**Equity Selection Model**

**1. Initial Assessment:** Do you believe this data to be equally predictive of all stocks? Provide a brief justification for your answer

The data is likely **not equally predictive for all stocks**. **Justification:***a) Sectoral Differences*: Stocks in different sectors, like technology vs. utilities, have unique drivers and risks, making features vary in predictiveness across sectors.

*b) Market Capitalization*: Large-cap stocks may be more influenced by macroeconomic factors, while small-caps are sensitive to company-specific events, affecting the predictive power of features like market cap and volatility.

*c) Index Membership*: Index inclusion impacts stock behavior due to factors like liquidity and institutional investment, making the index membership feature predictive.

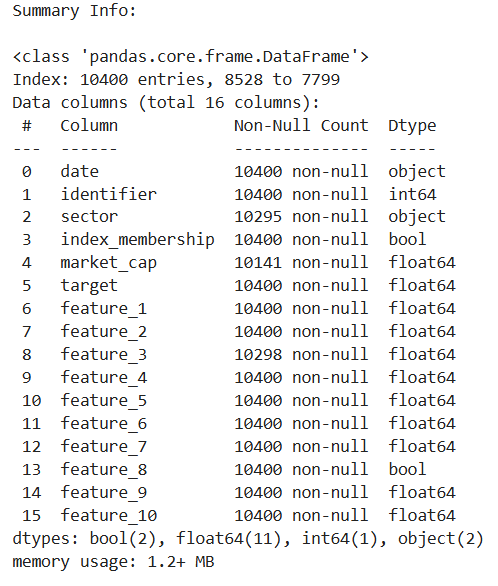
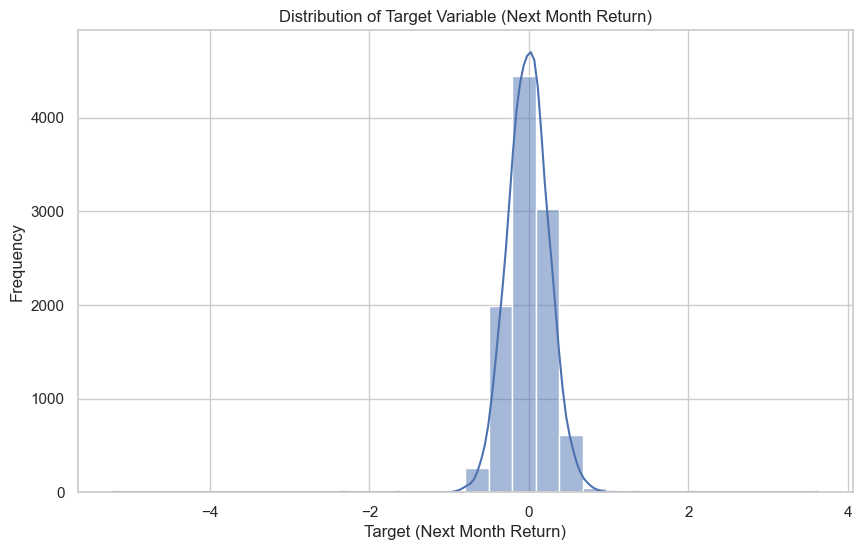
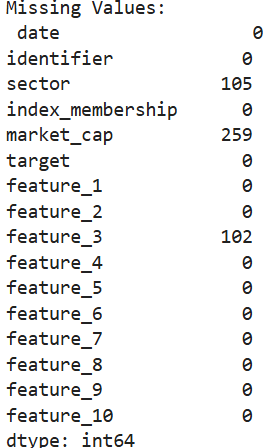
*d) Feature Relevance*: The 10 firm-specific features, based on price, fundamentals, and alternative data, may vary in relevance depending on stock characteristics, such as growth vs. value.

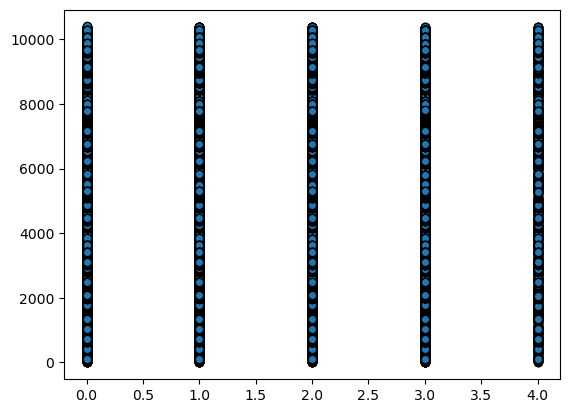
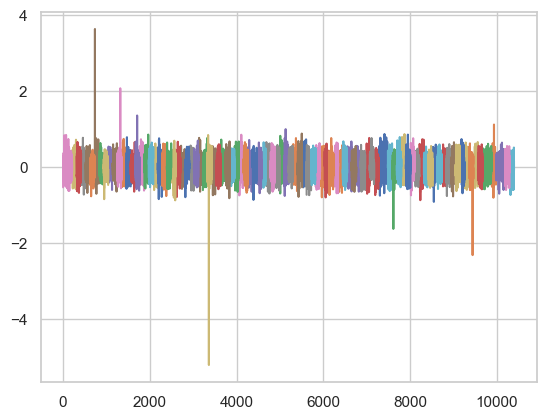
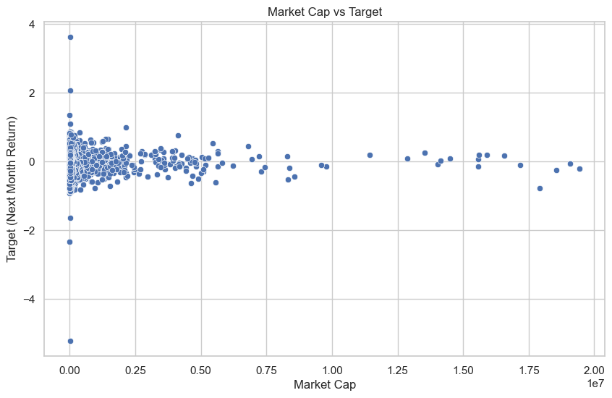
*e) Data Heterogeneity*: The 2000-2008 data span includes major events like the dot-com bubble and the early financial crisis, leading to different impacts on stocks and sectors, affecting predictiveness.

