Improving models for predicting trading volume on the Russian stock market

Bakhyshov Vakhid [*1], Dugin Maxim [†1], Samoilov Roman [‡1], and Chasov Nikita[§1]

[1]Moscow State University

October 29, 2024

Abstract

Данное исследование посвящено моделированию и улучшению моделей для прогнозирования внутридневных временных рядов объемов торгов на Московской фондовой бирже. В качестве методов прогнозирования используются SVWAP, PC-VWAP, а также модели машинного обучения: LSTM, XGBoost и Random Forest.

Эмпирические результаты исследования приводят к следующим выводам: классические методы предсказания временных рядов, предложенные в статье [YYL14], можно существенно улучшить с помощью ансамблевых методов и тщательного подбора гиперпараметров.

Keywords: ARFIMA; ARIMA; principal component; volume series; VWAP; SVWAP and PC-VWAP, LSTM, XGBoost, Random Forest

Introduction

На фондовом рынке трейдеры играют важную роль как одни из ключевых участников торгов, и минимизация торговых издержек является важной задачей как в теории, так и на практике. Существует множество стратегий для снижения затрат, как, например, в статье Обижаевой [Ann13], но большинство из них направлено на снижение рыночного импакта. При выполнении крупных заявок значительный рыночный импакт может привести к неблагоприятным изменениям цены, увеличивая издержки. Чтобы минимизировать такие издержки, используются алгоритмические стратегии, такие как VWAP, цель которых — максимально приблизить цену сделки к средней рыночной цене за торговый день. Крупные позиции часто разбиваются на более мелкие заявки, чтобы минимизировать влияние на цену и снизить затраты на их исполнение.

Иными словами, если на рынке нужно продать значительное количество акций, то одновременная продажа большой позиции может оказаться невыгодной, так как это повлечет за собой выполнение заявок далеко от текущей рыночной цены. Постепенная продажа акций позволяет рынку восстановиться, что даёт возможность вновь продавать позиции по более выгодной цене. В этой связи задача прогнозирования объемов торгов становится важной для оптимального исполнения крупных рыночных позиций.

В статье [YYL14] представлены алгоритмические стратегии для снижения подобных издержек. Она исследует эффективность модели ARFIMA, а также сравнивает динамический VWAP со статическим SVWAP.

[*]e-mail: vakhid.bakhyshov@math.msu.ru
[†]e-mail: maksim.dugin@math.msu.ru
[‡]e-mail: roman.samoilov@math.msu.ru
[§]e-mail: nikita.chasov@math.msu.ru

В этом исследовании сначала будут реализованы предложенные в статье эконометрические методы для прогнозирования внутридневных объемов. Затем будут применены модели нейронных сетей для предсказания и, наконец, проведено совмещение этих моделей для улучшения предсказательной точности.

1 Strategy (Econometric approach)

Основная стратегия, рассмотренная в статье, разделена на SVWAP и PC-VWAP методы, т.е исследование рассматривает две стратегии для VWAP — статическую SVWAP и динамическую PC-VWAP. Стратегия SVWAP базируется на предположении, что дневные распределения объемов торгов остаются стабильными и рассчитываются путем усреднения данных за фиксированный период (т.е SVWAP основана на фиксированном распределении объема торгов на протяжении дня). Напротив, PC-VWAP использует метод главных компонент (PCA), позволяющий разделить объем торгов на общую и специфическую компоненты, что позволяет учитывать сезонные колебания и случайные отклонения, связанные с конкретными событиями, что улучшает точность прогнозирования при включении новой рыночной информации. Также используется ARMA для краткосрочных прогнозов и ARFIMA для учета долгосрочной памяти в специфических компонентах объема. Стратегия PC-VWAP оказывается более гибкой и точной, так как учитывает внутридневные колебания и долгосрочную зависимость.

