**Slide 1: Title Slide**

Title: Stock Price Prediction in the Presence of High Volatility, Missing Data, and Extreme Events Subtitle: Evaluating the Performance of Different Prediction Models and Data Preparation Techniques Your Name Date

**Slide 2: Introduction**

* Introduce the challenges in stock price prediction, particularly in the presence of high volatility, missing data, and extreme events
* Explain the significance of addressing these challenges for investors and financial institutions
* State the research questions, emphasizing the need to find effective techniques for handling missing data, outliers, and high volatility

The main challenges in predicting stock prices, considering the presence of high volatility, missing data, and outliers due to extreme events:

* Difficulty in accurately estimating the impact of external factors such as news, economic events, and market sentiment on stock prices
* High-frequency data may be affected by noise, leading to challenging predictions and potentially impacting model performance
* Heteroskedasticity in return distributions complicates modelling and requires appropriate techniques to account for it

If handled correctly in a robust way, greater prediction accuracy may be achieved. Greater accuracy allows for trading systems to be implemented in volatile markets and robust to fluctuations in price, and consequently assist in risk mitigation.

Primary goal is investigate suitable approach for prediction when there is high volatility missing data and outliers due to extreme events

1. How does high volatility in stock price data impact the accuracy of traditional time series prediction models, and which models are most robust to high volatility, missing data, and outliers?
2. What are the most effective methods for handling missing data and outliers in the context of stock price prediction, considering high volatility and extreme events?
3. How do different pre-processing techniques, such as data imputation or outlier removal, affect the performance of prediction models in volatile stock price data?
4. What are the main challenges in predicting stock prices for the given dataset, and can these challenges be addressed using existing prediction methods or is there a need for new methods?
5. How do various data imputation techniques, such as linear interpolation and rolling mean, impact the accuracy of stock price predictions when dealing with missing data?
6. In what ways can outlier detection techniques be integrated into the stock price prediction process to enhance the model's robustness against extreme events and improve overall predictive performance?
7. How can the relationships between the two unnamed price series be leveraged to improve stock price predictions, and what is the potential for cross-series prediction, such as using one price series to predict the other or exploiting common trends and patterns between the two series?

considering the presence of high volatility, missing data, and outliers due to extreme events

* Challenges in stock price prediction: • Impact of external factors (news, economic events, market sentiment) • High-frequency data affected by noise • Heteroskedasticity in return distributions
* Significance of addressing challenges: • Greater prediction accuracy for trading systems • Robustness to fluctuations in price • Improved risk mitigation
* Research questions:
  1. Impact of high volatility on time series prediction models
  2. Effective methods for handling missing data and outliers
  3. Influence of preprocessing techniques on prediction performance
  4. Main challenges and potential need for new methods
  5. Impact of data imputation techniques on prediction accuracy
  6. Integration of outlier detection techniques for enhanced robustness
  7. Leveraging relationships between price series for improved predictions

investigate suitable approach for prediction when there is high volatility missing data and outliers due to extreme events

1. How does high volatility in stock price data impact the accuracy of traditional time series prediction models, and which models are most robust to high volatility, missing data, and outliers?
2. What are the most effective methods for handling missing data and outliers in the context of stock price prediction, considering high volatility and extreme events?
3. How do different preprocessing techniques, such as data imputation or outlier removal, affect the performance of prediction models in volatile stock price data?
4. What are the main challenges in predicting stock prices for the given dataset, and can these challenges be addressed using existing prediction methods or is there a need for new methods?
5. How do various data imputation techniques, such as linear interpolation and rolling mean, impact the accuracy of stock price predictions when dealing with missing data?
6. In what ways can outlier detection techniques be integrated into the stock price prediction process to enhance the model's robustness against extreme events and improve overall predictive performance?
7. How can the relationships between the two unnamed price series be leveraged to improve stock price predictions, and what is the potential for cross-series prediction, such as using one price series to predict the other or exploiting common trends and patterns between the two series?

**Slide 3:**  **Data Description:**

* A concise description of the data used in the study.
* This section would describe the dataset used for the research, including the source of the data, the time covered, the frequency of the data (e.g., daily, 2-minute intervals), and any pre-processing steps taken to clean or transform the data.
* Describe the dataset used for the research, including the source, time period, and frequency b. Explain any preprocessing steps taken to clean or transform the data

Source of data is unknown, two unnamed price series.