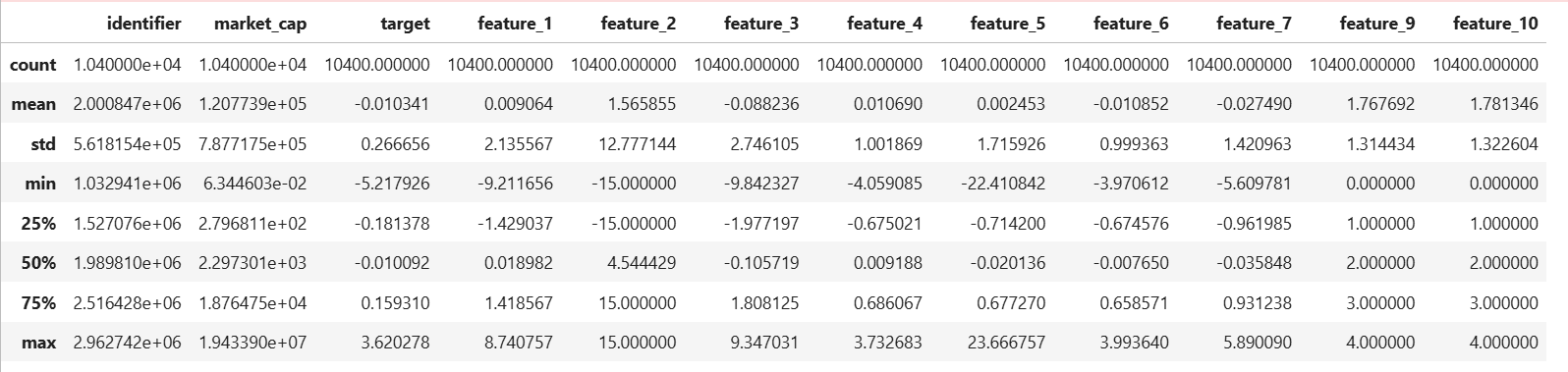
2. **Data Exploration:** Load the data and perform exploratory data analysis (EDA). This should include

summary statistics, visualizations, and any important observations about the data.

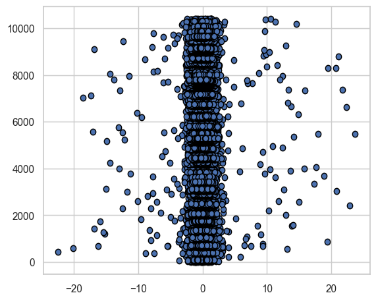
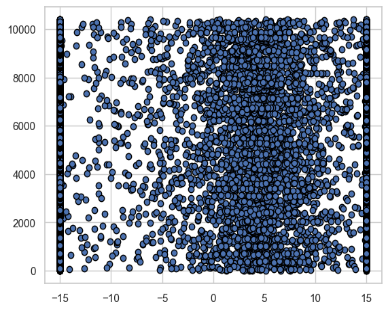
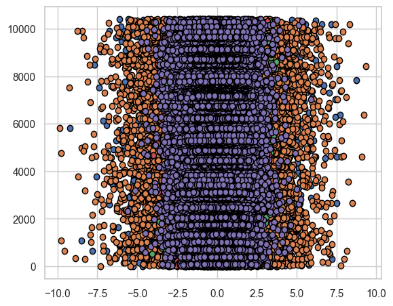
* Dataset Structure: **Entries**: 10,400 and **Features**: 16 columns with a mix of numeric, categorical, and boolean data types.
* Target Variable: The target variable is normally distributed (fig 2), with a mean close to zero and a range from -5.21 to 3.62, indicating the presence of outliers or extreme cases typical in financial data (fig-5 shows outliers). The distribution plot confirms that the target variable follows a normal distribution.
* Sector: This categorical variable has 105 missing values (fig 3) that cannot be filled.
* Market Cap: There are 259 missing values in market cap, which were **imputed using forward fill** for each identifier. Most stocks are clustered at a market cap below 5 million, suggesting a concentration of smaller companies fig 3&4.