1.1 SVWAP

Метод SVWAP (Static VWAP) основывается на статическом распределении торгового объема (т.е предполагает фиксированное распределение объема торгов в течение дня, исходя из исторических данных), которое рассчитывается путем усреднения дневных данных за определенный период, обычно 20 дней. Этот метод рассчитывает пропорции объема для каждого временного интервала и применяет их для прогнозирования. Данное усреднение определяет доли объема для каждой части дня. Однако недостатком SVWAP является ее статичность и неспособность учитывать новую рыночную информацию, что может приводить к значительным отклонениям в прогнозах, так как изменения в рыночных условиях игнорируются.

$$v^d_{L+1,t,m} = \frac{\sum_{i=1}^{L} v^d_{i,t,m}}{L}$$

1.2 PC-VWAP

Метод PC-VWAP использует метод главных компонент для разложения объемов на общую и специфическую компоненты, что позволяет учесть сезонные U-образные колебания и специальные события, которые могут влиять на объем торговли (общий компонент характеризует U-образное сезонное колебание, а специфический учитывает события, влияющие на объем). Предсказания по общей компоненте делаются путем усреднения исторических данных (т.е на основе средней исторической величины), тогда как специфическая компонента прогнозируется с помощью моделей ARMA, ARFIMA или SETAR. В данной статье также исследуется использование модели ARFIMA для специфического компонента, что может улучшить точность предсказаний в условиях долгосрочной памяти.

$$x_{i,t,m} = \frac{v_{i,t,m}}{N_{i,t,m}}$$

$$x_{i,t,m} = c_{i,t,m} + e_{i,t,m}$$

2 Neural networks and machine learning models

2.1 LSTM (Long Short-Term Memory)

LSTM (Long Short-Term Memory) нейронные сети - это тип рекуррентных нейронных сетей (RNN), которые особенно хорошо подходят для работы с временными рядами, где зависимость между предыдущими и последующими значениями может быть значительной. LSTM используются для предсказания будущих значений, таких как объем торгов, в зависимости от исторических данных. Соотвественно, в нашем проекте мы воспользовались нейронкой LSTM. Казалось бы, а почему мы используем нейронку LSTM, потому хорошо подходят для предсказания объема торгов: 1) Временная зависимость: Объемы торгов в финансовых данных часто являются автокоррелированными, т.е. значения в прошлом влияют на будущие значения. LSTM могут "запомнить" важные исторические данные и использовать их для предсказания; 2) Нелинейность: Рынки часто характеризуются нелинейными зависимостями. LSTM способны учитывать нелинейные паттерны в данных и могут более точно предсказывать объемы, чем линейные модели.

Разберём именно нашу модель:

1. Предобработка данных: Исходный временной ряд нормализуется методом Min-Max для ускорения сходимости модели. Мы разрезаем данные на обучающие и тестовые наборы, а также формируем последовательности данных для LSTM на основе скользящего окна.
2. Архитектура модели: Модель состоит из слоя LSTM и линейного выходного слоя, который преобразует выходные значения из скрытого состояния в прогнозы объема. Параметры модели, такие как количество скрытых узлов, слоев и скорость обучения, настраиваются под каждый тикер на основе лучших параметров.
3. Обучение модели: LSTM обучается минимизации ошибки предсказания (MSE) с использованием оптимизатора Adam. На каждом шаге обновляются веса, чтобы минимизировать разницу между прогнозом и реальным значением объема торгов. Процесс повторяется на нескольких эпохах, и результаты обучения контролируются по значению потерь.
4. Прогнозирование: После обучения модель тестируется на последних доступных данных. Мы подаем последовательности с известными значениями объема, сдвигая окно и добавляя реальные данные, чтобы LSTM могла корректировать прогноз на основе актуальных изменений. Прогнозируемые значения восстанавливаются из нормализованного диапазона.
5. Оценка: После прогноза модель оценивается с использованием корня среднеквадратичной ошибки (RMSE) по последнему отрезку данных, что позволяет количественно оценить точность модели для каждого тикера.

Такой подход позволяет получить надежные прогнозы объема торгов с учетом временных зависимостей в данных.

