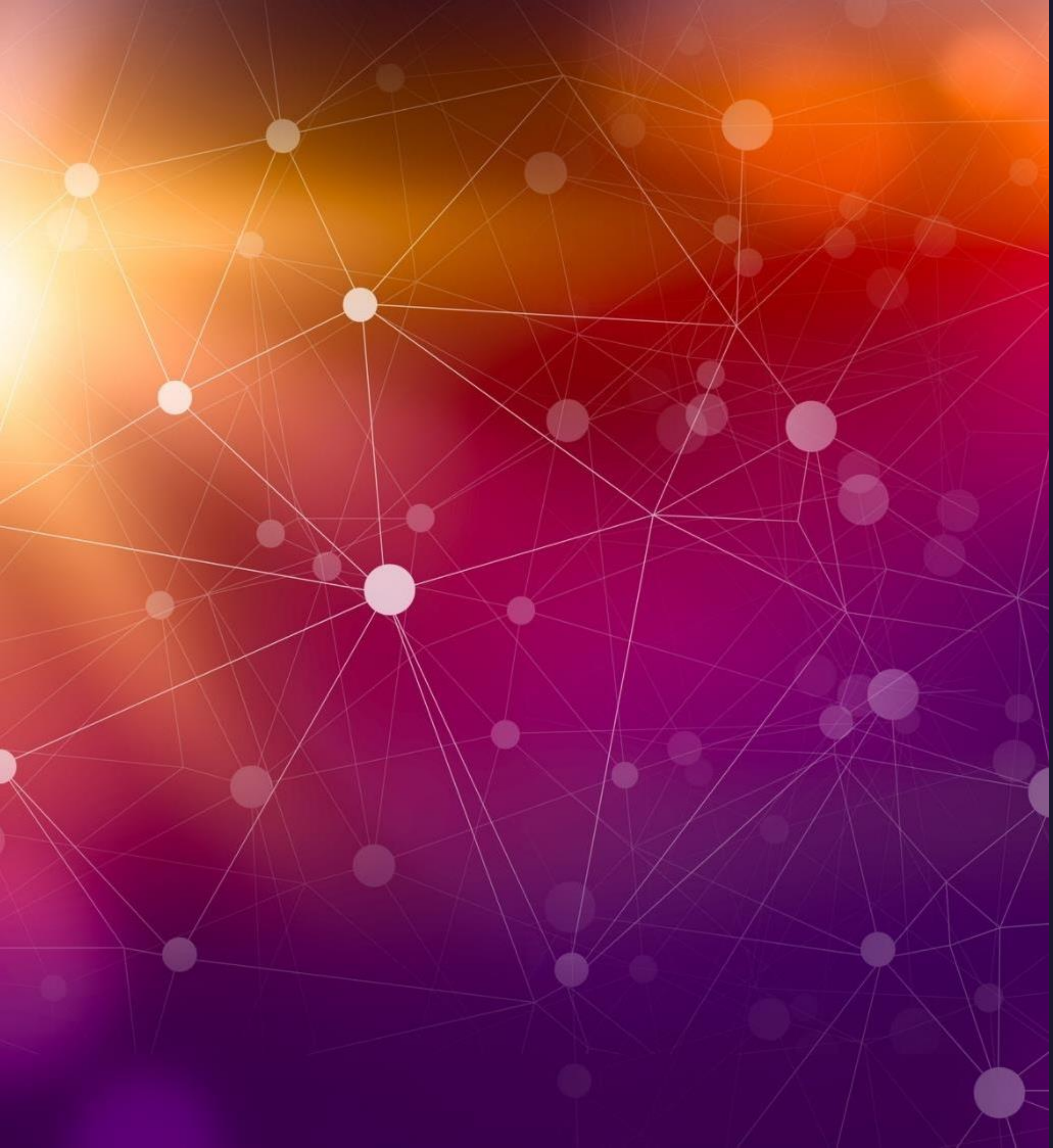


Stock Prediction Using News Headlines

Agenda

Problem Definition

Dataset Description

Project Overview

Algorithmic Solutions

Performance vs S&P 500

Data Leakage Lessons Learned

Future Research Suggestions

Problem Definition

- Can we build a machine learning model that can use natural language processing on news headlines to predict stock price movement?