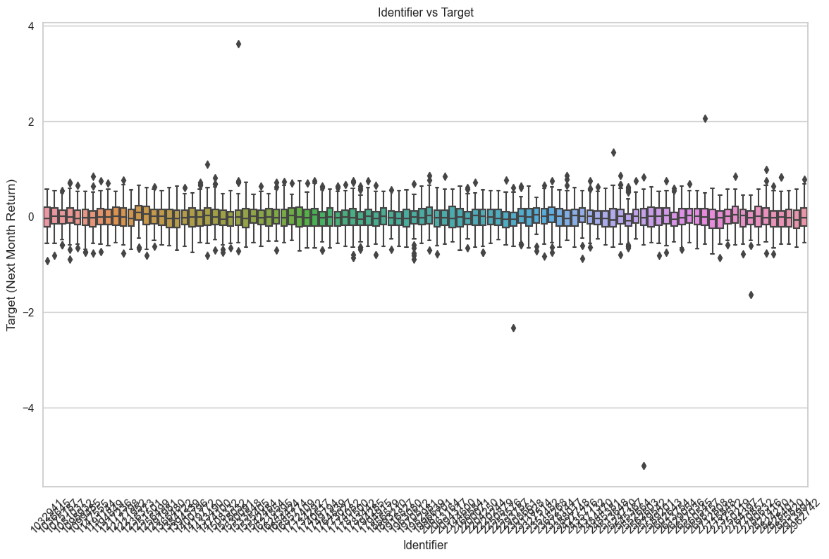
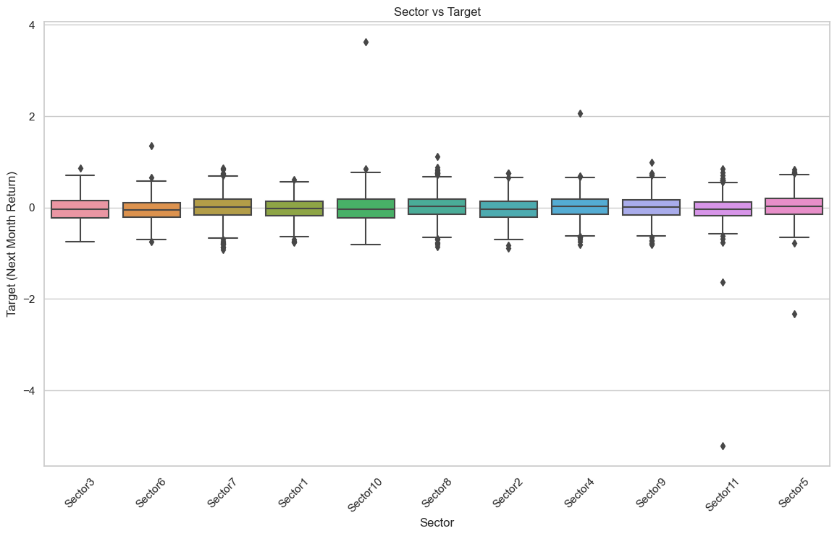


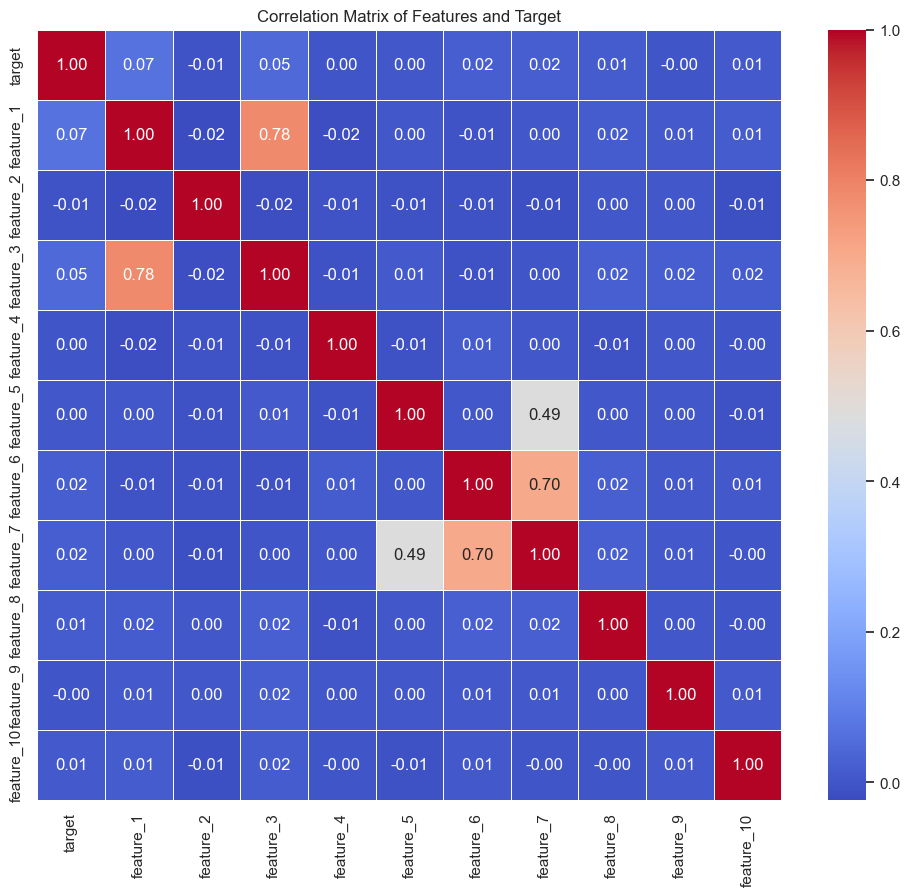
****

* Feature 1,3,4,6,7 are numeric features, the majority of the data points are distributed around its mean or median, with fewer extreme values, the distribution seems fairly symmetrical around zero and has approx normal distribution.
* Feature 2: Infinite values were replaced with 15, and negative infinite values with -15, possibly reflecting a transformation that created a clear distinction between two groups, same is visible in the plot.
* Feature 3: This feature had 102 missing values. Upon plotting, it resembled stationary random white noise, thus imputed the missing values using mean.
* Feature 5: This feature exhibited a significant number of outliers. However, due to the high frequency of these outliers, they might contain useful information. Also due to lack of information I have not removed the outliers. Non-linear models, such as Random Forests or Gradient Boosting Machines, can handle these outliers effectively.
* Features 8, 9 ,10 are categorical variables and the encoded version is plotted in fig-6 in previous page