They may possibly be futures data since trading begins early Sunday evening.

time period is 5 years 2008 to 2013 in two minute interval.

So all of the datapoints are equally spaced 2 minute intervals

Saturdays do not appears in the data since the market is closed. Sundays are just the late hours. Otherwise the data appears to be 24hr price excluding missing values

Monday 258

Tuesday 259

Wednesday 261

Thursday 260

Friday 253

Sunday 243

Name: datetime, dtype: int64

Monday 261

Tuesday 262

Wednesday 261

Thursday 261

Friday 260

Sunday 259

Name: datetime, dtype: int64

Chart, bar chart

Description automatically generated Chart, histogram

Description automatically generated

Missing values:

Non-uniform distribution: Instead of a single peak around the expected 2-minute interval, the histogram shows a wide distrution of time gaps

The frequency of streak lengths decreases as the streak length increases. This indicates that shorter streaks (smaller time gaps) are more common in the data, while longer streaks (larger time gaps) are less frequent.

Presence of outliers: The histogram show some bars far away from the main distribution, representing unusually large time gaps. These outliers could be due to holidays, market closed days, or other events that cause irregular time gaps in the data.

Skewness: The histogram might to the right , indicating that the time gaps are not symmetrically distributed around the expected 2-minute interval. This could suggest that there are more frequent occurrences of either shorter or longer time gaps than expected.

Long tails: The histogram has a long tail, indicating that there are a significant number of time gaps that are much larger than the expected 2-minute interval. This could be a sign of irregularities in the data collection process or other factors affecting the time intervals between data points.

Time Series 1 (ts1):

There are 104,817 time gaps in the data.

The average time gap is approximately 17.67 minutes.

The standard deviation of the time gaps is around 101.14 minutes, indicating a wide range of time gaps.

The minimum time gap is 2 minutes.

The 25th percentile (Q1) is 2 minutes, meaning 25% of the time gaps are 2 minutes or less.

The median (50th percentile) is 4 minutes, indicating that half of the time gaps are 4 minutes or less.

The 75th percentile (Q3) is 12 minutes, meaning 75% of the time gaps are 12 minutes or less.

The maximum time gap is 5,718 minutes, which is quite large compared to the other percentiles.

Time Series 2 (ts2):

There are 2,271 time gaps in the data.

The average time gap is approximately 168.24 minutes.

The standard deviation of the time gaps is around 453.51 minutes, indicating a wide range of time gaps.

The minimum time gap is 2 minutes.

The 25th percentile (Q1) is 2 minutes, meaning 25% of the time gaps are 2 minutes or less.

The median (50th percentile) is 2 minutes, indicating that half of the time gaps are 2 minutes or less.

The 75th percentile (Q3) is 4 minutes, meaning 75% of the time gaps are 4 minutes or less.

The maximum time gap is 2,838 minutes, which is quite large compared to the other percentiles.

From these statistics, we can observe that Time Series 1 has a higher average time gap and a larger maximum time gap compared to Time Series 2. However, both time series have a wide range of time gaps, as indicated by their large standard deviations.

From the distribution of missing value lengths in the time series data, it seems that most of the missing values are of shorter lengths, with a few longer gaps. The descriptive statistcs show that the values however are widely distributed. In this case, a combination of imputation methods might be appropriate to handle both short and long gaps..

For shorter gaps, can use time-weighted interpolation or moving mean imputation, as they can capture local trends and patterns in the data. Time-weighted interpolation assumes that the stock prices change linearly within short time gaps, while moving mean imputation assumes that the local average of the surrounding data points is a reasonable estimate for the missing values.

For longer gaps, consider using a more sophisticated method like the Expectation-Maximization (EM) algorithm, which iteratively estimates the missing values based on the available data and the fitted model. This method can be more appropriate for handling longer gaps, as it takes into account the overall structure and patterns in the data.

