only about 3% of the observations exceeded the threshold, the outliers were not removed from the dataset. Furthermore, the presence of outliers is necessary to reflect the high volatility of the stock market.

4.3.8 Step 8: Autocorrelation

To test for positive or negative autocorrelation between residuals, the Durbin-Watson statistic was calculated and was found to be 1.91. Because of the large n (3572), an exact upper and lower bounds were not available but are expected to converge around 2. Therefore, the authors accepted with reservation that there is not strong positive or negative autocorrelation present in the model.

4.3.9 Step 9: Residual Analysis

Upon review of the residuals vs. fitted values plot, there was a random scattering around a mean of zero, indicating no violation of the homoscedasticity assumption. Upon review of the histogram of residuals plot, there was conformity to a bell-shaped curve or a normal distribution, indicating no violation of the normality assumption. However, upon review of the Q-Q plot, there was some deviation in the extremes at both ends. This is likely due to the presence of outliers in the dataset and suggests some mild violation of the normality assumption.

4.3.10 Step 10: Final Model

After removing the predictors ACMA_log and VSPAP due to their multicollinearity issue, the final model was constructed with an R-squared of 0.538.

5 Conclusions and Recommendations

5.1 Results

5.1.1 Model Fit

The model's R-squared (0.538) indicates that 53.8% of the variability in the dependent variable (return) is explained by the predictors, suggesting a moderate level of explanation with room for improvement, as 46.2% of the variability remains unexplained due to possible omitted variables, nonlinear relationships, or randomness.

The Adjusted R-squared (0.534), which accounts for the number of predictors, is slightly lower, highlighting potential overfitting or the inclusion of unnecessary variables that do not significantly improve the model's fit.

The F-statistic (158.5, p-value = 0.000) confirms that the model as a whole is statistically significant, indicating that at least one predictor significantly explains the variability in return.

5.1.2 Predictors

The model indicates that CleverMoneyV1 has the greatest positive impact with a coefficient of 1.8810, while VolRebalancePriceV1 has the greatest negative impact with a coefficient of -11.1166. This means:

- A one unit increase in CleverMoneyV1 increases the return by 1.881 units, holding other variables constant. This predictor reflects "smart money" movements or strategic investments by well-informed traders. A large positive coefficient makes logical sense as well-informed traders are expected to show profitable investment patterns.
- A one unit increase in VolRebalancePriceV1 decreases the return by 11.1166 units, holding other factors constant. This predictor represents price rebalancing volumes based on market volatility levels. A large negative coefficient makes logical sense as large price corrections could be indicative of possible market inefficiencies.
- Also, other predictors, such as EarnVolumeRatio, IndayVol_log and so forth, show some extent of significant impact on return, but less influential than CleverMoneyV1 and VolRebalancePriceV1.
- The intercept has a coefficient of 1.2315 but is not statistically significant (p-value = 0.530), which means the baseline prediction of return when all predictors are zero is not meaningful.
- The monthly indicators (February to December) reveal significant and positive seasonality effects on returns for most months, except for May and August, which show no significant impact. This suggests that returns exhibit systematic seasonal patterns throughout the year, with certain months having stronger influences, while others, like May and August, display weaker or negligible seasonal effects.

5.1.3 Validation

The same analysis and modeling process was applied to stock data one year later (from 2022). The final model's R-squared differed from the 2021 data by only 0.007.

5.2 Discussion and Limitation

Stock market prediction remains inherently challenging due to high uncertainty and susceptibility to unforeseen events. This project operates under the following assumptions:

- Data points exhibit similar or identical patterns across different stocks, which may not hold for stock markets outside of China.
- Stock market prediction is inherently challenging due to high uncertainty.

Although the final model presented here has a lower R-squared than traditional regression models in other fields, the final R-squared of 0.538 is acceptable due to the inherent uncertainty of the stock market and the value is comparable to other models used in the industry that account for macroeconomic factors.

However, additional diagnostic metrics highlighted an issue with the normality assumption, a key component of any regression model.