Dataset Description

Headline data[†] and stock price data[*]

Jan 1, 2010 – Dec 31, 2016

Merged prices with headlines;

For each stock on each day:

- Previous day close
- Trading day close
- Previous days headline

[†] https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests

• https://www.kaggle.com/datasets/mosesmoncy/ny-stock-exchange-size-dataset

Project Overview

- Natural Language Processing
- Convert headlines into high-dimension vectors (one dimension per vocabulary word): TF-IDF Vectorizing
- Develop classification models to learn class labels based on finding separation between vectors
- Test model with unseen headlines to determine ability to generalize on new data

Algorithmic Solutions

Data Pre-processing

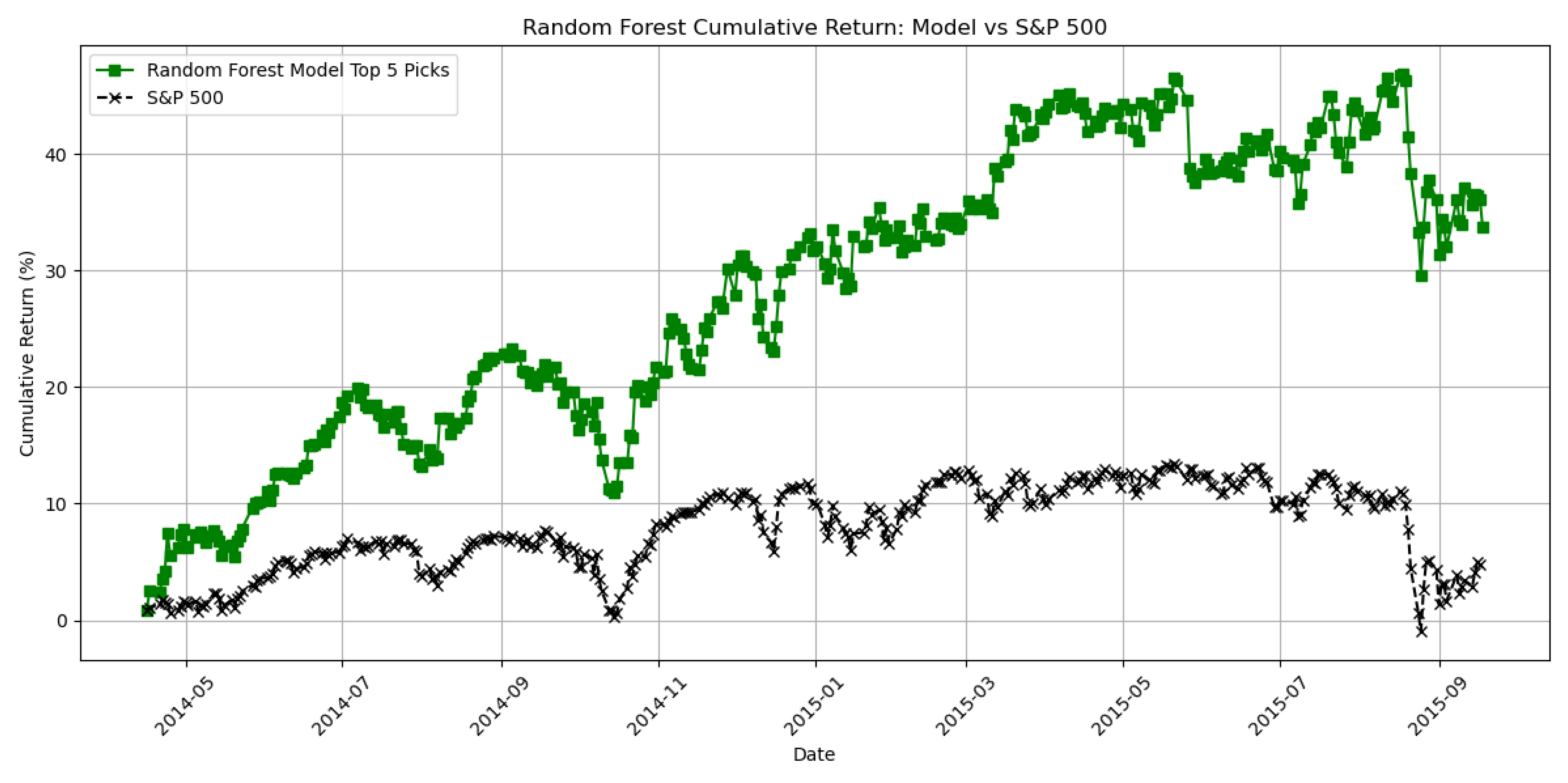
- Converted all text to lowercase; removed stopwords and punctuation, and lemmatize
- Referenced only the prior day headlines (earlier headlines were likely already priced into stock)
- Dropped rows with no headline information
- Considered adding quantitative features but modeling revealed no added predictive value from parameters such as weekly volatility and over/under priced versus a 52-week moving average