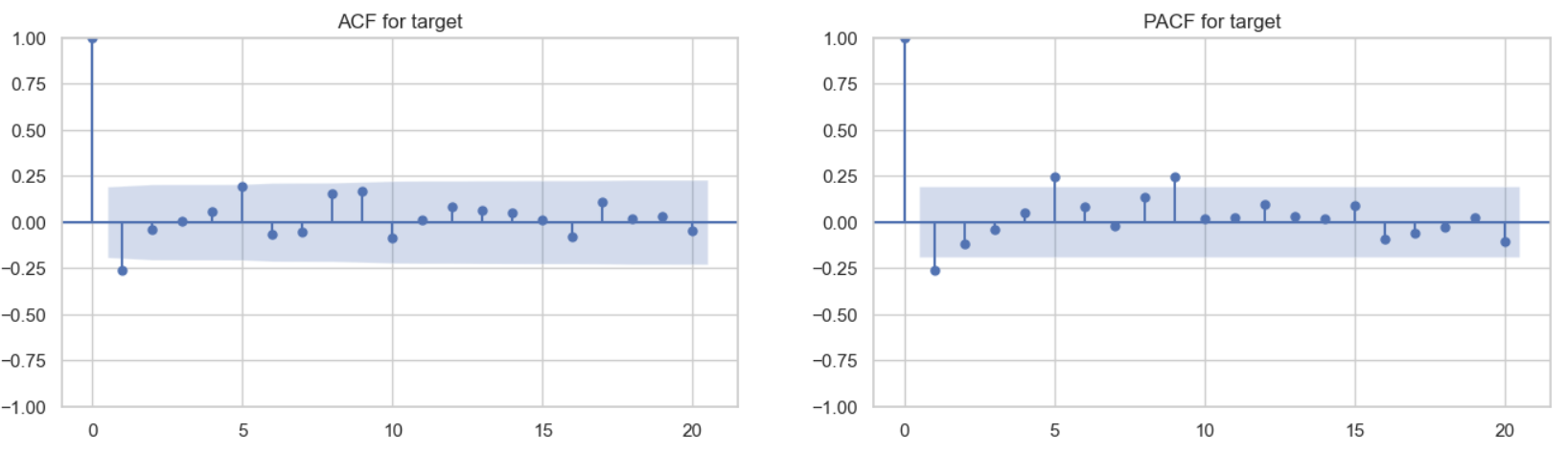
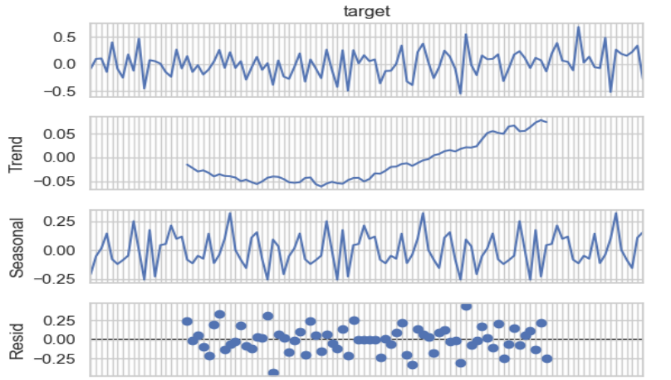
 **features 1,3,4,6,7 feature 2 feature 5**

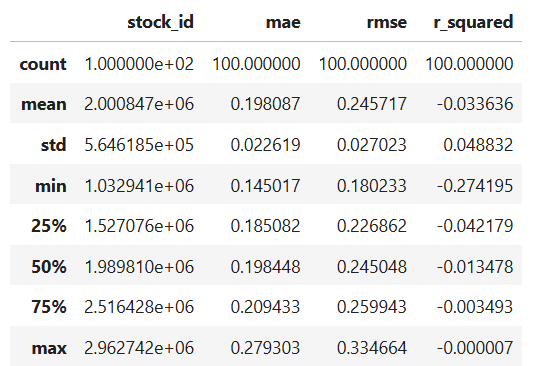
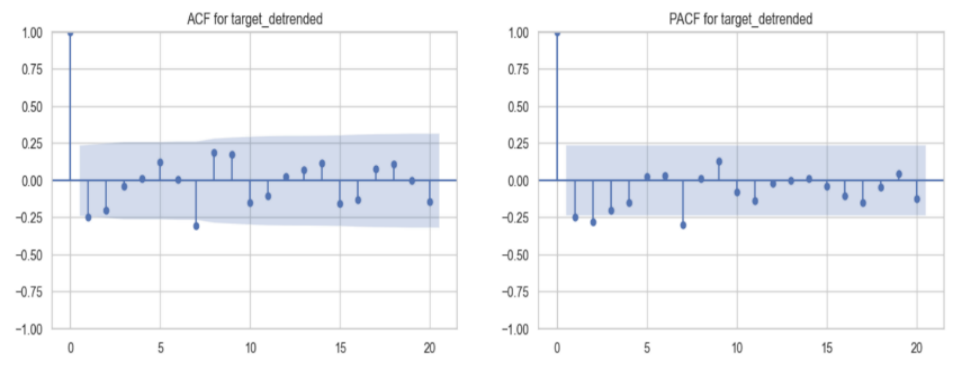
* *Sector and Stock box plots:*Median Returns: Box plots show that median returns across all sectors and stocks are near zero, indicating typically small average returns.
* Variability: The IQR is consistent across sectors and stocks, suggesting similar variability in returns.
* Outliers: Outliers are present in several sectors and stocks, potentially representing unusual market events or shocks.
* Outlier Analysis: Identified 1.4% outliers using IQR, and 1.35% and 0.3% outliers using Z-scores for sectors and stocks.
* Outlier Removal: Removed outliers based on stock identifier Z-scores, with a ±3 threshold for normally distributed data; adjusted thresholds for skewed distributions.