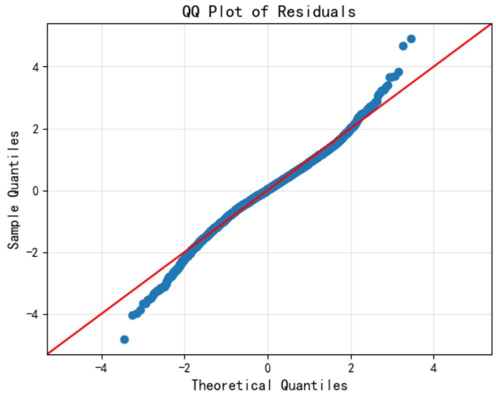


Figure 5.1: QQ Plot

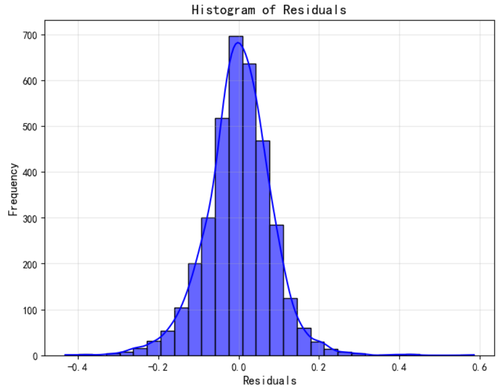


Figure 5.2: Histogram

The residuals are slightly left-skewed (-0.062) and have a high kurtosis (5.435), meaning they are more peaked and have heavier tails than a normal distribution. Furthermore, both the Omnibus test (210.581, p = 0.000) and Jarque-Bera test (884.408, p = 0.000) again show that the residuals are not normally distributed. These issues suggest that the model's p-values and confidence intervals might not be entirely reliable, which could lead to misleading conclusions about the predictors' significance. On the positive side, the Durbin-Watson statistic (1.91) suggested little to no autocorrelation in the residuals, indicating that errors are indeed independent, which is a good news to the model.

5.3 Conclusion

Using historical Chinese stock market data from 2021 and a diverse set of 20 features and variables, a multiple linear regression model to predict stock returns was developed. A modified backward regression approach was used to include transformations of certain predictors and confidential, outside knowledge. Residual and correlation analysis was performed to adjust the final model.

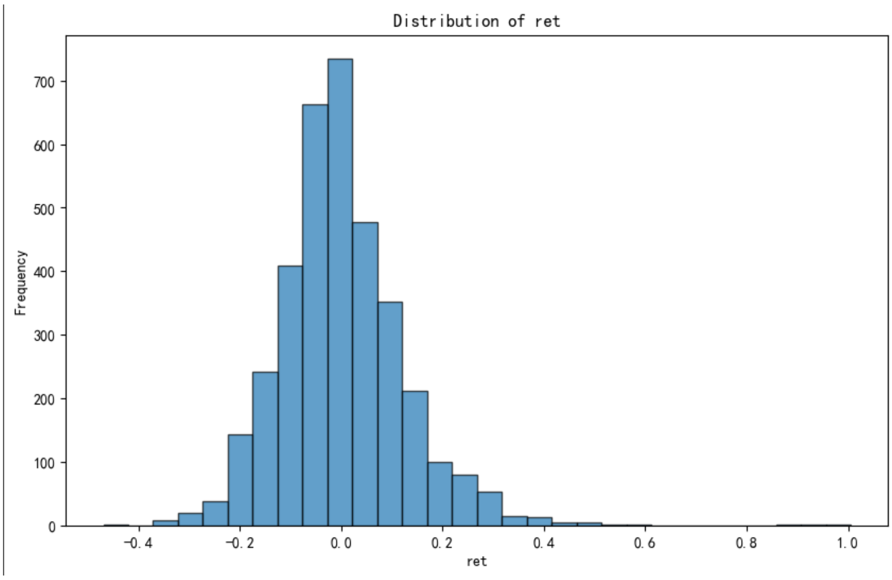
Potential improvements to this study include conducting a deeper literature review to better align methods with established practices and incorporating hypotheses to guide evaluation frameworks. Enhanced interpretation of how each predictor influences stock returns would provide stronger insights. Additionally, exploring diverse feature selection techniques could improve performance, ensuring a robust and adaptable forecasting approach.