**Exploratory data analysis:**

Data contains significant number of missing values

* **For the number of missing days, they are not likely to all be holidays (>100 per year) so reflect actually missing data. That should be considered as reducing the amount of relevant data points.**
* **However for 5 years of 2-minute interval data, it is unlikely that will have highly variable time gaps between data points. This is because the 2-minute interval is small enough to capture most of the variation in the data, and the regularity of the interval reduces the likelihood of gaps.**
* Assume that from visual analysis there are extreme market events that are meaningful to include
* Assume that outliers would be due to incorrect measurements and do not reflect the behaviour of the stock but verify this, so have to analyse the extreme values of the returns to clarify the nature of the outliers
* Assume noise is not due to measurement error but inherent in the data since cannot verify otherwise

Characteristics:

Stationary rice returns series

Non linear upwards trend

No seasonal components in the hourly, daily or weekly

Some periodic components in the daily and weekly

Non normal distributions of price returns with heteroskedasticity evident in the returns series and volatility clustering

Insert eda results

Seasonal decomposition:

Ts1

Weakening but strong correlation between price and trend. So there is evidence of trend in all frequencies. Possible seasonality likely only in the monthly time frame where there is an increase at the start of the year. Noise is random up to monthly time frame

Residuals are strongly corelated so suggests that there is underlying structure that is not captured by seasonal decomposition.

Ts2

More likely to be seasonal at higher timescales, but noise and lack of seasonality at lower timescales.

strong upward trend across all timescales. There is an underlying structure that is not captured by seasonal decomposition, as suggested by the close relationship between residuals.Each exhibits an annual seasonality of an increase in price and a decrease in price, respectively.Both exhibit substantial nonlinear trends.

Periodogram

Both show some seasonality or periods appearing in the intra day timescale and daily. 2 hours, 16 hours, 48 hours

Significant lags in the hourly and daily frequencies indicates serial correlation.

Statistics on log returns

Descriptive statistics:

Mean:

ts1 -3.164616e-08

ts2 7.095238e-10

dtype: float64

Variance:

ts1 4.598756e-09

ts2 2.315885e-07

dtype: float64

Covariance:

ts1 ts2

ts1 4.598756e-09 -5.300196e-09

ts2 -5.300196e-09 2.315885e-07

Normality tests:

Kolmogorov-Smirnov test for ts1: Null hypothesis = 0.4995612806420975, p-value = 0.0

Kolmogorov-Smirnov test for ts2: Null hypothesis = 0.4978485727663148, p-value = 0.0

Skewness for ts1: 3.9155086367800345

Kurtosis for ts1: 261.99977845222173

Skewness for ts2: -0.7193960568098621

Kurtosis for ts2: 61.32295854243591

* it appears that the logarithmic returns of both time series (ts1 and ts2) do not follow a normal distribution. The p-values for the Kolmogorov-Smirnov tests are 0.0, which indicates that we can reject the null hypothesis that the data is normally distributed.
* Additionally, the skewness and kurtosis values provide further evidence of non-normality. For ts1, the skewness is close to 0, which suggests that the data is approximately symmetric. However, the kurtosis value of 138.45 is significantly higher than 3 (which is the kurtosis of a normal distribution), indicating that the distribution has heavy tails and a high peak. For ts2, the skewness is 0.99, which suggests that the data is positively skewed, and the kurtosis value significantly different from 3 as well.
* Since it is not normally distributed, then there will likely be fat tails and thus much of the data would be classified as an outlier if considered normally distributed.
* The mean values for both time series (ts1 and ts2) are very close to zero, indicating that the average returns are close to zero over the given time period.
* The variances for both time series are quite small, suggesting that the price changes are relatively small on average. However, the variance of ts2 is larger than that of ts1, indicating that ts2 has more variability in its price changes.
* The covariance between ts1 and ts2 is negative, which implies that the price changes in the two time series tend to move in opposite directions. However, the magnitude of the covariance is quite small, suggesting that this relationship may not be very strong.
* The Kolmogorov-Smirnov test results show that the null hypothesis is rejected for both time series (p-value = 0.0), indicating that the data does not follow a normal distribution.
* The skewness values for ts1 and ts2 are -0.53 and 2.94, respectively. A negative skewness for ts1 indicates that the distribution is skewed to the left, with a longer tail on the left side. A positive skewness for ts2 indicates that the distribution is skewed to the right, with a longer tail on the right side.
* The kurtosis values for ts1 and ts2 are 469.29 and 878.89, respectively. These values are significantly higher than the kurtosis of a normal distribution (which is 3), indicating that both time series have heavy tails and a high probability of extreme values (outliers).
* Given these observations, it's clear that both time series exhibit non-normal behavior, with heavy tails and a high likelihood of extreme values
* Given these results, it's important to consider models that do not rely on the normality assumption when analyzing your stock price data. GARCH models for volatility, and LSTM
* From rolling mean and standard deviation plots, using window size of 4 hours, can see that the mean is fairly constant whilst there is flucatuuation around volatility in the std
* Augmented Dickey-Fuller (ADF) test you provided, the test statistic is -102.5273, which is much smaller than all the critical values at the 1%, 5%, and 10% significance levels. The p-value is 0, indicating strong evidence against the null hypothesis.
* The null hypothesis of the ADF test is that there is a unit root, meaning the time series is non-stationary. Since the test statistic is smaller than the critical values and the p-value is 0, you can reject the null hypothesis. This suggests that your time series is stationary.
* In summary, the results of the ADF test indicate that your 5 years of 2-minute interval stock price data is stationary.
* Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test you provided, the test statistic is 0.07857, which is smaller than all the critical values at the 1%, 2.5%, 5%, and 10% significance levels. The p-value is 0.1, which is greater than the common significance levels of 0.01, 0.025, 0.05, and 0.1.
* The null hypothesis of the KPSS test is that the time series is stationary around a deterministic trend. Since the test statistic is smaller than the critical values and the p-value is greater than the common significance levels, you fail to reject the null hypothesis. This suggests that your time series is stationary around a deterministic trend.
* In summary, the results of the KPSS test also indicate that your 5 years of 2-minute interval stock price data is stationary, which is consistent with the results of the Augmented Dickey-Fuller test you provided earlier.
* For both time series, the test statistics are highly negative, indicating strong evidence against the presence of a unit root. The p-values are also very low, suggesting that the null hypothesis of a unit root can be rejected at any reasonable significance level.
* In general, the results suggest that the time series are stationary and do not exhibit any trend or structural change over time.
* tests indicate that the data is stationary and exhibits no significant autocorrelation, it may suggest that the noise is an inherent part of the data-generating process.