*Correlation with Target Variable:* The correlation matrix shows very low correlations between the target variable and all features, with the highest being only 0.07 (with feature\_1), suggesting a lack of strong linear relationships and the potential need for complex, non-linear models. *Multicollinearity* is evident in feature pairs like Feature 1 & Feature 3 (0.78), Feature 5 & Feature 7 (0.49), and Feature 6 & Feature 7 (0.70), with VIF values ranging from 1.001 to 3.788, indicating moderate multicollinearity. However, this is not severe enough to require immediate action, though it reinforces the potential value of non-linear models for capturing relationships in the data.  
*Feature Engineering*: Various techniques were employed, including creating lagged variables, sector returns, rolling window features, interaction and polynomial features, and extracting date/time features. These enhancements aimed to capture temporal dependencies, sector-wide movements, hidden interactions, non-linear relationships, and potential seasonal effects in the data.  
*Feature Selection*: A combination of correlation thresholds, Lasso regression, and Random Forest was used to refine the feature set. This approach identified the most impactful predictors, ensuring that both linear and non-linear relationships were effectively captured.  
*Stationarity Assessment*: The Augmented Dickey-Fuller (ADF) test was conducted to evaluate the stationarity of the target variable, using a 0.01 p-value threshold for determining stationarity. The analysis identified two stocks (IDs 2378144 and 2657036) as non-stationary, as indicated by their higher p-values in the initial test results.  
*Impact of Detrending*: After applying detrending to the non-stationary series, the ACF and PACF plots showed a reduction in significant lags, and the ADF test results indicated a lower p-value, suggesting improved stationarity. This demonstrates that detrending effectively mitigated underlying trends, enhancing the suitability of the data for time series modeling.

**Time Series Analysis**: Modeled the target variable using ARIMA to see if I can get any predictability from time series*Model Performance*: The negative R-squared value indicates that the model is performing worse than a simple mean-based model. The relatively high MAE and RMSE values suggest that the model is not accurately predicting the target variable.  
*Conclusion*: The time series model seems to have failed in capturing any meaningful predictability within the target variable. This outcome aligns with the ACF and PACF analysis, which did not show significant lags, implying that the target variable might not have strong temporal dependencies or patterns that can be exploited by a time series model like ARIMA.



**3. Model Development:** *Build a predictive model to forecast the target variable for the period from 2006-2008. It is up to you to decide on the model and how to structure the study.*

Key Exploratory Data Analysis Insights:

* Low Correlation with Target: Suggests non-linear relationships, favoring non-linear models.
* Multicollinearity: Present in some features, but less problematic for non-linear models like tree-based methods.
* Poor performance of the time series model: I switched to a cross-sectional regression using a machine learning model, assuming consistent feature-return relationships across stocks to allow data pooling. With only 104 data points per stock, pooling was necessary for reliable predictions.

*A) Explain why you chose this particular model: Why Random Forest?*

1)Robust to Outliers: Reduces the influence of outliers, making it ideal for data where feature information is not known.

2)Handles Non-Linear Relationships: Captures non-linear interactions without needing explicit feature engineering

3)Feature Importance: Provides insights into the most influential features, aiding in understanding return drivers. 4)Robust to Multicollinearity: Less affected by multicollinearity.

5)Strong Predictive Performance: Generalizes well in noisy environments, making it effective for financial forecasting.

*B) Discuss the strengths and weaknesses of your chosen model. Strengths of Random Forest:*

1)Robustness to Overfitting: Utilizes multiple decision trees, reducing overfitting and ensuring good generalization. 2)Handles High-Dimensional Data: Manages many features without overfitting, ideal for complex financial datasets. 3)Non-Linear Data Compatibility: Captures complex relationships as it doesn’t assume linearity.

4)Feature Importance: Ranks feature importance, aiding in selection and understanding key drivers.

5)Handles Missing Values and Outliers: Manages missing data and outliers effectively through imputation or different data subsets.

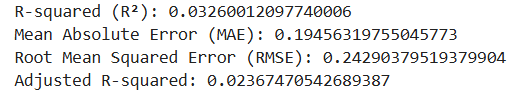
*Weaknesses of Random Forest:*

1)Computationally Intensive: Requires significant resources, leading to longer training times.

2)Model Interpretability: Difficult to interpret individual predictions, making it a "black box" model.

3)Memory Usage: Consumes considerable memory, which can be limiting in resource-constrained environments.