6 Appendix

	index	date	code	IntroClain	CleverMoneyV1	CleverMoneyV2	CORAV1	CORAV2	InDayVol	EndAmountPortion	...	HVRealizedSkew	IDhf	RTVV1	VCVV1	ret
0	2646913	2021-01-04	000001.SZ	-2.307922e-08	1.001415	0.998645	0.388472	0.403585	0.771950	0.067540	...	-0.406428	-0.062500	0.000020	0.771950	-0.017704
1	2646914	2021-01-04	000002.SZ	-6.410019e-25	1.001249	0.998955	0.315000	0.287930	0.653615	0.069361	...	-0.065703	-0.016667	0.000007	0.653615	0.000850
46	2646959	2021-01-04	000063.SZ	-1.572439e-07	1.004285	0.993326	0.631008	0.418536	1.020093	0.133937	...	1.112980	-0.020833	0.000036	1.020093	0.023990
48	2646961	2021-01-04	000066.SZ	-2.987916e-21	0.998321	1.015908	0.378624	0.192933	1.785623	0.063742	...	0.411036	-0.020833	0.000171	1.785623	0.020562
50	2646963	2021-01-04	000069.SZ	-9.805961e-07	1.001192	0.998321	0.033850	-0.044665	1.003282	0.080880	...	-0.060512	-0.025000	0.000015	1.003282	-0.003101
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3408270	6055183	2023-12-29	688363.SH	-4.994732e-06	0.997821	1.013145	0.264262	0.305864	1.182238	0.110618	...	0.343820	-0.066667	0.000019	1.182238	-0.011713
3408299	6055212	2023-12-29	688396.SH	-3.515678e-06	0.998415	1.004779	0.119409	0.130016	0.688588	0.089914	...	0.636949	0.050000	0.000011	0.688588	-0.028737
3408387	6055300	2023-12-29	688561.SH	-7.929759e-07	0.997959	1.006442	0.250577	0.444521	1.018495	0.060102	...	-0.670061	-0.025000	0.000023	1.018495	-0.033881
3408418	6055331	2023-12-29	688599.SH	-2.737843e-13	0.999824	1.001942	0.265656	0.273097	1.209220	0.072041	...	-0.220050	-0.008333	0.000090	1.209220	-0.027671
3408519	6055432	2023-12-29	688981.SH	-3.454954e-08	0.999416	1.001617	0.189005	0.102018	0.768916	0.079691	...	0.377975	-0.012500	0.000011	0.768916	-0.003206

217064 rows × 40 columns

1. Snapshot of the raw dataset prior to down-sampling.



2. Distribution of the dependent variable, ret.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ret   R-squared:                       0.484
Model:                            OLS   Adj. R-squared:                  0.481
Method:                 Least Squares   F-statistic:                     158.5
Date:                Sat, 23 Nov 2024   Prob (F-statistic):               0.00
Time:                        11:19:47   Log-Likelihood:                 3635.7
No. Observations:                3572   AIC:                            -7227.
Df Residuals:                    3550   BIC:                            -7091.
Df Model:                          21
Covariance Type:            nonrobust
===========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                    17.0133      2.183      7.793      0.000      12.733      21.294
CleverMoneyV1                            -8.9125      2.093     -4.258      0.000     -13.017      -4.808
CORAV1                                    0.1704      0.045      3.787      0.000       0.082       0.259
InDayVol                                  0.0912      0.012      7.489      0.000       0.067       0.115
EndAmountPortion                         -0.4494      0.148     -3.039      0.002      -0.739      -0.159
ACMA                                     -0.0101      0.041     -0.249      0.804      -0.090       0.070
VSPlow                                   -9.5493      0.408    -23.425      0.000     -10.349      -8.750
IntradayExtremeReturn                     2.2778      0.931      2.447      0.014       0.452       4.103
IndayRsi                                  0.0381      0.002     23.462      0.000       0.035       0.041
IndayHighLow                             -0.0263      0.012     -2.244      0.025      -0.049      -0.003
EarnVolumeRatio                          -0.2141      0.034     -6.262      0.000      -0.281      -0.147
VolRebalancePriceV1                     -13.7068      1.173    -11.682      0.000     -16.007     -11.406
VSPAPP                                  -10.8656      0.658    -16.509      0.000     -12.156      -9.575
CDPPV1                                    0.5764      0.057     10.095      0.000       0.464       0.688
HVRealizedSkew                           -0.0028      0.009     -0.300      0.764      -0.021       0.015
JumpIndicator_ogn                        -0.0641      0.033     -1.968      0.049      -0.128      -0.000
JumpIndicator_1                          -0.0027      0.016     -0.169      0.866      -0.034       0.029
MorningAfternoonVolatilityIndicator_1     0.0077      0.005      1.614      0.107      -0.002       0.017
MarketCapIndicator_2                      0.0080      0.005      1.615      0.106      -0.002       0.018
MarketCapIndicator_3                     -0.0029      0.005     -0.586      0.558      -0.012       0.007
LiquidityIndicator_1                     -0.4071      0.070     -5.803      0.000      -0.545      -0.270
ConsecutiveLimitIndicator_1               0.0147      0.005      3.066      0.002       0.005       0.024
==============================================================================
Omnibus:                      356.397   Durbin-Watson:                   1.874
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2798.653
Skew:                          -0.024   Prob(JB):                         0.00
Kurtosis:                       7.336   Cond. No.                     1.03e+05
==============================================================================
```

3. Step 1 - Preliminary regression model summary.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ret   R-squared:                       0.481
Model:                            OLS   Adj. R-squared:                  0.479
Method:                 Least Squares   F-statistic:                     235.3
Date:                Sat, 23 Nov 2024   Prob (F-statistic):               0.00
Time:                        11:20:24   Log-Likelihood:                 3625.0
No. Observations:                3572   AIC:                            -7220.
Df Residuals:                    3557   BIC:                            -7127.
Df Model:                          14
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          15.3426      2.132      7.198      0.000      11.163      19.522
CleverMoneyV1                  -7.3982      2.047     -3.615      0.000     -11.411      -3.386
CORAV1                          0.1703      0.023      7.468      0.000       0.126       0.215
InDayVol                        0.0805      0.010      8.299      0.000       0.062       0.100
EndAmountPortion               -0.6073      0.131     -4.651      0.000      -0.863      -0.351
VSPlow                         -9.3934      0.403    -23.319      0.000     -10.183      -8.604
IndayRsi                        0.0390      0.002     24.545      0.000       0.036       0.042
EarnVolumeRatio                -0.2192      0.034     -6.542      0.000      -0.285      -0.153
VolRebalancePriceV1           -13.6112      1.145    -11.889      0.000     -15.856     -11.367
VSPAPP                        -10.6850      0.649    -16.457      0.000     -11.958      -9.412
CDPPV1                          0.6117      0.054     11.362      0.000       0.506       0.717
JumpIndicator_ogn              -0.0558      0.032     -1.720      0.085      -0.119       0.008
JumpIndicator_1                -0.0042      0.016     -0.264      0.792      -0.036       0.027
LiquidityIndicator_1           -0.4362      0.068     -6.390      0.000      -0.570      -0.302
ConsecutiveLimitIndicator_1     0.0150      0.005      3.120      0.002       0.006       0.024
==============================================================================
Omnibus:                      352.522   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2730.754
Skew:                          -0.008   Prob(JB):                         0.00
Kurtosis:                       7.283   Cond. No.                     1.00e+05
==============================================================================
```