Swaured and absolute returns indicate some periods of high volatility, that large price changes are followed by large price changes

Null hypothesis not rejected so suggests that the data is not homoskedastic

reject the null hypothesis and conclude that there is evidence of heteroskedasticity in the time series.

Both methods of deecting outliers are consistent with the extreme market event of 2008, but also are consistent in identifying ismilar outliers. Including outliers may affect the stability of some models, by removing them could improve the stability and improve prediciton. Would benefit from applying a sensitivty analysis and ross validating the outlier detectionn with the accuracy of the models

Ts1

Resampled data for frequency: H

Correlation between Price and Trend: 0.9996736884969083

Resampled data for frequency: D

Correlation between Price and Trend: 0.9990496544718067

Resampled data for frequency: W

Correlation between Price and Trend: 0.8266779487916374

Resampled data for frequency: M

Correlation between Price and Trend: 0.8353758128930284

Chart

Description automatically generated Chart, line chart

Description automatically generated

Chart

Description automatically generated Chart

Description automatically generated with medium confidence

Ts2

Resampled data for frequency: H

Correlation between Price and Trend: 0.999490609174061

Resampled data for frequency: D

Correlation between Price and Trend: 0.9979701711481833

Resampled data for frequency: W

Correlation between Price and Trend: 0.8171025108286337

Resampled data for frequency: M

Correlation between Price and Trend: 0.8149782321598528

Chart, histogram

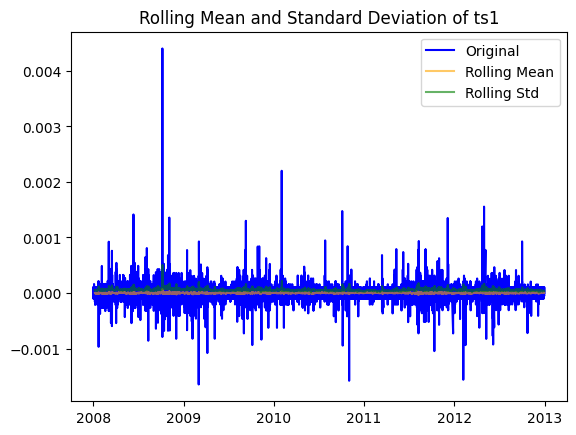
Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generatedDiagram

Description automatically generated with medium confidence

 A picture containing text, antenna, screenshot

Description automatically generated

**Slide: Modelling outline**

* Outline rationale and assumptions made for answer the research questions.
* Outline the structure of the approach used and how the research questions will be investigated.
* Explain the reasoning for the choices made

Conduct experiments on how volatility, outliers due to extreme events and missing values can be incorporated into developing suitable predictions for price returns, and the suitability of various machine learning models with the simple choice of single step forecasting for univariate and multivariate cases.