2.2 XGBoost and Random forest

В рамках задачи предобработки данных и создания прогнозов по объему торгов применяются следующие этапы:

1. Загрузка и предобработка данных: Временной ряд для каждого тикера загружается, включая закрывающую цену и объем торгов. На основе данных о цене закрытия рассчитываются основные технические индикаторы, что помогает выделить тренды и волатильность, а также предсказывать будущие изменения.
2. Создание признаков с техническими индикаторами:** Вычисляются такие индикаторы, как:
   - SMA и EMA (простой и экспоненциальный скользящие средние), чтобы уловить краткосрочные и долгосрочные тенденции.
   - RSI (индекс относительной силы) для измерения перепроданности или перекупленности.
   - MACD (конвергенция и дивергенция скользящих средних) для выявления силы тренда и потенциальных разворотов.
   - Bollinger Bands для оценки волатильности.
   - ROC (скорость изменения), Stochastic Oscillator и Momentum для отслеживания скорости изменений и трендов.
   - DPO (осциллятор детрендированных цен), чтобы устранять долгосрочные тренды и фокусироваться на цикличности.

   После расчета индикаторы объединяются в датафрейм для создания полноценного набора признаков.
3. Выбор признаков: Используется регрессор XGBoost, чтобы определить наиболее значимые признаки для прогнозирования объема торгов. Функция визуализирует важность признаков, что позволяет сфокусироваться на тех, которые имеют наибольшее влияние.
4. Оценка: После прогноза модель оценивается с использованием корня среднеквадратичной ошибки (RMSE) по последнему отрезку данных, что позволяет количественно оценить точность модели для каждого тикера.

   Каждая модель обучается на данных и делает прогноз для тестовой выборки. Результаты предсказаний объединяются в стеке, включающем предсказания каждого из алгоритмов и реальные значения.

Такой подход, объединяющий технические индикаторы и машинное моделирование, позволяет учесть особенности и шум данных, обеспечивая более точные и устойчивые прогнозы.

3 Empirical analysis

3.1 Data

Для анализа были использованы данные о часовых значениях объема торгов и цены закрытия акций по конкретному тикеру.

Для анализа с помощью модели PC-ARMA данные были преобразованы в 14 отдельных временных рядов, где каждый ряд соответствует одному часу в течение торгового дня. Например, файл с именем $GAZP\_10.csv$ содержит данные об объеме торгов акций «Газпром» за десятый час торгов (с 10:00 до 10:59) за каждый день в период с 1 июня 2023 года по 23 октября 2024 года. Такой подход позволил выделить особенности временных рядов для каждого часа отдельно, что способствует повышению точности прогнозов при анализе внутридневных паттернов.

Чтобы применить методы машинного обучения, изначальный временной ряд, включающий данные об объеме торгов за доступные торговые часы, был сохранен в полном объеме.

3.2 Initial models

Изначально мы брали немыслемое количество моделей машиного обучения. Но в конечном варианте оставили случайный лес, градиентный бустинг и $LSTM$.

Их метрики можно видеть в таблице C.1 по колонкам $rmse\_lstm$ (Корень среднеквадратической ошибки (RMSE) модели LSTM), $rmse\_fe$ (Корень среднеквадратической ошибки (RMSE) модели PC-ARMA), $rmse\_rf$ (Корень среднеквадратической ошибки (RMSE) модели случайного леса (RF)), $rmse\_gb$ (Корень среднеквадратической ошибки (RMSE) модели градиентного бустинга (GB)).

3.3 Improving the model

Для улучшения исходных моделей была реализована кроссвалидация 3.3.1, подобраны гиперпараметры 3.3.2 и реализован стейкинг с особыми весами 3.3.3.

3.3.1 Cross-validation

Для модели PC-ARMA реализована кросс-валидация с расширяющимся окном **??**; первая обучающая выборка - дни от первого до двадцать первого с конца, последняя - все дни, представленные в датасете, кроме последнего (то есть окно 20 раз расширялось на один день). Тестовые данные - это следующий день за тренировочными.