*C) Evaluate the performance of your model and describe your evaluation method.*

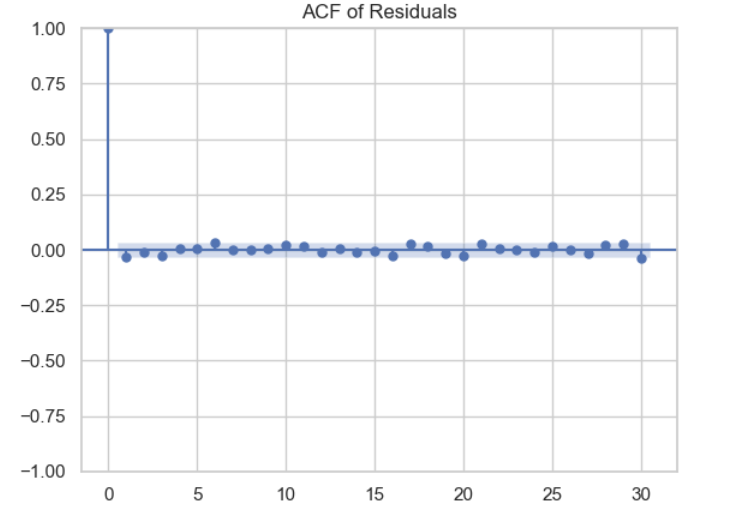
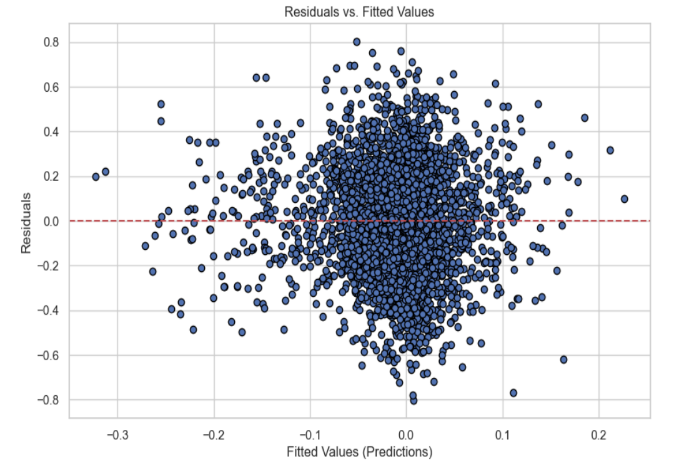
 Random Forest model yielded improved results: **R-squared (R²)**: 3.3%, **MAE**: 0.19, **RMSE**: 0.24, **Adjusted R²**: 2.4%. While the R² is not very high, it’s a reasonable outcome given the inherent complexity and noise in stock return data. This suggests that Random Forest is an effective choice for this dataset.

*Ensure you avoid any potential forward-looking bias.* **Chronological Split**: I used a strict chronological split, where earlier data before 2006 was used for training and later data 2006 to 2008 was used for testing. I used TimeSeriesSplit function to maintain the chronological order in the data, ensuring no future information leaked into the training process.

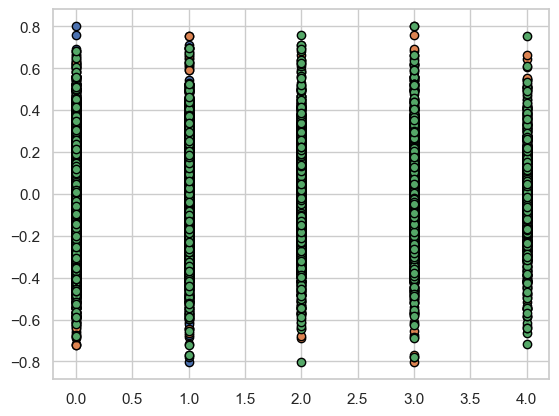
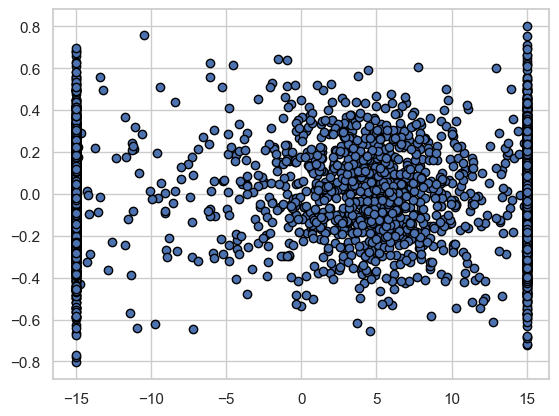
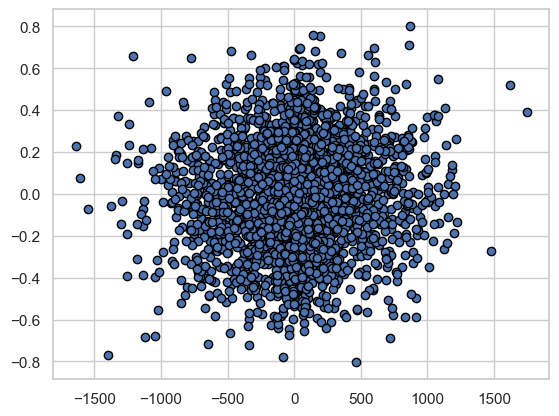
**Feature Engineering**: Lagged features and rolling statistics were created using only past data, avoiding any look-ahead bias.

*D) Residual Analysis:* **Residual vs fitted values**: The residuals are mostly random and symmetric around zero, with a consistent variance, suggesting that the model is making errors in a balanced and unbiased way.

**Residual ACF plot** shows residuals are not autocorrelated. It indicates that the model has adequately captured any temporal relationships in the data, and the residuals are behaving as expected—independently and randomly. This further confirms that there is no significant structure in the residuals that the model has failed to capture.