Models evaluated:

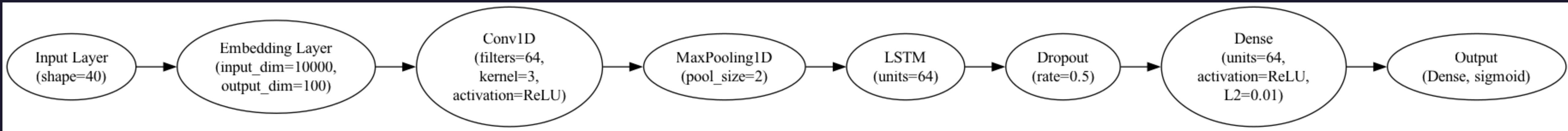
- Random Forest, Logistic Regression, XGBoost Classifier, Support Vector Machine, and various Neural Network architectures
- XGBoost, SVM, and Logistic Regression performed worse than Random Forest and Neural Networks
- Used Random Forest and Neural Network as base models and Logistic Regression as meta-model in stacking
- Although overall F1 accuracy scores did not suggest good accuracy on all stocks, the models did not need to get all predictions correct; instead, the strategy involved selecting the most confident picks and then evaluate cumulative returns based on daily average returns
  - Prioritizes recall: % of True UP that model predicts

Performance Summary

Selected models compared to cumulative returns of S&P 500


Random Forest Cumulative Return: Model vs S&P 500
Random Forest Model Top 5 Picks
S&P 500
Cumulative Return (%)
0
10
20
30
40
2014-05
2014-07
2014-09
2014-11
2015-01
2015-03
2015-05
2015-07
2015-09
Date

Hybrid CNN-LSTM Classifier Architecture



Input Layer (shape=40)

Embedding Layer (input_dim=10000, output_dim=100)

Conv1D (filters=64, kernel=3, activation=ReLU)