Hypotheses:

* + H1: ts1 can predict ts1
  + H2: ts2 can predict ts2
  + H3: ts1 can predict ts2
  + H4: ts2 can predict ts1
* the motivation behind testing these hypotheses is interest of investigating the relationships between the two time series and their predictability and what uncertainty is reduced or introduced in predictions by applying various imputation and feature engineering methods. ie can past volatility and outliers due to extreme events be used as features to predict future volatility and extreme events. It can enhance the model's robustness against extreme events and improve overall predictive performance.
* Experiments:
  + Original data with missing values removed and outliers not identified
  + Original data with missing values removed and outliers identified
  + Imputed data with outliers not identified
  + Imputed data with outliers identified

For each hypothesis four experiments will be conducted by modifying the data preparation method will be implemented

Discussion:

* The choice of imputation technique can lead to different levels of preservation of the original data structure and noise levels, which can subsequently impact model performance
* The success of imputation techniques may depend on the specific characteristics of the data, such as the extent of missing data and the nature of the underlying patterns
* Want to account for extreme market conditions
* Black swan events are unpredictable, but current economic climate is very uncertain and there is significant instability in many banking institutions that could lead to extreme market behaviour. (find a reference of this)
* Test volatility based models and deep learning models that are more robust to outliers
* Given that the distribution of returns shows volatility clustering and large kurtosis for both data series, and visualizing the returns reveals very large changes in price over a very narrow range in times, it suggests that the outliers in your dataset are likely to represent genuine extreme events in the stock market. These extreme events could be due to market news, earnings announcements, or other factors that cause sudden and significant price movements.
* In this case, it's generally better to keep the outliers in the dataset, as they provide valuable information for your time series and LSTM models. Removing or adjusting these outliers might lead to a loss of important information that could help your models better understand and predict future stock price movements.
* There are a significant number of outliers in both stock price series, with the first series having a higher number of outliers than the second one.
* The number of common outliers between the two series is relatively smaller compared to the total number of outliers in each series. This suggests that the outliers in the two series are mostly independent of each other, and it is less likely that the outliers are driven by a common factor or event.
* The presence of outliers in both series could be attributed to various factors, such as stock-specific news, market-wide events, or data errors. However, the smaller number of common outliers suggests that stock-specific factors may have a more significant impact on the occurrence of outliers than market-wide events.
* Since the non nomralrity of the data, it might be better suited to use other methods to filter for the extrme value outliers due to the very large tails.

Rationale:

1. **Influence of imputation**: Imputation methods can sometimes introduce artificial outliers or mask real outliers, depending on the method used and the characteristics of the data. By testing for outliers using the imputed log return data, you can assess the impact of the imputation method on the distribution of returns and make any necessary adjustments to your imputation approach or outlier detection methods.
2. **Model performance**: Outliers can have a significant impact on the performance of your models, especially if they are sensitive to extreme values. By testing for outliers using the imputed log return data, you can ensure that your models are robust to the presence of outliers and that your results are not unduly influenced by extreme values in the imputed data.

Arguments for:

Continuity of the time series is essential for accurate outlier detection for several reasons:

1. Temporal dependencies: Financial time series data, like stock returns, often exhibit temporal dependencies, meaning that the value at a given time point is influenced by the values at previous time points. To accurately detect outliers, it is essential to maintain the correct order and spacing of the data points to preserve these dependencies. Ignoring missing data or leaving gaps in the time series can disrupt the underlying temporal structure, leading to misleading results.
2. Statistical tests and measures: Many outlier detection methods, such as moving averages or z-scores, rely on the calculation of summary statistics or other measures that require a continuous time series. Discontinuities in the time series can distort these calculations and result in incorrect identification of outliers. For example, missing values can artificially lower the mean or standard deviation, causing data points to appear more extreme than they actually are.
3. Context: Outliers are usually identified by comparing individual data points to their surrounding context. In a time series, this context is often based on the immediate past or future values. Discontinuities in the time series can make it difficult to accurately assess the context of a given data point, leading to incorrect outlier detection.