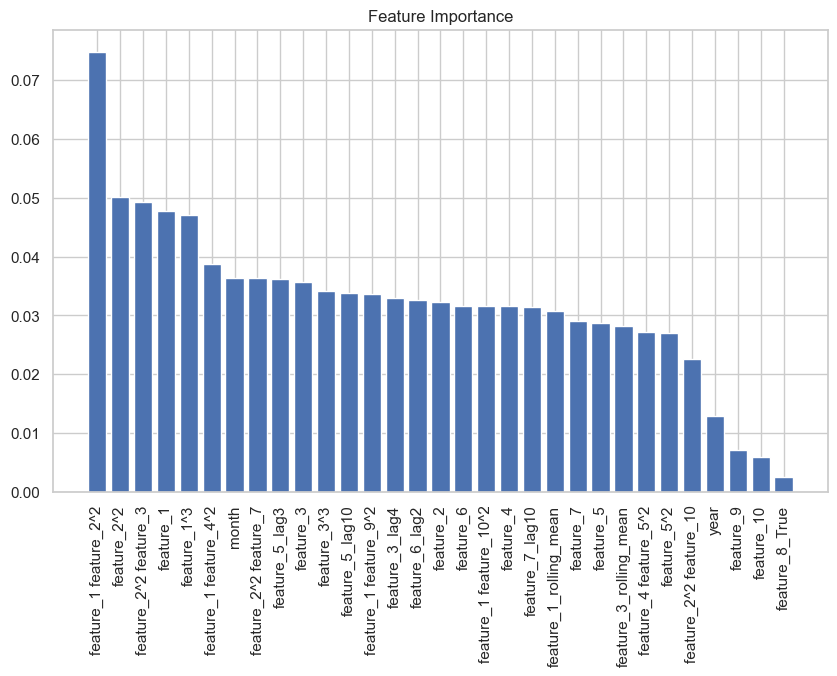


**Residual vs features** plots indicate that key features like 'feature\_1 feature\_2^2', 'feature 1', and 'feature 3' are effectively captured in the model, as shown by the even distribution of residuals. The residuals reveal that feature\_2 independently is not well modeled. However, feature\_2 is only indirectly captured through 'feature\_1 feature\_2^2', highlighting the importance of feature engineering. Additionally, categorical variables (features 8 to 10) are not fully captured, suggesting that adding interaction terms could further improve the model.



***Residuals vs feature\_1xfeature\_2^2 Residuals vs feature 2 Residuals vs categorical features***

***4. Feature Importance:*** Identify and discuss which features are important for your chosen model from Task 3.

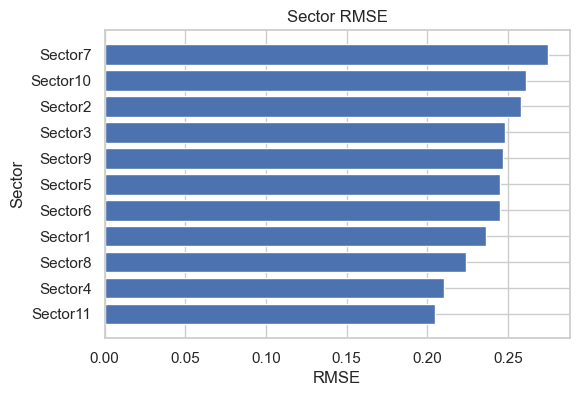
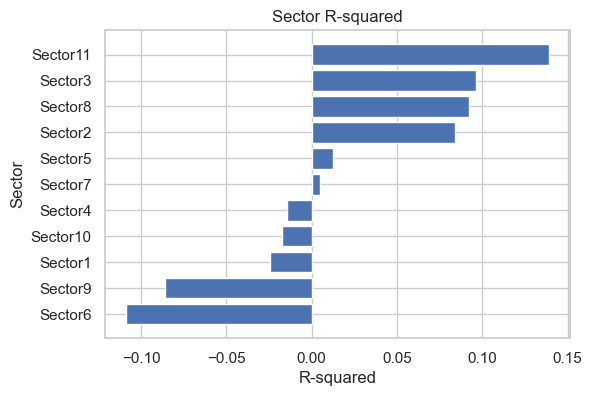
Feature importance plot from the Random Forest model highlights several key contributors to the model's predictions:

* **‘feature\_1 feature\_2^2’** emerges as the most significant feature, underscoring the critical role of interactions and polynomial terms in capturing complex relationships.
* **‘feature\_2^2’**, **‘feature\_1’**, and **‘feature\_3’** are also highly important, indicating that both quadratic and standalone features are pivotal for the model's accuracy.
* Temporal aspects like **‘month’** and lagged features such as **‘feature\_6\_lag2’** are significant, showing that past values and temporal patterns are crucial in forecasting future returns.
* Categorical features like **‘feature\_10’** and **‘feature\_8\_True’** are less impactful, suggesting that they contribute less to the model’s overall prediction power.

In summary, the model relies heavily on interaction terms, polynomial features, and key features like **‘feature\_1’** and **‘feature\_3’**, alongside temporal dynamics, to effectively predict stock returns.

5. **Sector Analysis:** *Analyze whether your model from Task 3 is equally predictive across all sectors.*

The model's performance varies significantly across different sectors:



* **R-squared Analysis**: The R-squared values range widely across sectors, with some sectors like Sector 11 and Sector 9 showing higher R-squared values, indicating that the model explains a relatively larger portion of the variance in these sectors. On the other hand, sectors like Sector 7 and Sector 6 exhibit negative or near-zero R-squared values, suggesting poor model fit for these sectors.
* **RMSE Analysis**: The RMSE values are fairly consistent across sectors, indicating that while the model's errors are similar in magnitude, the ability to explain variance (as seen in the R-squared) differs. This consistency in RMSE could imply that the model is making similar magnitudes of errors across sectors, but the nature of the errors or the variance captured varies.