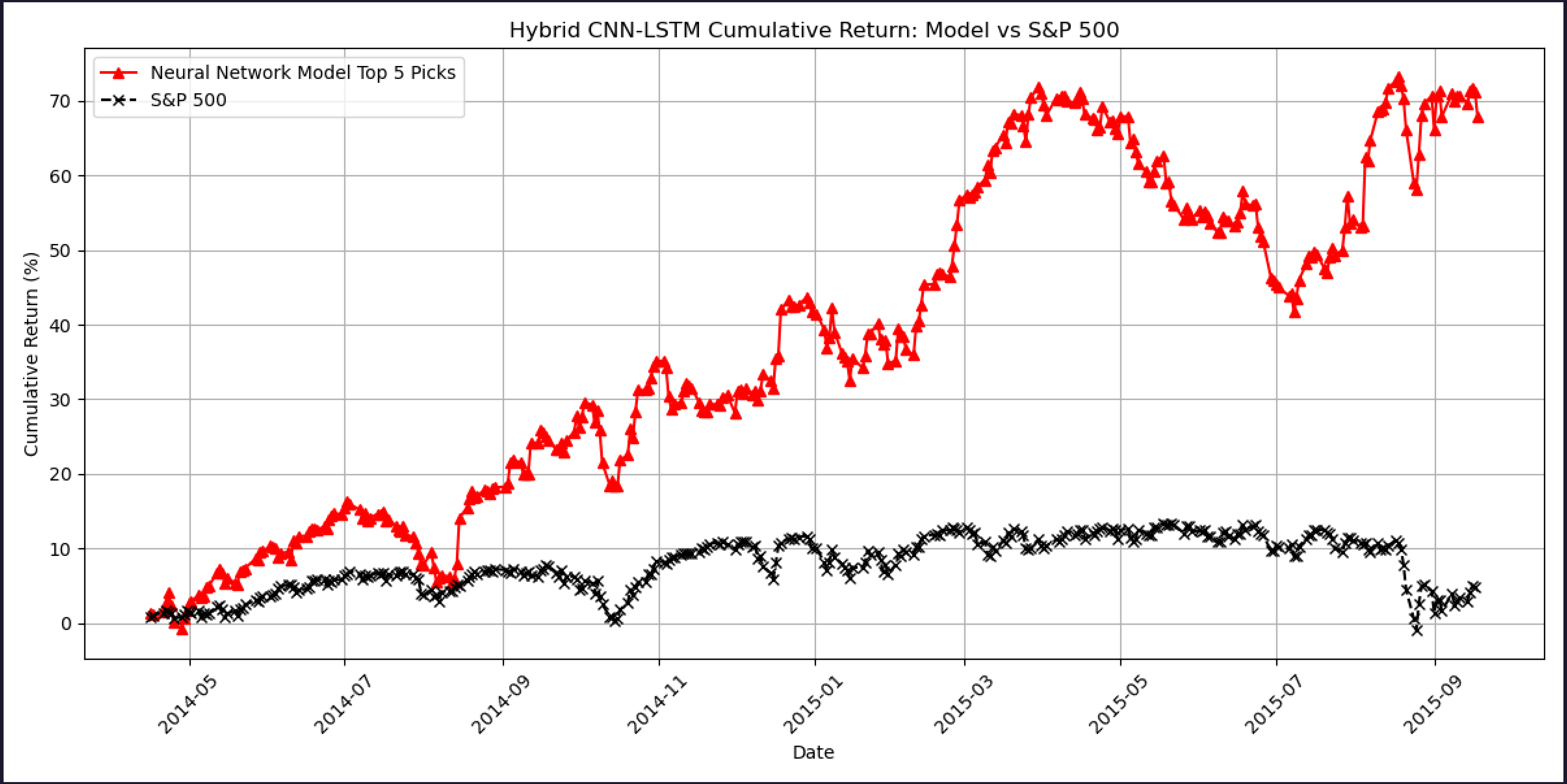
MaxPooling1D (pool_size=2)

LSTM (units=64)