4. Step 2 - Regression model summary (after removal of 6 predictors).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ret   R-squared:                       0.530
Model:                            OLS   Adj. R-squared:                  0.528
Method:                 Least Squares   F-statistic:                     267.8
Date:                Sat, 23 Nov 2024   Prob (F-statistic):               0.00
Time:                        11:20:27   Log-Likelihood:                 3804.4
No. Observations:                3572   AIC:                            -7577.
Df Residuals:                    3556   BIC:                            -7478.
Df Model:                          15
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                        10.6674      2.049      5.205      0.000       6.649      14.686
InDayVol_log                  0.0455      0.022      2.076      0.038       0.003       0.088
EndAmountPortion_log         -0.2151      0.139     -1.549      0.121      -0.487       0.057
CleverMoneyV1                -4.4020      1.956     -2.250      0.024      -8.237      -0.567
IndayRsi_squ                  0.0015   7.56e-05     19.453      0.000       0.001       0.002
CORAV1                        0.2112      0.022      9.714      0.000       0.169       0.254
VSPlow                       -6.4578      0.414    -15.601      0.000      -7.269      -5.646
IndayRsi                     -0.0660      0.006    -11.840      0.000      -0.077      -0.055
EarnVolumeRatio              -0.3155      0.032     -9.809      0.000      -0.379      -0.252
VolRebalancePriceV1          -5.9714      1.155     -5.171      0.000      -8.235      -3.707
VSPAPP                       -7.8089      0.637    -12.268      0.000      -9.057      -6.561
CDPPV1                        0.5494      0.051     10.702      0.000       0.449       0.650
JumpIndicator_ogn            -0.0710      0.031     -2.298      0.022      -0.132      -0.010
LiquidityIndicator_1         -0.0392      0.065     -0.608      0.543      -0.166       0.087
JumpIndicator_1               0.0033      0.015      0.219      0.827      -0.026       0.033
ConsecutiveLimitIndicator_1   0.0132      0.005      2.899      0.004       0.004       0.022
==============================================================================
Omnibus:                      203.981   Durbin-Watson:                   1.824
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              831.935
Skew:                           0.061   Prob(JB):                    2.23e-181
Kurtosis:                       5.361   Cond. No.                     5.02e+06
==============================================================================
```

5. Step 3 - Regression model summary (transformation based on scatter plot).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ret   R-squared:                       0.532
Model:                            OLS   Adj. R-squared:                  0.530
Method:                 Least Squares   F-statistic:                     237.6
Date:                Sat, 23 Nov 2024   Prob (F-statistic):               0.00
Time:                        11:20:31   Log-Likelihood:                 3810.4
No. Observations:                3572   AIC:                            -7585.
Df Residuals:                    3554   BIC:                            -7473.
Df Model:                          17
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                        10.3674      2.063      5.025      0.000       6.323      14.412
ACMA_log                     -0.1388      0.050     -2.771      0.006      -0.237      -0.041
VolRebalancePriceV1_squ     2.772e+04   1.39e+04     1.995      0.046     480.281     5.5e+04
InDayVol_log                  0.0327      0.022      1.459      0.145      -0.011       0.077
EndAmountPortion_log         -0.2792      0.140     -1.988      0.047      -0.555      -0.004
CleverMoneyV1                -4.0789      1.969     -2.072      0.038      -7.939      -0.218
IndayRsi_squ                  0.0015   7.77e-05     19.064      0.000       0.001       0.002
CORAV1                        0.3067      0.040      7.607      0.000       0.228       0.386
VSPlow                       -6.3908      0.415    -15.384      0.000      -7.205      -5.576
IndayRsi                     -0.0674      0.006    -11.802      0.000      -0.079      -0.056
EarnVolumeRatio              -0.3445      0.033    -10.338      0.000      -0.410      -0.279
VolRebalancePriceV1          -7.5801      1.363     -5.561      0.000     -10.253      -4.907
VSPAPP                       -7.5137      0.642    -11.700      0.000      -8.773      -6.255
CDPPV1                        0.5508      0.051     10.743      0.000       0.450       0.651
JumpIndicator_ogn            -0.0770      0.031     -2.486      0.013      -0.138      -0.016
LiquidityIndicator_1         -0.0235      0.065     -0.363      0.716      -0.150       0.103
JumpIndicator_1               0.0047      0.015      0.306      0.760      -0.025       0.034
ConsecutiveLimitIndicator_1   0.0124      0.005      2.723      0.006       0.003       0.021
==============================================================================
Omnibus:                      205.650   Durbin-Watson:                   1.824
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              836.787
Skew:                           0.076   Prob(JB):                    1.97e-182
Kurtosis:                       5.366   Cond. No.                     2.46e+10
==============================================================================
```