Seasonality and trends: Many time series exhibit seasonality or underlying trends, which can influence the identification of outliers. Discontinuities in the time series can obscure these patterns or introduce artificial ones, complicating outlier detection and potentially leading to false-positive or false-negative results.

**Methodology:**

Data Preparation

* Describe the data imputation techniques used, such as linear interpolation and rolling mean
* Explain the methods used for outlier detection, for example, clustering-based methods

Given the nature of the missing values, Mean imputation, linear interpolation and median interpolation were considered.

* Choose to impute using two different methods to handle short gaps and long gaps in the data because will preserve structure of time series and not affect stability of time series modelling. This assumes that the price continues to trend over missing days which may be holidays.
* Impute the issing values using the two methods for short time frame and long time frame sectoins of missing values. Short time gaps is consider less than 4 hours (240 minutes), since the data is very dense in ery short intervals of 2 minutes. hus any time gaps longer will be imputed using the expectation maximzation. Will use a rolling window of 4 hours for the moving mean imputation.
* Will use the orignal data including saturdays
* For the machine learning models, the gaps equal to or longer than a day in duration can be removed as the models are more robust to irregularities.
* The result of the forecasting with the models and testing accuracy will give insight on how

Linear interpolation proved to be the least sensitive to the underlying volatility of the price. It introduced no new sources of volatility whereas median and mean did. Mean and median introduced artificial volatility. It may be possible that the volatility would have been present, but it is not likely that these methods are robust enough to account for the underlying distribution of the returns.

Outliers were identified based on the squared price returns series. Peak over threshold and cluster methods were compared. The DBSCAN cluster method was found to be more appropriate as it identified outliers by the cluster size relative to their neighbours, rather than the peak over threshold method which applied a threshold to every point calculated from the entire data set.

**Data Resampling**

* Describe the rationale behind resampling the time series to daily frequency, including capturing potential daily and weekly periodicities and computational efficiency
* Explain the resampling methods used, such as resampling to daily frequency using the mean of the original data

The data was resampled to hourly and daily frequencies. These were chosen given the clustering of the missing values in shorter time frames, uncertainty in the true value would be reduced by down sampling to longer time steps, but also in consideration of computational demand. Training time would be significantly lengthened for higher frequencies due to the large amount of data points, consequently algorithmic efficiency would be another factor to include, but it was not considered here.

**Slide 5: Benchmark Model**

* Explain the random walk simulation as a benchmark for comparison
* Describe how the random walk model is used to simulate the unpredictability of stock prices in the short term
* Discuss the importance of comparing the performance of prediction models against this benchmark

The simple assumption that the price series behaved like a random walk would serve as a sufficient benchmark prediction to compare the chosen models against. A simple random walk was applied using the mean and standard deviation of the returns price series for both the univariate (H1H2) case and the multivariate (H3H4) case. Comparison to the random walk RMSE will indicate whether the chosen models, or modelling in general is appropriate for prediction or whether the series is inherently random and there is no clear underlying structure or pattern that can be exploited for forecasting purposes.

**Slide 7: Prediction Models**

* Introduce the chosen prediction models (e.g., ARIMA, GARCH, LSTM, etc.)
* Describe the key characteristics of each model and why they were chosen for the study
* Mention that these models will be applied to test the hypotheses on the four versions of the dataset

Because of the non-linearity and fat tailed distribution of price returns, traditional modelling would not likely be able to account for the heteroskedasticity of the returns (ie changing volatility over time). Classical machine learning regression methods are employed as a first approach: KNN clustering, Random Forest, Decision Tree, Gradient Boosting Tee and Light Gradient boosting tree and Gaussian process

**Slide 8: Model Evaluation**

* Describe the evaluation metrics (MSE, MAE, R-squared) and their significance in assessing the prediction models' performance
* Explain the process of comparing the performance of different models on different dataset versions, including the original and imputed data with and without outlier identification
* Discuss the investigation of the impact of data imputation techniques and outlier detection techniques on prediction accuracy and model robustness

The train and test sets are split at 31 12 2012 so the training set is 4 years and the test set is 1 year

RMSE is used to compared the results on the training and test sets

**Cross validation**

walk forward cross validation was implemented to fine tune model hyperparameters using a gridsearch. The optimal number of splits was found to be 30, where the model would not be overfitted to the training set, but also include sufficient previous data points to train the model on changes in volatility. A test forecast horizon window was chosen to be 7 days for daily resampled series and 24 hours for hourly resampled series

we use k-fold cross-validation for model tuning, that is, finding the optimal

hyperparameter values that yields a satisfying generalization performance.

