CSCI E109B PROJECT – S&P 500 PREDICTIONS

Pluto, Venkatesh, Artem, Dhimant – Spring’2023

Project Group # 25

AGENDA

Problem Statement

Visualization/EDA/Analysis

Modeling

Training details: Time taken, number of epochs, batch size, learning rate etc.

Results

Conclusion/ Inferences

Future Work/ Scope of improvement

PROBLEM STATEMENT

Problem statement

- Can ML, Probabilistic or DL models predict S&P 500 accurately enough to outperform stock indexes?

History

- Not too distant in the past, the financial firms were using faster servers, network speed and latency as a competitive advantage in beating the stock market indexes.
- With the advancement of technology and data science, financial firms are analyzing and learning more about using ML and deep learning (DL) techniques to have sustained advantage.

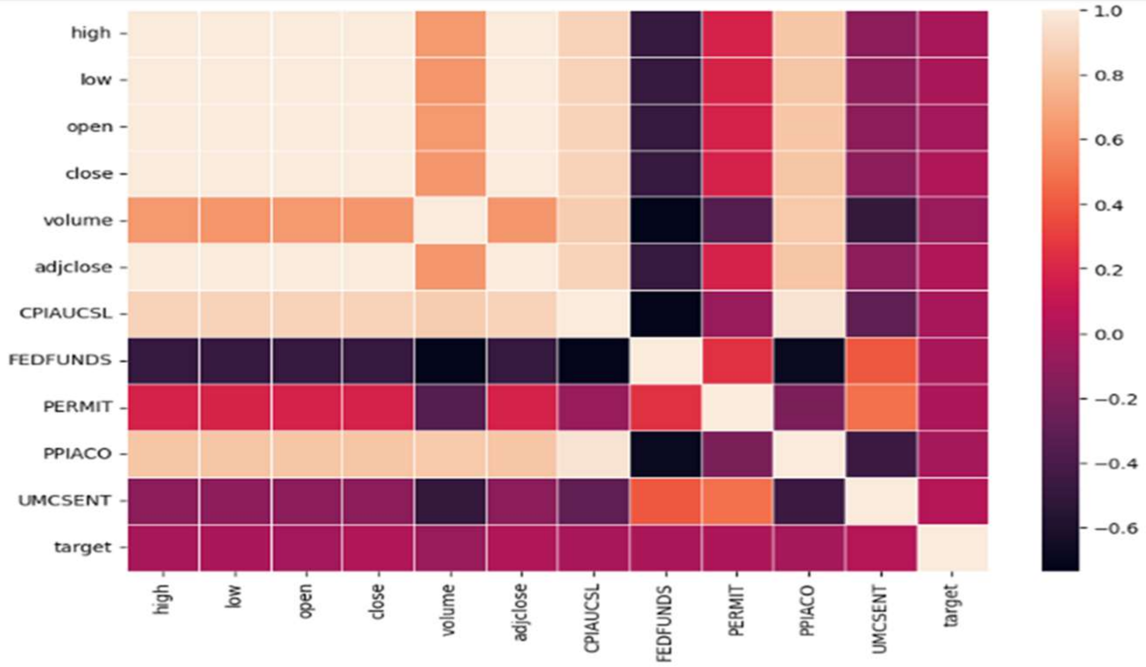
Measurement

- Our measurement of success is to predict if the opening price for the next month, high/low, w.r.t. opening price of the current month given the economic indicators

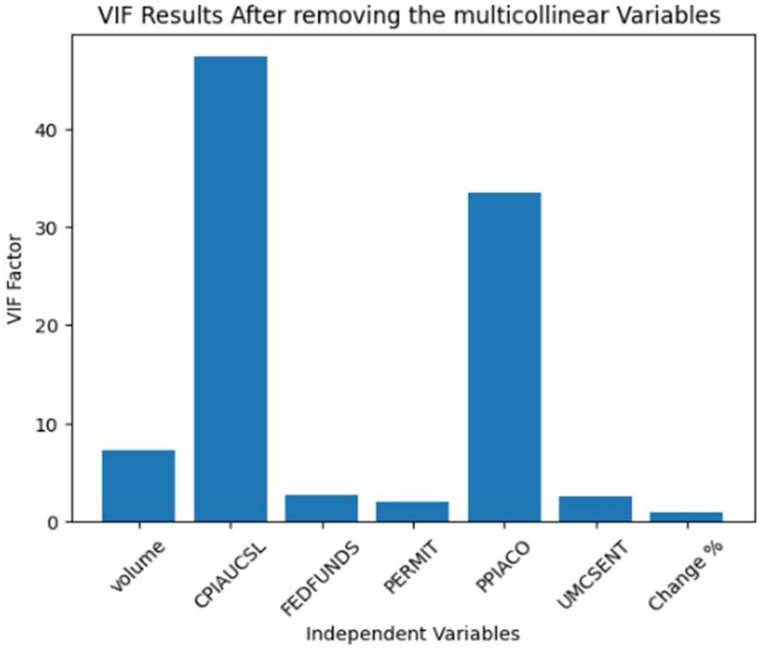
Note: Opening price of the next day is same as close price of the previous day on most days.