6. Step 4 - Regression model summary (transformation based on outside knowledge).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ret   R-squared:                       0.560
Model:                            OLS   Adj. R-squared:                  0.556
Method:                 Least Squares   F-statistic:                     160.8
Date:                Sat, 23 Nov 2024   Prob (F-statistic):               0.00
Time:                        11:20:34   Log-Likelihood:                 3919.4
No. Observations:                3572   AIC:                            -7781.
Df Residuals:                    3543   BIC:                            -7602.
Df Model:                          28
Covariance Type:            nonrobust
=================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                           9.8871      2.046      4.831      0.000       5.875      13.899
ACMA_log                       -0.1113      0.049     -2.277      0.023      -0.207      -0.015
VolRebalancePriceV1_squ      3.944e+04   1.36e+04      2.902      0.004    1.28e+04    6.61e+04
InDayVol_log                    0.0228      0.022      1.031      0.303      -0.021       0.066
EndAmountPortion_log           -0.3271      0.139     -2.348      0.019      -0.600      -0.054
CleverMoneyV1                  -3.2842      1.953     -1.682      0.093      -7.113       0.545
IndayRsi_squ                    0.0015   7.68e-05     20.095      0.000       0.001       0.002
CORAV1                          0.2938      0.040      7.409      0.000       0.216       0.372
VSPlow                         -6.7018      0.408    -16.436      0.000      -7.501      -5.902
IndayRsi                       -0.0715      0.006    -12.741      0.000      -0.082      -0.060
EarnVolumeRatio                -0.3576      0.033    -10.728      0.000      -0.423      -0.292
VolRebalancePriceV1            -9.8839      1.389     -7.116      0.000     -12.607      -7.160
VSPAPP                         -8.0176      0.629    -12.740      0.000      -9.252      -6.784
CDPPV1                          0.5649      0.051     11.140      0.000       0.465       0.664
JumpIndicator_ogn              -0.0659      0.030     -2.185      0.029      -0.125      -0.007
LiquidityIndicator_1           -0.0022      0.063     -0.034      0.973      -0.126       0.122
JumpIndicator_1                 0.0040      0.015      0.270      0.787      -0.025       0.033
ConsecutiveLimitIndicator_1     0.0131      0.004      2.929      0.003       0.004       0.022
month_factor_2                  0.0246      0.007      3.542      0.000       0.011       0.038
month_factor_3                  0.0487      0.007      7.044      0.000       0.035       0.062