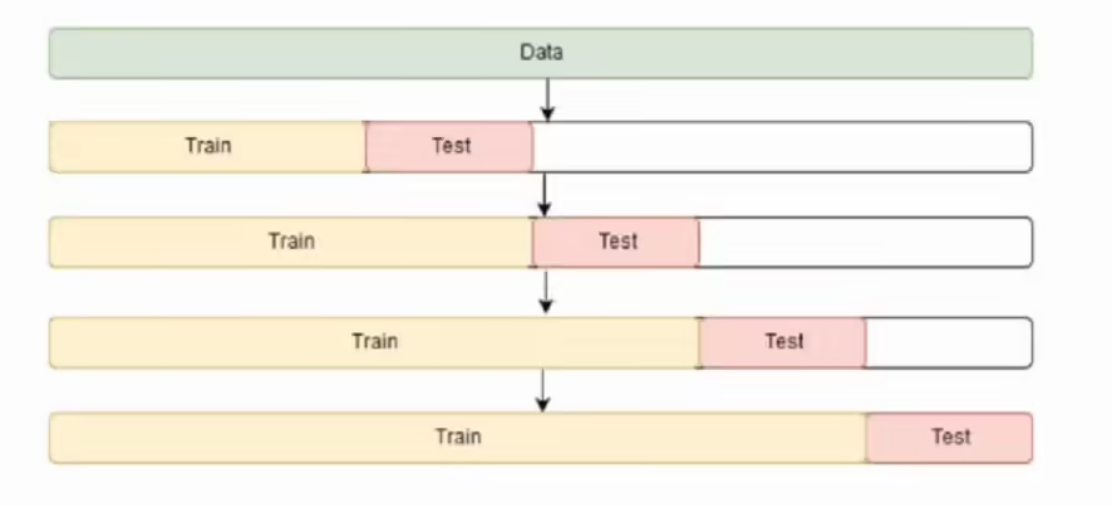

Figure 3.1: Кросс-валидация с расширяющимся окном

Для моделей машинного обучения использовалась кросс-валидация со скользящим окном **??**: размер окна является гиперпараметром в модели LSTM, в остальных моделях первая обучающая выборка - дни от первого до двести восьмидесятого с конца, последняя - дни от 280-го до 14-го с конца (то есть окно 20 раз двигалось на 14 дней).

3.3.2 Optimization of hyperparametrs

Для модели PC-ARMA был выбран размер временного окна так, чтобы метрика MSE, усреднённая по всем акциям и батчам, была минимальной.

Для моделей машинного обучения выбор гиперпараметров реализован с помощью библиотеки optuna.

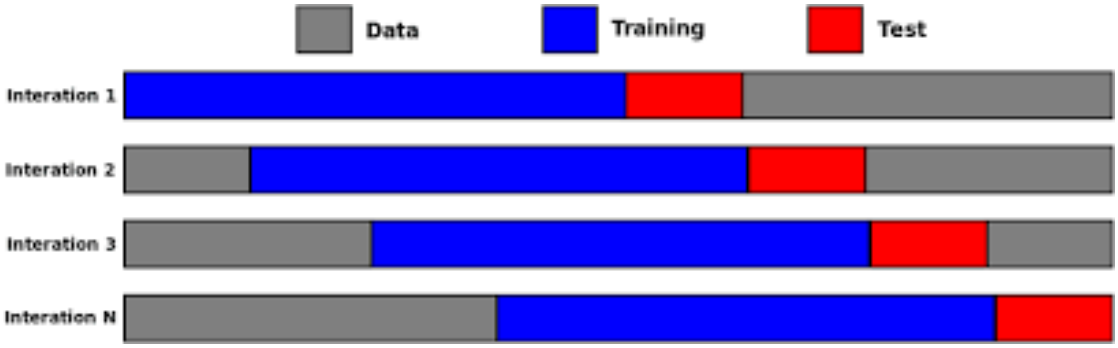

Figure 3.2: Кросс-валидация со скользящим окном

3.3.3 Stacking

Описать реализацию стейкинга можно в простых шагов:

