# 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

* 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 pre-processing 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
  + 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?
  + 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?
  + How do different preprocessing techniques, such as data imputation or outlier removal, affect the performance of prediction models in volatile stock price data?
  + 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?
  + 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?
  + 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?
  + 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

* Dataset overview:
  + • Source: Unknown (two unnamed price series)
  + • Potential type: Futures data (trading begins early Sunday evening)
  + • Time period: 5 years (2008-2013)
  + • Frequency: 2-minute intervals
* Data characteristics:
  + • 24-hour price data (excluding Saturdays)
  + • Missing values and non-uniform distribution of time gaps
  + • Shorter time gaps more common; longer time gaps less frequent
  + • Presence of outliers (e.g., holidays, market closed days)
  + • Significant number of missing values (not all likely to be holidays)
  + • Wide range of time gaps (Time Series 1: avg. 17.67 min, max. 5,718 min; Time Series 2: avg. 168.24 min, max. 2,838 min)
* Data irregularities:
  + • Skewness in time gap distribution
  + • Long tails indicating significant deviations from expected 2-minute intervals

Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generated

# Slide 4: Exploratory Data Analysis

* Characteristics of the data:
  + • Stationary price returns series
    - Stationarity tests:
      * • Augmented Dickey-Fuller (ADF): p-value = 0, time series stationary
      * • Kwiatkowski-Phillips-Schmidt-Shin (KPSS): p-value = 0.1, time series stationary around deterministic trend
  + • Nonlinear upwards trend
  + • No seasonal components in hourly, daily, or weekly frequencies
  + • Periodic components in daily and weekly frequencies
  + • Non-normal distributions with heteroskedasticity and volatility clustering
* Seasonal decomposition:
  + Time Series 1 (ts1):
    - • Strong correlation between price and trend
    - • Possible seasonality in monthly time frame
    - • Residuals strongly correlated (underlying structure not captured)
  + Time Series 2 (ts2):
    - • More likely to be seasonal at higher timescales
    - • Strong upward trend across all timescales
    - • Residuals strongly correlated (underlying structure not captured)
* Periodogram analysis:
  + • Both time series show seasonality or periods in intra-day timescale and daily (2 hours, 16 hours, 48 hours)
  + • Significant lags in hourly and daily frequencies indicate serial correlation

Daily/Weekly

Chart

Description automatically generated Chart, line chart

Description automatically generated

Chart

Description automatically generated Diagram

Description automatically generated

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

Chart, histogram

Description automatically generated Diagram

Description automatically generated

# Slide 3: Data Description

* Key observations:
  + • Extreme market events visually identified
  + • Outliers: Require further analysis to determine nature (measurement error or actual stock behavior)
  + • Noise assumed to be inherent in data (not due to measurement error)

Chart

Description automatically generatedChart

Description automatically generated

* Logarithmic returns: not normally distributed (Kolmogorov-Smirnov test p-value = 0.0)
* Skewness and kurtosis: evidence of non-normality • ts1: skewness ≈ 0, kurtosis = 138.45 • ts2: skewness = 0.99, kurtosis significantly different from 3
* Mean values close to zero, variances small but different between ts1 and ts2
* Covariance between ts1 and ts2: negative, but small magnitude
* Non-normal behavior: heavy tails, high likelihood of extreme values
* Rolling mean and standard deviation (window size: 4 hours)
* Squared and absolute returns: periods of high volatility
* Heteroskedasticity present in time series
* Chart

  Description automatically generatedChart

  Description automatically generated

# Outlier detection:

extreme market events (e.g., 2008) and sensitivity analysis for model stability and prediction improvement

Chart

Description automatically generatedChart, histogram

Description automatically generated

Chart

Description automatically generatedChart, histogram

Description automatically generated

In general, the results suggest that the time series are stationary and do not exhibit any trend or structural change over time.

* Non-normality: consider alternative methods to filter extreme value outliers

# Slide: Modelling Outline

* Research goal: Develop suitable predictions for price returns, accounting for volatility, extreme events, and missing values
* Hypotheses:
  1. • H1: ts1 can predict ts1
  2. • H2: ts2 can predict ts2
  3. • H3: ts1 can predict ts2
  4. • H4: ts2 can predict ts1
* Motivation: Investigate relationships between time series and predictability; impact of imputation and feature engineering on uncertainty
* Experiments (4 data preparation methods):
  1. Original data, missing values removed, outliers not identified
  2. Original data, missing values removed, outliers identified
  3. Imputed data, outliers not identified
  4. Imputed data, outliers identified