EDA - MultiCollinearity Test

- Correlation heat map reveals that PERMIT, UMSCENT, FEDFUNDS,PPIACO, CPIAUCSL, VOLUME, High, Open, adjclose, and low are having relationship with close.
- High, Open, low and adjclose are same as close variables and identified as multicollinear variables.
- VIF Plot shows low for PERMIT, FEDFUNDS, UMSCENT, volume, Change % after removing multicollinear variables
- Although CPIAUCSL,PPICAO shows multicollinearity, these multicollinearity will not impact the deep learning models.

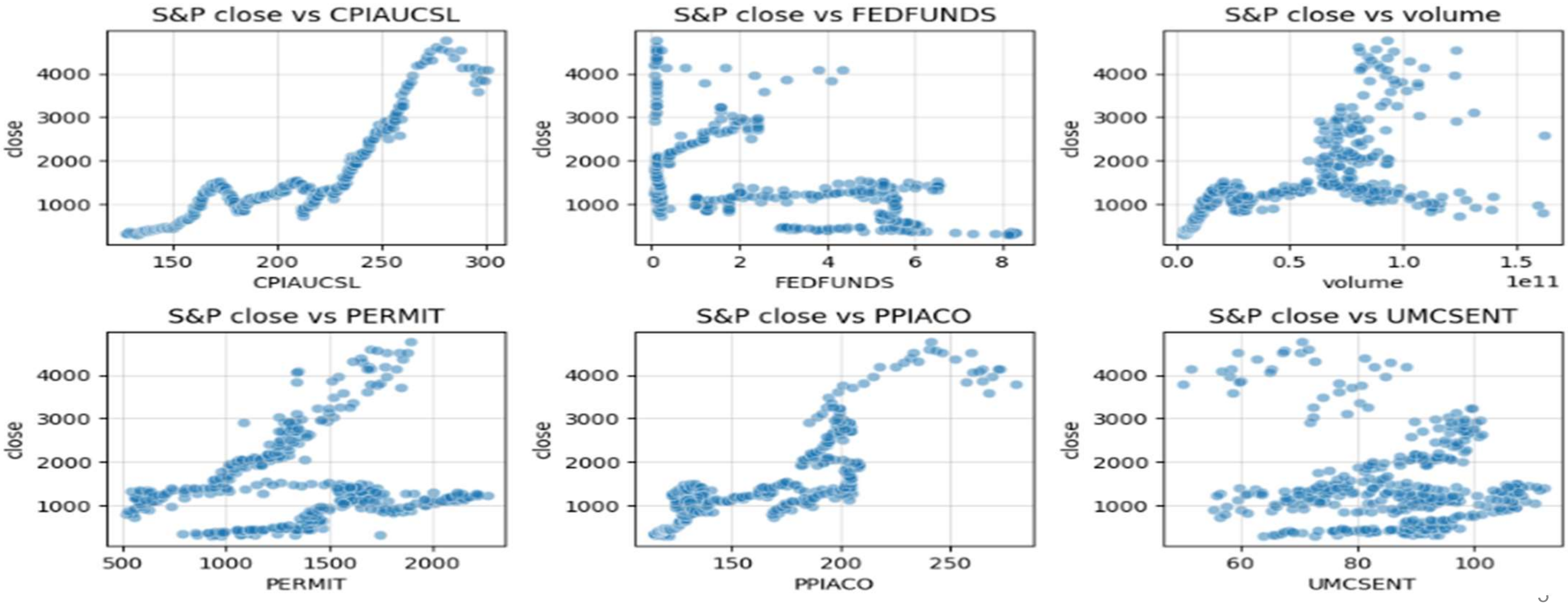


Correlation Matrix



EDA - Trend Analysis

- Strong linear trends observed between the close and CPIAUCSL, FEDFUNDS, PERMIT, PPIACO, UMCSENT.
- No potential Outliers identified in the data set.
- No missing data observed in the dataset.
- Converted time series data of 30 years (monthly) data into input-output sequence of 12 time steps.
- Splitted the 12 time steps data into training and testing for modeling.