*Could you use this sector-specific information to your advantage, if so, how?*

* **Separate Models for Each Sector:** Instead of using a single model for all sectors, train individual models for each sector. This allows each model to specialize in the specific patterns within its sector.
* **Ensemble of Sector-Specific Models**: Combining predictions from sector-specific models using an ensemble approach. This could help capture both sector-specific and general patterns.
* **Feature Engineering**: Incorporating sector-specific features. Using domain knowledge to engineer new features that are relevant to those sectors might enhance model performance.

**6. Model Improvement:** *Suggest additional data sources that could potentially improve the performance of your model*

To enhance the model's predictive accuracy, we can consider integrating below data:

* *Macroeconomic Indicators*: Interest rates, inflation, and GDP growth rates to reflect broader economic influences.
* *Sector-Specific Data*: Industry metrics and commodity prices to capture sector-specific dynamics.
* *Sentiment and News Data*: social media and news sentiment to incorporate investor behavior and market reactions.
* *Alternative Data*: Satellite data and web scraping to obtain real-time indicators of company performance.
* *Earnings and Analyst Estimates*: Earnings announcements and analyst revisions to capture stock price movements.
* *Liquidity and Volatility Data*: Bid-ask spread and historical volatility to account for market liquidity and risk profiles.

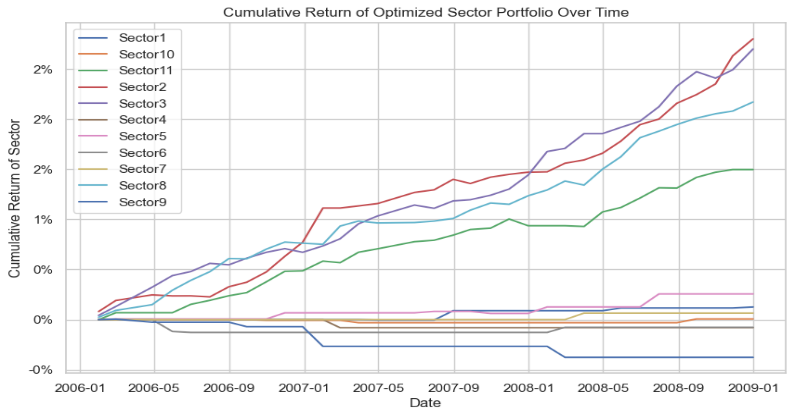
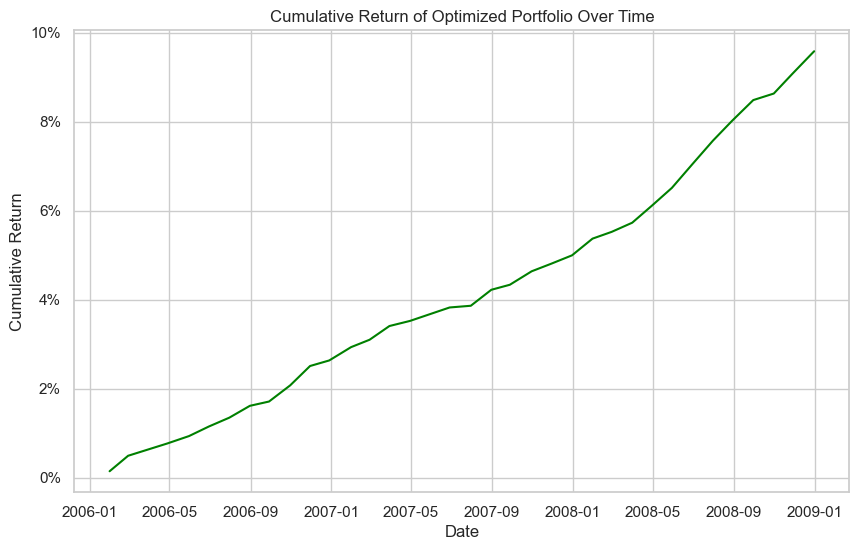
7. **Trading Strategy:** *Propose a trading strategy based on the signal generated by your model*

* I developed a long-short strategy using predicted monthly returns to select stocks and determine positions. The strategy goes long on the top 5 and shorts the bottom 5 stocks by predicted returns, which yielded the best results compared to other stock numbers tested (1, 10, 20, 40, 50).
* For portfolio weights I used an equal-weighted approach and also a portfolio optimization aiming to maximize the Sharpe ratio. The equal-weighted method provided the best performance. Can also consider Market-cap weighting but discarded in this case due to high variability in market cap data.

*Include any performance metrics you consider important.*

The performance metrics that I used were Sharpe ratio, annualized return and max drawdown.

*Construct a graph showing the cumulative generated PnL (Profit and Loss) over time*



*Ensure your trading strategy is coherent with your answers from Task 5.*

I compared the R-squared values with the cumulative returns plots and observed that sectors with positive R-squared values are performing well. Specifically, **Sector 2 (red), Sector 3 (purple), Sector 8 (light blue), and Sector 11 (green)** are the top-performing sectors, reflecting their ability to explain a significant portion of returns, as indicated by the R-squared plots. Even Sector 5 (pink) shows a positive cumulative return, which aligns with its positive R-squared value. (figure-2 in sector analysis Task 5 shows the Sector R squared)