Guess-Timate Gang: Proposal and Progress

Proposal for Modeling Approach at the Erdos Institute – Stock prices

Introduction: We are excited to present a comprehensive modeling approach aimed at predicting stock prices, leveraging various time-series forecasting techniques. Our team has utilized sophisticated models such as ARIMA, double exponential smoothing, and ensemble methods to enhance the accuracy and reliability of stock price predictions.

Key Predictors and Outcomes:

1. Predictors:
    ○ Historical stock prices (Opening, Closing, High, Low)
    ○ Time series features
    ○ Economic indicators and external factors
2. Outcomes:
    ○ Closing prices for future time points
    ○ Price differentials (Closing - Opening)

Modeling Techniques:

1. ARIMA (AutoRegressive Integrated Moving Average):
    ○ ARIMA is a powerful time-series forecasting model that captures the temporal dependencies in stock price data.
    ○ It considers the auto-regressive, integrated, and moving average components to make predictions.
2. Double Exponential Smoothing:
    ○ Double Exponential Smoothing, or Holt's method, is utilized to capture trends and seasonality in stock prices.
    ○ The method incorporates both level and trend smoothing to enhance the accuracy of predictions.
3. Ensemble Methods:
    ○ We have employed ensemble methods, combining the predictions from multiple models to mitigate individual model weaknesses.
    ○ Boosting and bagging techniques are implemented to create a robust ensemble that can adapt to different market conditions.

Methodological Approach:

1. Data Preprocessing:
    ○ Historical stock price data is cleaned and preprocessed to handle missing values and outliers.
    ○ Feature engineering is performed to extract relevant information from the time series.
2. Model Training:
    ○ The ARIMA model is trained on historical data to capture the underlying patterns and temporal dependencies.
    ○ Double Exponential Smoothing is optimized for smoothing parameters to enhance its forecasting capabilities.

- ○ Ensemble models are constructed using boosting and bagging to combine the strengths of multiple forecasting techniques.
3. Validation and Evaluation:
   - ○ The models are rigorously validated using time-series cross-validation techniques to ensure robust performance.
   - ○ Evaluation metrics such as root mean square error (RMSE) are employed to quantify the accuracy of predictions.

Expected Impact: The proposed modeling approach aims to provide accurate and reliable predictions of stock prices, aiding investors and financial analysts in making informed decisions. By combining the strengths of various modeling techniques, our approach seeks to enhance forecasting accuracy and contribute to the understanding of market dynamics.

We look forward to the opportunity to collaborate with the Erdos Institute, leveraging our expertise in time-series forecasting and predictive modeling to contribute valuable insights to the financial domain.

Thank you for considering our proposal.

Sincerely,

Guestimate Gang

Guess-Timate Gang: Proposal and Progress

Progress Report (Nov 16)

We have applied a variety of forecasting methods to analyse historical S&P 500 data. The project's primary objectives were to build a pipeline to predict future S&P price (and possibly to other indices and stocks).

Data: https://www.nasdaq.com/market-activity/index/spx/historical?page=254&rows_per_page=10&timeline=y10

Data Preprocessing: Clean the data by handling missing values on holidays because there is no trading on these days; Reorganise the data; Convert date to a timestamp, time series format; Rename the column name for convenience; Define new variables based on need and intuition

Baseline Models: We calculated the average cross-validation root mean squared error for the 4 baseline models average, naive, tend and random walk with drift.

Methods Applied: Moving average, (Double) Exponential smoothing, ARIMA, Ensemble model such as Gradient Boosting, XGBoost, Simple Ensemble ARIMA

Parameter Selection: For some methods, we tuned the hyperparameters either by hand or by an optimization routine to improve the model performance.

Model Comparison: We use the RMSE (root mean squared error) of the test data and the predicted data to measure a model's performance and to compare different models.

Current Challenges: During the course of the project, we encountered a few challenges:
- It is hard to understand how good of a prediction is good enough, i.e. in our measure, how small the RMSE should be?
- How can we merge or massage the currently used methods to increase our performance, i.e. decrease the RMSE?