S&P close vs CPIAUCSL

S&P close vs FEDFUNDS

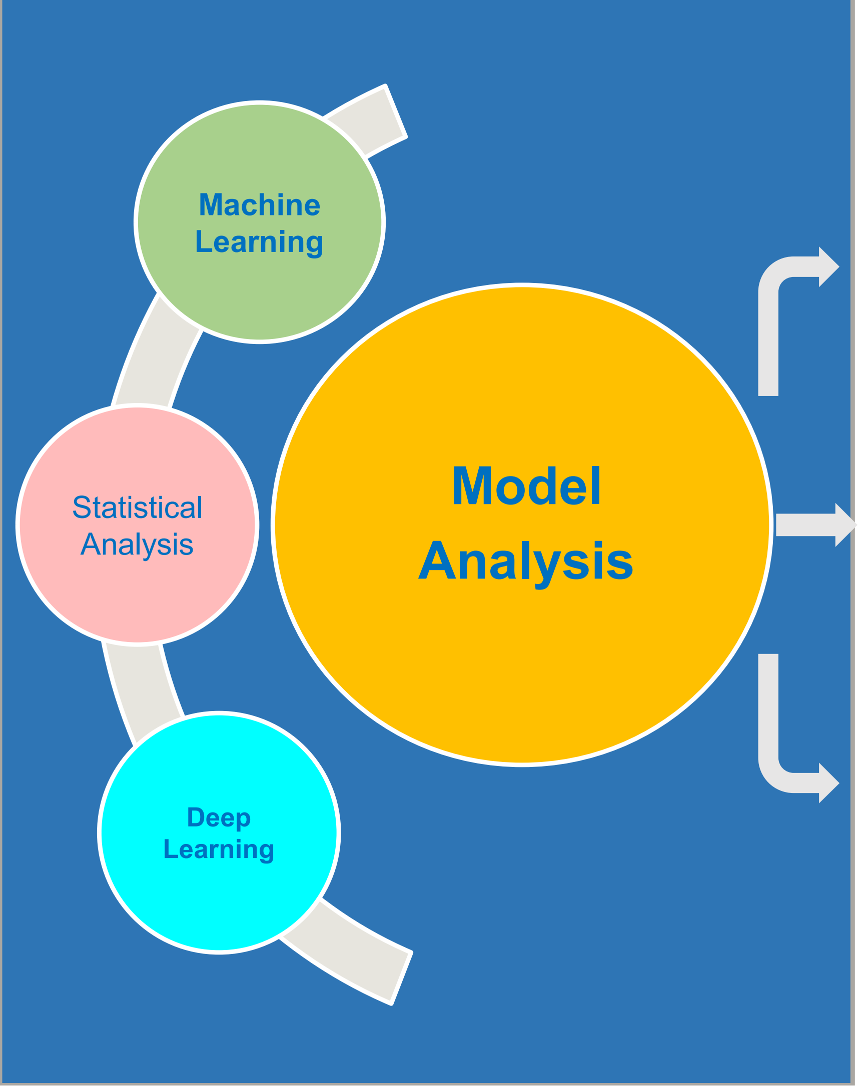
S&P close vs volume

S&P close vs PERMIT

S&P close vs PPIACO

S&P close vs UMCSENT

MODEL SELECTION RATIONALE



- **Machine Learning: Random Forest**
  - Ensemble method: Combines multiple decision trees, reducing variance.
  - Handling non-linearities: Captures complex patterns in time series data without requiring transformation.
  - Robust to outliers: Insensitive to noise, leading to more stable forecasts.
  - Feature importance: Identifies relevant variables, allowing for better understanding of relationships in data.
- **Statistical Analysis: Bayesian model**
  - Probabilistic approach: Quantifies uncertainty in forecasts, useful for risk assessment.
  - Incorporation of prior knowledge: Improves accuracy with limited/noisy data.
  - Model updating: Easily adapts to new data for more relevant predictions.
- **Deep Learning – FNN, RNN, LSTM, GRU and Transformer**
  - Special emphasis on Recurrent Neural Networks (RNNs) and Transformer models…
  - …as they are particularly well-suited for time series forecasting tasks.
  - RNNs are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps.
  - Transformer models leverage self-attention mechanisms to weigh the importance of different time steps in the input sequence…
  - …which has demonstrated exceptional performance in various seq-to-seq tasks.