Once we have found satisfactory hyperparameter values, we can retrain the model

on the complete training set and obtain a final performance estimate using the

independent test set. The rationale behind fitting a model to the whole training

dataset after k-fold cross-validation is that providing more training samples to a

learning algorithm usually results in a more accurate and robust model.

**Slide 9: {Insert Results and Graphs}**

* Show the results of the experiments, including graphs and tables comparing the performance of the prediction models on different dataset versions and for different hypotheses
* Highlight key observations and insights derived from the results

Insert RMSE for random walk on each hypothesis slide

**Slide 10: Results and Conclusions**

* Summarize the findings, emphasizing the most effective techniques for handling missing data, outliers, and challenges of stock price prediction in the presence of high volatility and extreme events
* Provide insights into the relationships between the two time series and their predictability based on the tested hypotheses
* Discuss the limitations of the study, such as data limitations, model assumptions, and parameter tuning
* Suggest potential avenues for future research, including exploring other prediction models, imputation techniques, or outlier detection methods, as well as investigating different financial markets and assets

**Since forecasting for daily accuracy, information lost or bias introduced by imputation has a significant affect on trend prediction, but will be evident in forecasting**

Squared residuals of a fitted model to a stock price returns series: If there is heteroskedasticity in the squared residuals of a fitted model, it indicates that the model has not adequately captured the changing volatility in the returns series. This can lead to inefficient parameter estimates and incorrect inferences.

It's essential to acknowledge the limitations of forecasting and focus on risk management

Future

Missing values

* Can try Kalman filter
* Time weighted imputation
* Cluster methods
* Drawing the missing return value from a distribution generated from the historical distribution of returns at the same time relative to day/week/month

Outlier detection:

* Local order factor
* autoencoder

Models:

1. Regime-switching models: These models allow for changes in the underlying data-generating process, which can be useful when dealing with non-normal data that exhibits different behaviors in different periods. Examples include Markov-switching models and hidden Markov models.
2. Wavelet-based methods: Wavelet analysis can be used to decompose a time series into different frequency components, which can help capture non-normal features in the data. Wavelet-based methods can be applied to various tasks, such as denoising, forecasting, and volatility estimation.
3. Bayesian methods: Bayesian techniques can be used to model non-normal data by specifying flexible prior distributions and updating them with observed data. Examples include Bayesian hierarchical models, Bayesian state-space models, and Bayesian non-parametric methods like Gaussian process regression.
4. Addressing challenges using existing prediction methods or developing new methods:

* Existing prediction methods, such as ARIMA, GARCH, and LSTM, can be adapted to address these challenges by incorporating additional features or modifying model parameters.
* However, it may be necessary to develop new methods specifically designed for this context, especially when dealing with high-frequency data, extreme events, and non-normal return distributions.

1. Techniques for handling missing data and outliers:

* Data imputation techniques like linear interpolation, rolling mean, or more advanced methods such as K-nearest neighbors imputation can be used to fill in missing data points.
* Outliers can be handled by using robust prediction models, such as robust regression, or incorporating outlier detection methods, like Z-score or IQR-based techniques, into the prediction process.

1. Performance of different prediction methods in the presence of high volatility and extreme events:

* Factors contributing to the performance of prediction methods include the model's ability to capture non-linear relationships, adapt to changing market conditions, and account for heteroskedasticity.
* Machine learning models, such as LSTMs or recurrent neural networks (RNNs), might perform better due to their ability to capture complex patterns and relationships in the data.

1. What role does the granularity of the data play in the effectiveness of stock price prediction methods and the applicability of one-step forecasting?

* Critically examine the impact of data granularity on the performance of various prediction methods and their ability to capture short-term and long-term market dynamics
* Investigate the potential limitations of one-step forecasting when applied to high-frequency or low-frequency data, considering the influence of market microstructure and macroeconomic factors
* Assess the role of data aggregation and resampling techniques in mitigating the challenges posed by different levels of data granularity and improving the performance of stock price prediction methods