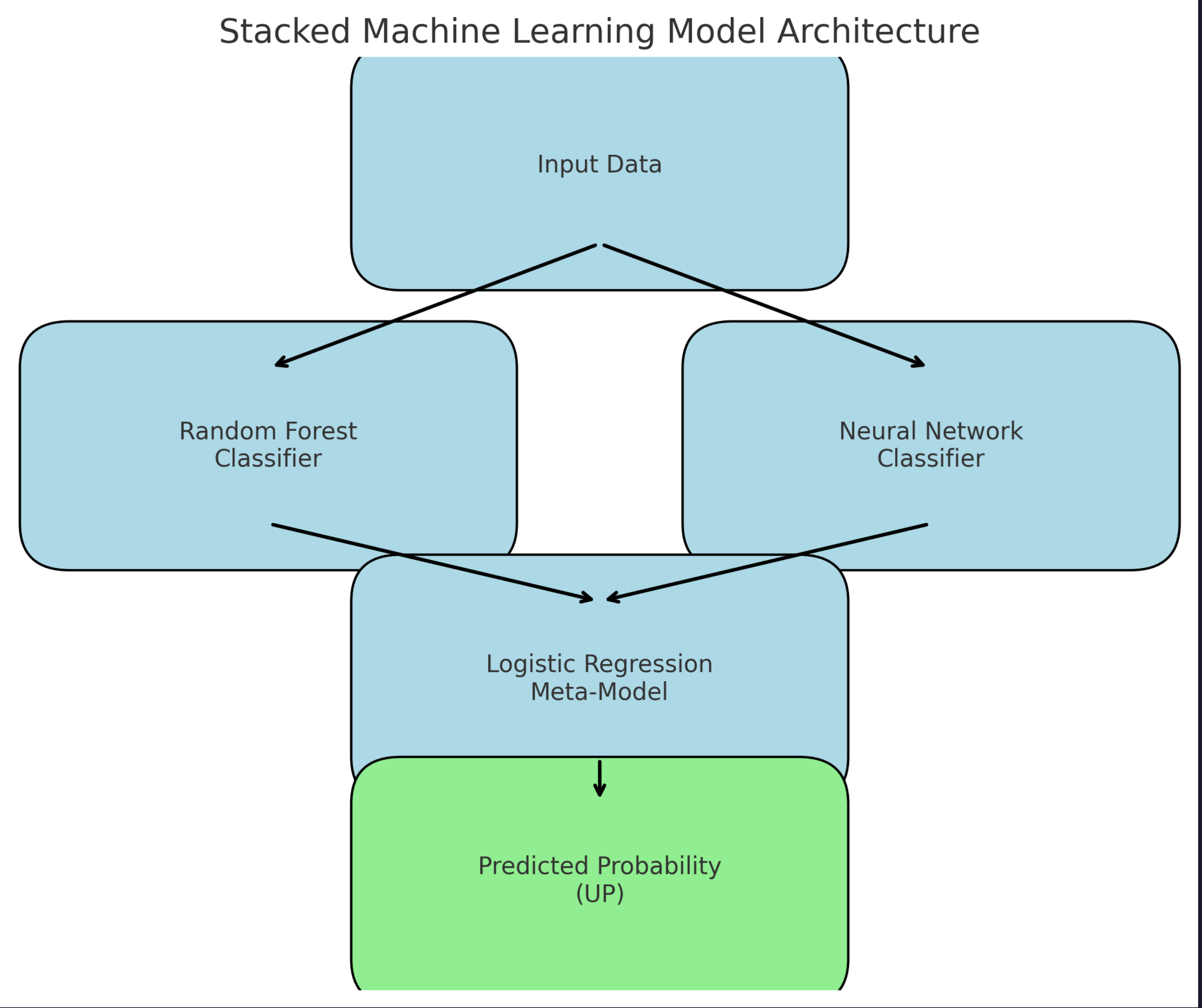
Dropout (rate=0.5)