month_factor_4                  0.0160      0.007      2.299      0.022       0.002       0.030
month_factor_5                 -0.0146      0.007     -2.117      0.034      -0.028      -0.001
month_factor_6                  0.0403      0.007      5.912      0.000       0.027       0.054
month_factor_7                  0.0568      0.007      8.381      0.000       0.043       0.070
month_factor_8                  0.0040      0.007      0.592      0.554      -0.009       0.017
month_factor_9                  0.0114      0.007      1.682      0.093      -0.002       0.025
month_factor_10                 0.0362      0.007      5.311      0.000       0.023       0.050
month_factor_11                 0.0399      0.007      5.789      0.000       0.026       0.053
month_factor_12                 0.0427      0.007      6.126      0.000       0.029       0.056
==============================================================================
Omnibus:                      217.327   Durbin-Watson:                   1.925
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              957.668
Skew:                           0.002   Prob(JB):                    1.11e-208
Kurtosis:                       5.537   Cond. No.                     2.48e+10
==============================================================================
```

7. Step 5 - Regression model summary (seasonality).

	Feature	VIF
0	const	2.236322e+06
1	ACMA_log	1.014065e+01
2	VolRebalancePriceV1_squ	1.573089e+00
3	InDayVol_log	2.401442e+00
4	EndAmountPortion_log	1.776113e+00
5	CleverMoneyV1	1.300611e+00
6	IndayRsi_squ	1.008371e+00
7	CORAV1	1.045809e+01
8	VSPlow	6.321363e+00
9	IndayRsi	1.830605e+00
10	EarnVolumeRatio	3.683209e+00
11	VolRebalancePriceV1	4.264131e+00
12	VSPAPP	8.302695e+00
13	CDPPV1	1.185976e+00
14	JumpIndicator_ogn	3.298077e+00
15	LiquidityIndicator_1	1.520165e+00
16	JumpIndicator_1	2.825774e+00
17	ConsecutiveLimitIndicator_1	1.670098e+00
18	month_factor_2	1.976893e+00
19	month_factor_3	1.970092e+00
20	month_factor_4	1.995950e+00
21	month_factor_5	1.967730e+00
22	month_factor_6	1.928267e+00
23	month_factor_7	1.904073e+00
24	month_factor_8	1.883515e+00
25	month_factor_9	1.902009e+00
26	month_factor_10	1.934738e+00
27	month_factor_11	1.977518e+00
28	month_factor_12	2.014689e+00

8. VIF analysis (after Step 5).

	Feature	VIF
0	const	1.988450e+06
1	VolRebalancePriceV1_squ	1.615412e+00
2	InDayVol_log	2.299526e+00
3	EndAmountPortion_log	1.731196e+00
4	CleverMoneyV1	1.233649e+00
5	IndayRsi_squ	2.864153e+01
6	CORAV1	2.305276e+00
7	VSPlow	3.375835e+00
8	IndayRsi	2.328386e+01
9	EarnVolumeRatio	3.616638e+00
10	VolRebalancePriceV1	4.953930e+00
11	CDPPV1	1.190085e+00
12	JumpIndicator_ogn	3.232333e+00
13	LiquidityIndicator_1	1.803774e+00