# Comparing model performance:

• Assess models on different dataset versions (original, imputed, with and without outlier identification)

• Investigate the impact of data imputation and outlier detection techniques on prediction accuracy and model robustness

Considerations:

• Choice of imputation technique: preservation of data structure, noise levels, model performance

• Accounting for extreme market conditions and black swan events

• Testing volatility-based models and deep learning models (e.g., LSTM) for robustness to outliers

* Imputation technique considerations: preservation of data structure and noise, impact on model performance
* Importance of accounting for extreme market conditions and black swan events

# Slide: Methodology - Data Preparation

* Pre-processing steps:
* Imputation techniques considered:
  + • Mean imputation
  + • Linear interpolation
  + • Median interpolation
* Imputation strategy:
  + • Different methods for short gaps (< 4 hours) and long gaps
  + • Preserve time series structure and stability
  + • Rolling window of 4 hours for moving mean imputation
* Data used:
  + • Original data including Saturdays
* Handling gaps for machine learning models:
  + • Remove gaps equal to or longer than a day
  + • Assess forecasting results to gain insights
* Linear interpolation:
  + • Least sensitive to underlying volatility
  + • No artificial volatility introduced (unlike mean and median)
* Outlier detection:
  + • Based on squared price returns series
  + • Methods compared: Peak over Threshold and DBSCAN clustering
  + • DBSCAN chosen: identifies outliers by cluster size relative to neighbors
* Outlier observations:
  + • Higher number of outliers in first series
  + • Smaller number of common outliers, suggesting stock-specific factors have significant impact

# Slide: Data Resampling

* Rationale for resampling:
  + • Capture potential daily and weekly periodicities
  + • Reduce uncertainty in true value due to missing values clustering in shorter time frames
  + • Improve computational efficiency
* Resampling methods:
  + • Resampled to hourly and daily frequencies
  + • Used mean of the original data for resampling

# Slide 5: Benchmark Model

* Random walk simulation as a benchmark:
  + • Simple assumption of price series behaving like a random walk
  + • Used for both univariate (H1H2) and multivariate (H3H4) cases
* Simulating unpredictability of stock prices:
  + • Applied using the mean and standard deviation of the returns price series
  + • Captures short-term unpredictability in stock prices
* Importance of comparison:
  + • Helps determine if chosen models or modelling approach is appropriate for prediction
  + • Indicates if series is inherently random or has underlying structure/pattern for forecasting purposes
  + • Compare model performance against random walk RMSE

# Slide 7: Prediction Models

* Chosen prediction models:
* • KNN clustering
* • Random Forest
* • Decision Tree
* • Gradient Boosting Tree
* • Light Gradient Boosting Tree
* • Gaussian Process
* Key characteristics and rationale:
  + • Non-linearity and fat-tailed distribution of price returns
  + • Traditional models unable to account for heteroskedasticity (changing volatility over time)
  + • Classical machine learning regression methods as a first approach
* Applying models to test hypotheses:
  + • Models will be applied on four versions of the dataset
  + • Evaluating performance and suitability for predicting price returns

# Slide 8: Model Evaluation

* Evaluation metrics:
  + • RMSE: Root Mean Squared Error
  + • Significance: Assess prediction accuracy and model performance
* Train and test sets:
  + • Split at 31-12-2012 (4 years of training data, 1 year of test data)
* Cross-validation:
  + • Walk forward cross-validation with grid search
  + • Optimal number of splits: 30
  + • Test forecast horizon: 7 days for daily resampled series, 24 hours for hourly resampled series
  + • Prevents overfitting while including sufficient data for training
  + • K-fold cross-validation for model tuning and hyperparameter optimization
  + • Retrain the model on the complete training set after k-fold cross-validation

# Slide 11: Discussion

* Limitations of the study:
  + • Data limitations
  + • Model assumptions
  + • Parameter tuning
* Focus on risk management in forecasting

Future Research Avenues:

1. Missing values:
   1. • Kalman filter
   2. • Time-weighted imputation
   3. • Cluster methods
   4. • Distribution-based imputation
2. Outlier detection:
   1. • Local order factor
   2. • Autoencoder
3. Explore other prediction models:
   1. • Regime-switching models (Markov-switching, hidden Markov models)
   2. • Wavelet-based methods
   3. • Bayesian methods (Bayesian hierarchical models, Bayesian state-space models, Gaussian process regression)
4. Adapt existing prediction methods or develop new ones for high-frequency data, extreme events, and non-normal return distributions
5. Investigate the role of data granularity:
   1. • Impact on prediction method performance
   2. • Limitations of one-step forecasting for high-frequency or low-frequency data