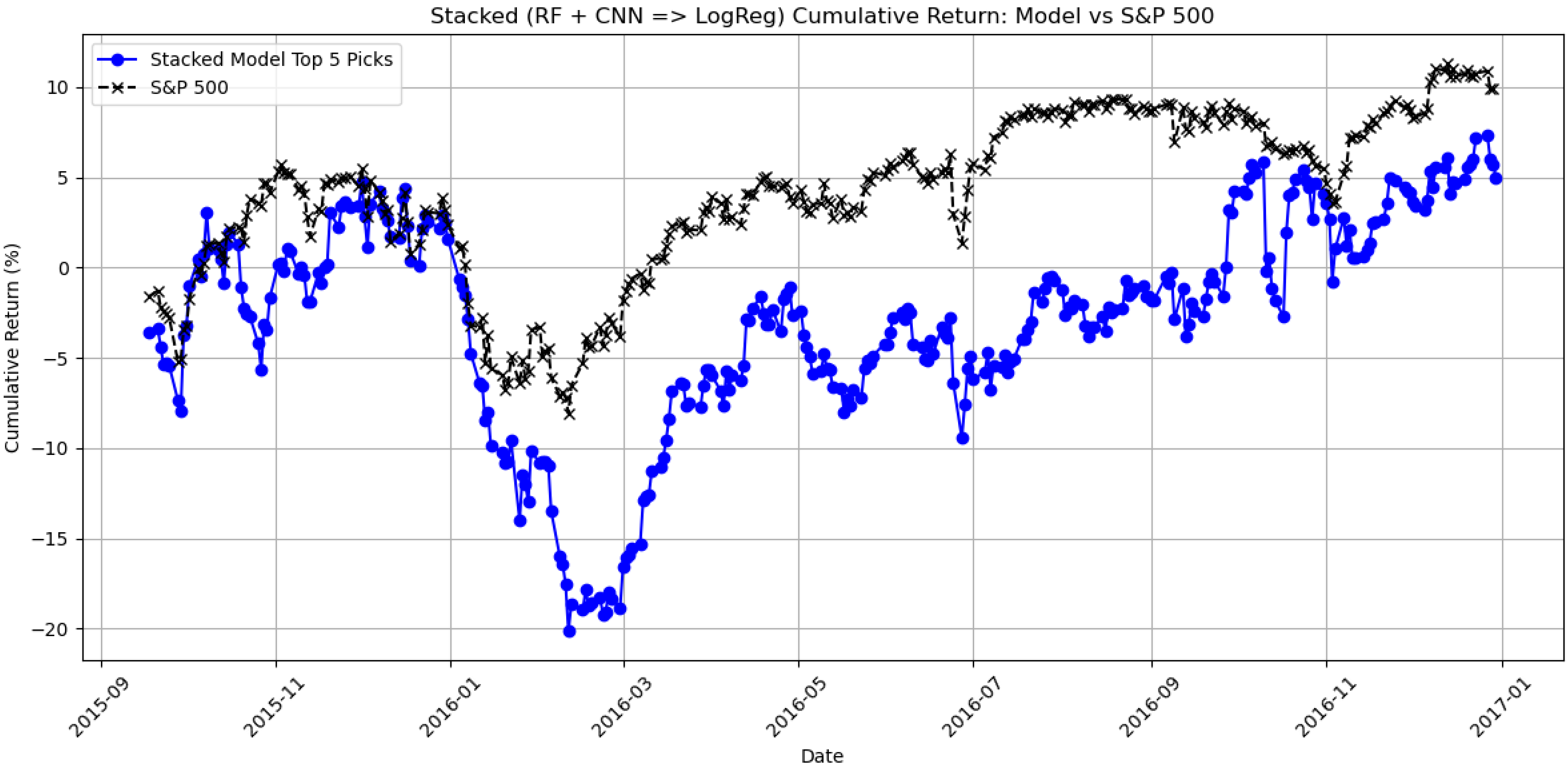
Dense (units=64, activation=ReLU, L2=0.01)

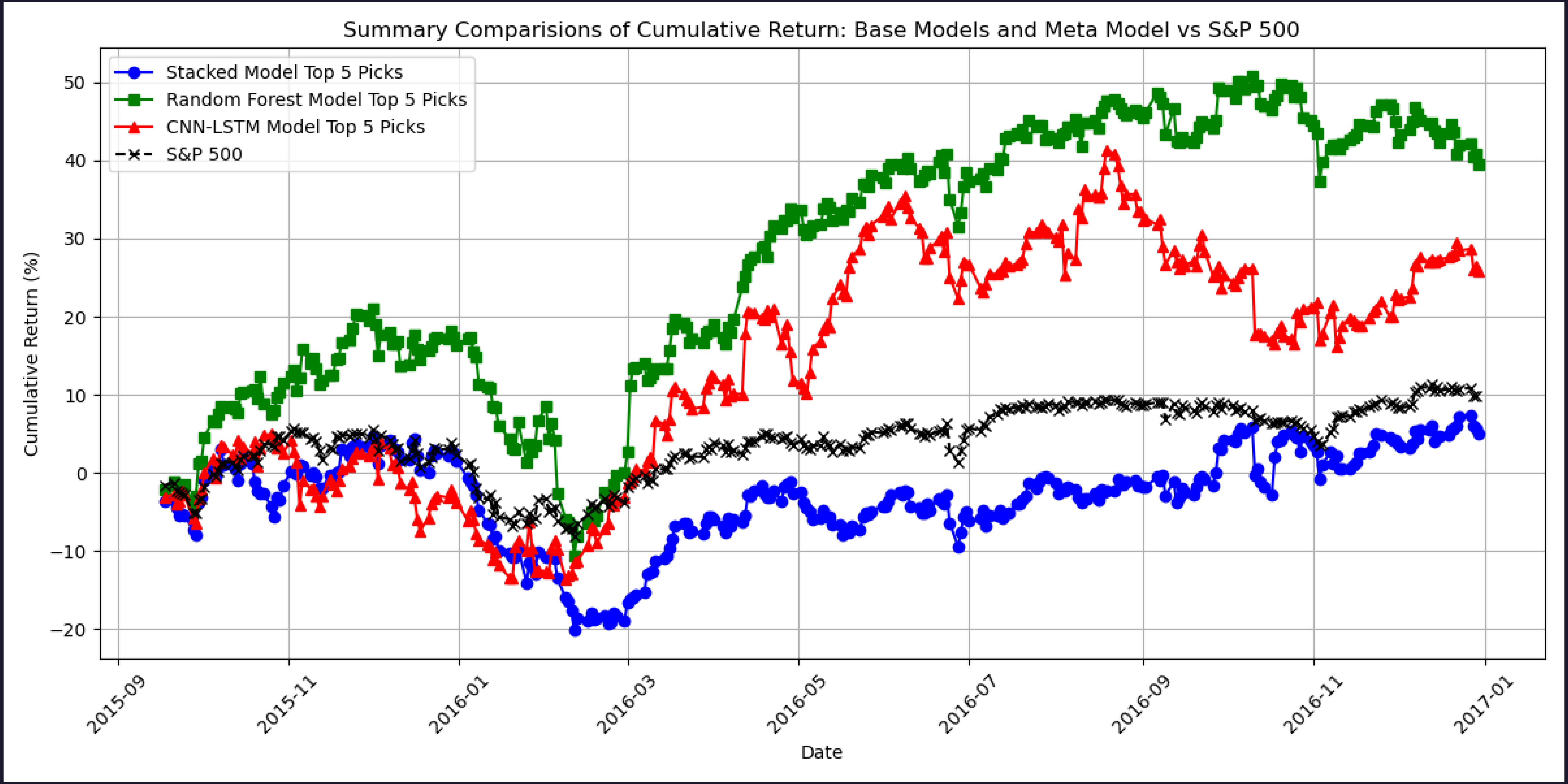
Output (Dense, sigmoid)