14	JumpIndicator_1	2.810311e+00
15	ConsecutiveLimitIndicator_1	1.656734e+00
16	month_factor_2	1.975670e+00
17	month_factor_3	1.969185e+00
18	month_factor_4	1.984942e+00
19	month_factor_5	1.961966e+00
20	month_factor_6	1.918189e+00
21	month_factor_7	1.902093e+00
22	month_factor_8	1.876907e+00
23	month_factor_9	1.889800e+00
24	month_factor_10	1.923150e+00
25	month_factor_11	1.959558e+00
26	month_factor_12	2.000648e+00

9. VIF analysis (after removal of ACMA_log and VSPAPP).

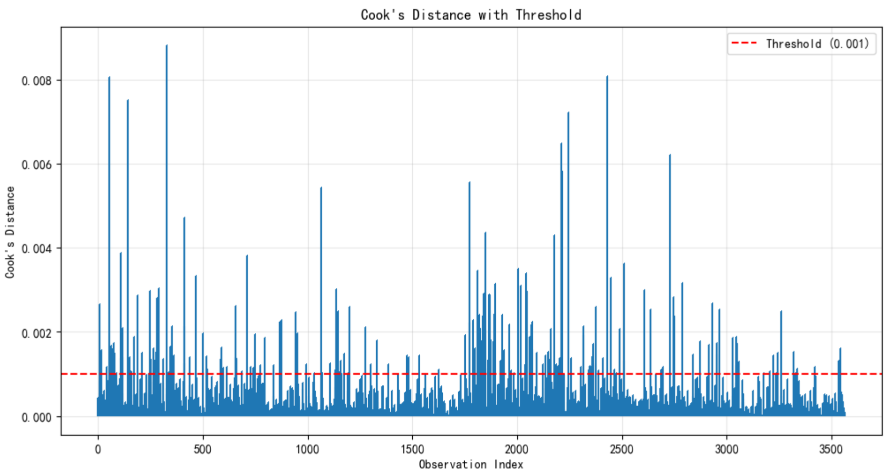
OLS Regression Results

Dep. Variable:	ret	R-squared:	0.538
Model:	OLS	Adj. R-squared:	0.534
Method:	Least Squares	F-statistic:	158.5
Date:	Sat, 23 Nov 2024	Prob (F-statistic):	0.00
Time:	14:33:01	Log-Likelihood:	3831.8
No. Observations:	3572	AIC:	-7610.
Df Residuals:	3545	BIC:	-7443.
Df Model:	26		
Covariance Type:	nonrobust		

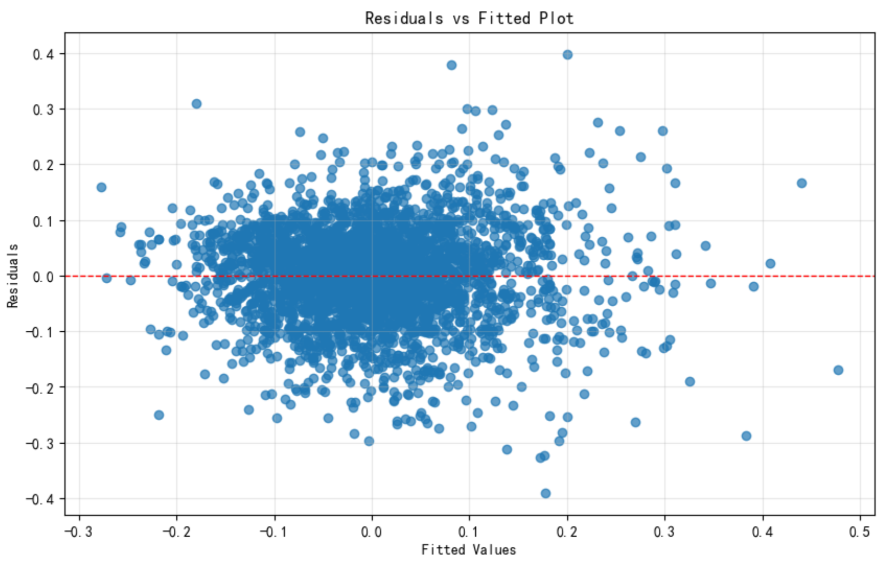
	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2315	1.960	0.628	0.530	-2.612	5.075
VolRebalancePriceV1_squ	4.604e+04	1.39e+04	3.309	0.001	1.88e+04	7.33e+04
InDayVol_log	-0.0580	0.021	-2.773	0.006	-0.099	-0.017
EndAmountPortion_log	0.0231	0.139	0.166	0.868	-0.250	0.296
CleverMoneyV1	1.8810	1.941	0.969	0.332	-1.924	5.686
IndayRsi_squ	0.0017	7.6e-05	22.605	0.000	0.002	0.002
CORAV1	0.0715	0.019	3.821	0.000	0.035	0.108
VSPlow	-2.8864	0.284	-10.155	0.000	-3.444	-2.329
IndayRsi	-0.0856	0.006	-15.556	0.000	-0.096	-0.075
EarnVolumeRatio	-0.3818	0.034	-11.346	0.000	-0.448	-0.316
VolRebalancePriceV1	-11.1166	1.419	-7.836	0.000	-13.898	-8.335
CDPPV1	0.5919	0.052	11.402	0.000	0.490	0.694
JumpIndicator_ogn	-0.1205	0.031	-3.939	0.000	-0.181	-0.061
LiquidityIndicator_1	-0.0189	0.064	-0.294	0.769	-0.145	0.107
JumpIndicator_1	0.0179	0.015	1.183	0.237	-0.012	0.048
ConsecutiveLimitIndicator_1	0.0088	0.005	1.932	0.053	-0.000	0.018
month_factor_2	0.0284	0.007	4.005	0.000	0.014	0.042
month_factor_3	0.0518	0.007	7.322	0.000	0.038	0.066
month_factor_4	0.0247	0.007	3.478	0.001	0.011	0.039
month_factor_5	-0.0059	0.007	-0.839	0.402	-0.020	0.008
month_factor_6	0.0448	0.007	6.435	0.000	0.031	0.058
month_factor_7	0.0575	0.007	8.298	0.000	0.044	0.071
month_factor_8	0.0093	0.007	1.357	0.175	-0.004	0.023
month_factor_9	0.0184	0.007	2.664	0.008	0.005	0.032
month_factor_10	0.0405	0.007	5.822	0.000	0.027	0.054
month_factor_11	0.0469	0.007	6.677	0.000	0.033	0.061
month_factor_12	0.0509	0.007	7.173	0.000	0.037	0.065

Omnibus:	210.581	Durbin-Watson:	1.910
Prob(Omnibus):	0.000	Jarque-Bera (JB):	884.408
Skew:	-0.062	Prob(JB):	8.98e-193
Kurtosis:	5.435	Cond. No.	2.47e+10

10. Step 10 - Regression model summary (final).



11. Cooks distance analysis of all observations.



12. Plot of residuals vs fitted values.