TRAINING
Deep Learning

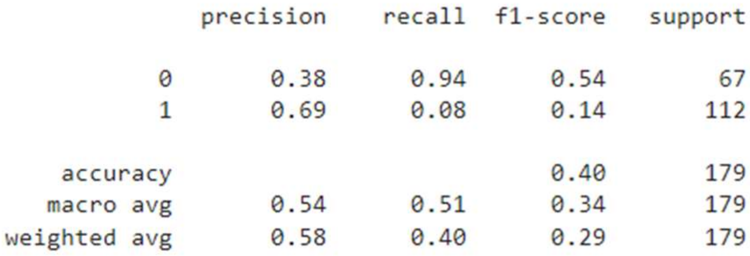
- **Deep learning models - key hyperparameters:**
  - *Sequence Length:* 12 time steps for prediction.
  - *Recurrent Layer Units:* 50 neurons for complex patterns.
  - *Activation Functions:* 'relu' & 'sigmoid' for non-linearity.
  - *Loss Function:* 'binary_crossentropy' to measure performance.
  - *Optimizer:* 'adam' for efficient weight updates.
  - *Epochs:* 50 for training, watch for overfitting.
  - *Batch Size:* 32 to balance speed & convergence.

- **Transformer Model:**
  - *Attention Heads: 4 for diverse patterns.*
  - *Key Dimension: 32 for self-attention capacity.*

RESULTS

Machine learning:

- Random Forest – walk-forward



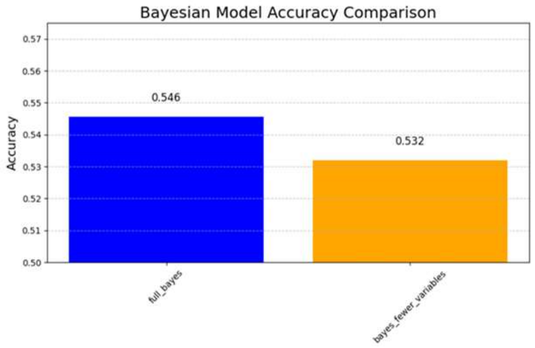
```
              precision    recall  f1-score   support

           0       0.38      0.94      0.54        67
           1       0.69      0.08      0.14       112

    accuracy                           0.40       179
   macro avg       0.54      0.51      0.34       179
weighted avg       0.58      0.40      0.29       179
```

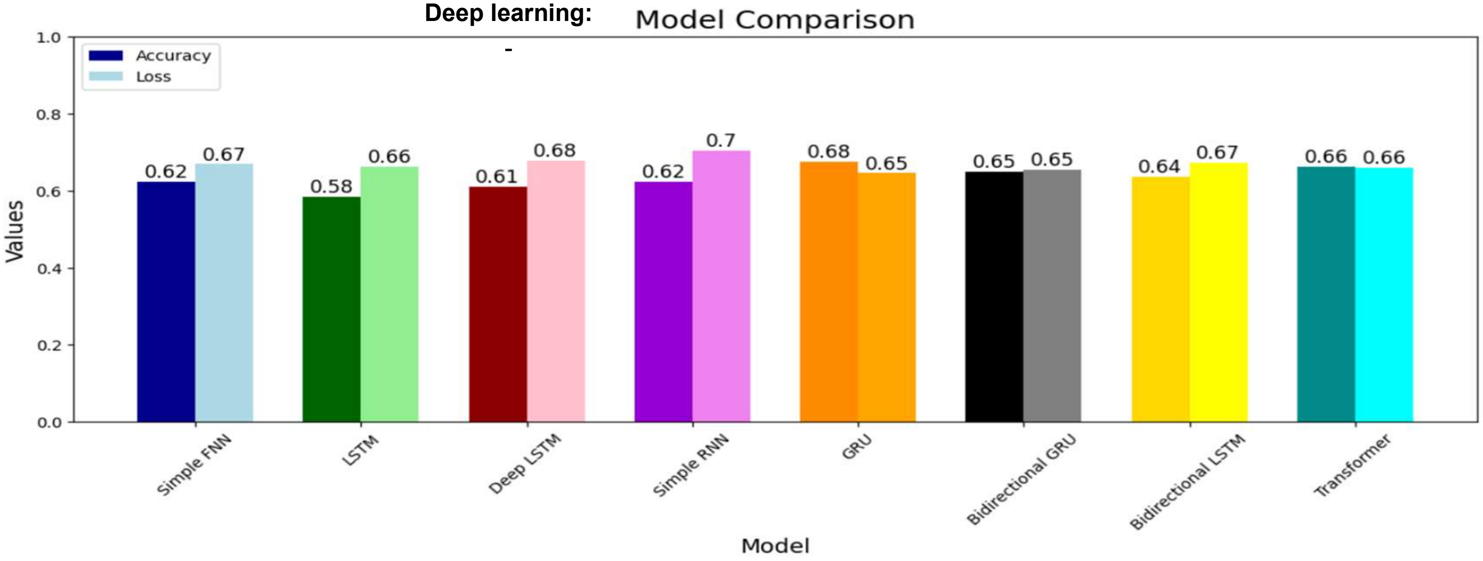
Statistical analysis:

- Bayesian model

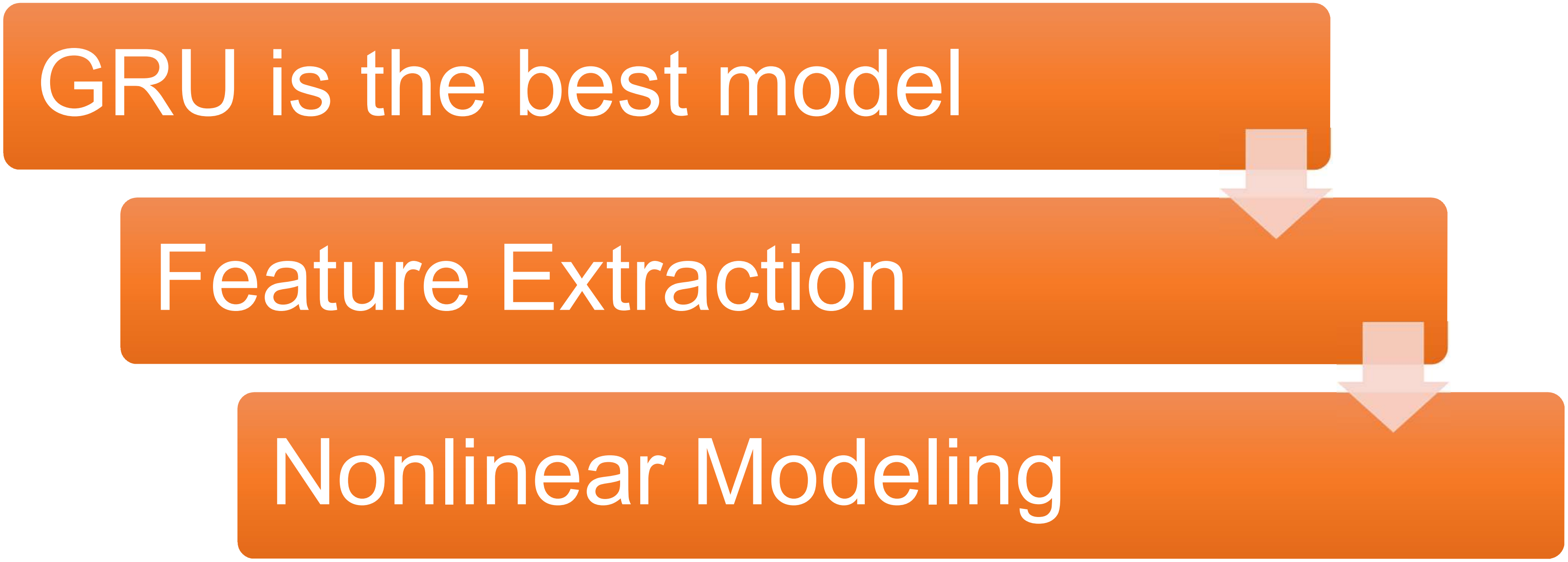


Deep learning:

-



CONCLUSION/INFERENCES



GRU is the best model

Feature Extraction

Nonlinear Modeling

WHERE DO WE GO FROM HERE?

- Incorporate S&P 500 individual company’s performance indicators
- Sentiment analysis – social media, e.g, tweets, Facebook messages, news messages…

Ideally, we would like to be able to predict opening price on the next business day which may include indicators from around the globe