Hybrid CNN-LSTM Cumulative Return: Model vs S&P 500
Neural Network Model Top 5 Picks
S&P 500
Cumulative Return (%)
0
10
20
30
40
50
60
70
Date
2014-05
2014-07
2014-09
2014-11
2015-01
2015-03
2015-05
2015-07
2015-09

Stacked Model Structure


Stacked Machine Learning Model Architecture
Input Data
Random Forest Classifier
Neural Network Classifier
Logistic Regression Meta-Model
Predicted Probability (UP)


Stacked (RF + CNN => LogReg) Cumulative Return: Model vs S&P 500
Stacked Model Top 5 Picks
S&P 500
Cumulative Return (%)
10
5
0
−5
−10
−15
−20
Date
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01


Summary Comparisions of Cumulative Return: Base Models and Meta Model vs S&P 500
Stacked Model Top 5 Picks
Random Forest Model Top 5 Picks
CNN-LSTM Model Top 5 Picks
S&P 500
Cumulative Return (%)
50
40
30
20
10
0
−10
−20
Date
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01

Conclusion

- First, predicting stock price changes is extremely challenging!
  - Price reflects buying behavior
  - Buying behavior = macro and micro market factors/forces + human intuition, logic, irrationality
  - Predictions of this complex cognitive and behavioral pipeline that use weak/noisy headline signals has low accuracy
  - Base models outperformed S&P 500; further exploration with more recent data is needed
  - Stacked model that combines predictions of both base models performed worse than either bas model alone
- Second, the benefit of a buy/hold strategy with an index fund is underappreciated
  - Performance charts understate the S&P500 performance by comparing the compounded gains the same as our stick picking model
  - An S&P 500 index fund with 0.05% annual fee would have returned 125%* over the same period as this analysis
- Finally, ML and data science fundamentals are critical – data leakage can be difficult to spot and must be accounted for early and often

* incl. Dividends; 96% w/o dividends

Future Research

- Need more data!
- NLP w/ headlines might one of many features to model
- Other context-aware NLP methods
- Event-driven modeling
- Sector-specific context and risk
- Quantifiable stock valuation measurements
- SEC Filings
- Earnings calls tone
- Pipeline assessments

Thank you