```
[ ] from statsmodels.stats.stattools import durbin_watson
    residuals = model.resid
    dw_stat = durbin_watson(residuals)
    print(f"Durbin-Watson statistic: {dw_stat}")

    Durbin-Watson statistic: 1.9098715018285335
```

13. Durbin-Watson statistic calculation.

OLS Regression Results

Dep. Variable:	ret	R-squared:	0.545
Model:	OLS	Adj. R-squared:	0.541
Method:	Least Squares	F-statistic:	163.1
Date:	Sat, 23 Nov 2024	Prob (F-statistic):	0.00
Time:	11:36:08	Log-Likelihood:	4444.1
No. Observations:	3295	AIC:	-8838.
Df Residuals:	3270	BIC:	-8686.
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4049	2.783	-0.145	0.884	-5.862	5.052
VolRebalancePriceV1_squ	-8377.0977	3.82e+04	-0.219	0.827	-8.34e+04	6.66e+04
InDayVol_log	-0.0576	0.016	-3.613	0.000	-0.089	-0.026
EndAmountPortion_log	0.1919	0.112	1.720	0.085	-0.027	0.411
CleverMoneyV1	-1.1263	2.092	-0.538	0.590	-5.229	2.976
IndayRsi_squ	-0.0011	0.001	-1.375	0.169	-0.003	0.000
CORAV1	0.0564	0.017	3.367	0.001	0.024	0.089
VSPlow	-4.3425	0.287	-15.118	0.000	-4.906	-3.779
IndayRsi	0.1775	0.079	2.233	0.026	0.022	0.333
EarnVolumeRatio	-0.3611	0.028	-12.862	0.000	-0.416	-0.306
VolRebalancePriceV1	-13.6111	1.527	-8.911	0.000	-16.606	-10.616
CDPPV1	0.4220	0.043	9.778	0.000	0.337	0.507
JumpIndicator_ogn	-0.1402	0.021	-6.645	0.000	-0.182	-0.099
JumpIndicator_1	0.0213	0.011	1.996	0.046	0.000	0.042
ConsecutiveLimitIndicator_1	0.0024	0.004	0.582	0.561	-0.006	0.010
month_factor_3	0.0355	0.006	6.162	0.000	0.024	0.047
month_factor_4	-0.0039	0.006	-0.645	0.519	-0.016	0.008
month_factor_5	-0.0343	0.005	-6.459	0.000	-0.045	-0.024
month_factor_6	-0.0685	0.006	-12.371	0.000	-0.079	-0.058
month_factor_7	0.0545	0.006	9.506	0.000	0.043	0.066
month_factor_8	0.0108	0.005	2.028	0.043	0.000	0.021
month_factor_9	0.0668	0.006	11.360	0.000	0.055	0.078
month_factor_10	0.0281	0.006	4.728	0.000	0.016	0.040
month_factor_11	-0.0456	0.005	-8.537	0.000	-0.056	-0.035
month_factor_12	0.0403	0.006	7.205	0.000	0.029	0.051

Omnibus:	346.017	Durbin-Watson:	1.896
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1043.723
Skew:	-0.548	Prob(JB):	2.28e-227
Kurtosis:	5.530	Cond. No.	8.55e+10

14. Regression model summary (validation on stock market data from 2022).