1. Расчет ошибок: Для каждой модели (LSTM, PC-ARMA, Random Forest, Gradient Boosting) были рассчитаны абсолютные ошибки прогнозов по отношению к фактическому объему торгов. Ошибки определяют, насколько сильно каждая модель отклоняется от истинных значений.
2. Веса на основе ошибок: Для каждой модели был рассчитан вес, обратно пропорциональный ошибке, что позволяет более точным моделям оказывать большее влияние на итоговый прогноз. Вес для каждой модели определялся как:

$$weight = \frac{1}{error}$$

3. Нормализация весов: Чтобы суммарный вес всех моделей был равен единице, веса нормализовались, делением веса каждой модели на сумму всех весов. Это позволяло создать сбалансированный ансамбль, где каждая модель учитывалась пропорционально своей точности.
4. Объединенный прогноз: После нормализации весов итоговый прогноз рассчитывался как взвешенная сумма прогнозов всех моделей, что создавало метапрогноз. Формула комбинированного прогноза выглядела следующим образом:

$$\begin{aligned} combined\_forecast = (&\text{norm.weight LSTM} \times \text{forecast LSTM}) + \\ +(&\text{norm.weight PC-ARMA} \times \text{forecast PC-ARMA}) + \\ +(&\text{norm.weight RF} \times \text{forecast RF}) + \\ +(&\text{norm.weight GB} \times \text{forecast GB}) \end{aligned}$$

5. Оценка точности объединенной модели: Для анализа точности метапрогноза были рассчитаны метрики: MSE, RMSE, MAE и MAPE, а также RMSE для каждой из моделей и комбинированного прогноза.

3.4 Results

Результаты примения можно видеть на графиках в анендиксе.

Conclusion

Исследовав метрики предсказательной способности моделей, можно сделать следующие выводы:

1. Модели машинного обучения показывают некоторую предсказательную способность, успешно улавливая поведение данных и их распределение. Однако при наличии выбросов и резких "скачков" в реальных данных модели часто не учитывают такие аномалии и, как следствие, не способны их предсказать, что приводит к значительному росту ошибок, таких как MSE и других метрик.
2. Применение предложенных в статье моделей к российским данным показало худшую предсказательную способность по сравнению с результатами, полученными на данных китайского фондового рынка.
3. Совмещение моделей позволяет уменьшить абсолютное значение ошибок. В наших данных это привело к снижению ошибок примерно в два раза.

References

[Rob00] Neil Chriss Robert Almgren. “Optimal Execution of Portfolio Transactions”. In: (2000).

[Ann13] Jiang Wang Anna Obizhaeva. “Optimal trading strategy and supply/demand dynamics”. In: (2013).

[YYL14] Xunyu Ye, Rui Yan, and Handong Li. *Forecasting trading volume in the Chinese stock market based on the dynamic VWAP*. 2014.

[WS21] Liang Zhao Wei Li Ruihan Bao Keiko Harimoto Yunfang Wu and Xu Sun. “Long-term, Short-term and Sudden Event: Trading Volume Movement Prediction with Graph-based Multi-view Modeling”. In: (2021).

Appendix A Github

Весь код опубликован у нас на Github: Forecasting Trading Volume

Помимо этого там находятся все расчёты дополнительных статистик как для моделей, используемых в этой статье, так и для моделей, которые в это статью не были включены из-за их плохих предсказательных способностей.

Appendix B Measures to assess the predictive capabilities of models

Средняя абсолютная ошибка (Mean Absolute Error, MAE): $MAE = \frac{1}{H}\sum_{t=1}^{H}|y_t - \hat{y_t}|$

Среднеквадратическая ошибка (Mean Square Error, MSE): $MSE = \frac{1}{H}\sum_{t=1}^{H}(y_t - \hat{y_t})^2$

Корень из среднеквадратической ошибки (RMSE): $RMSE = \sqrt{\frac{1}{H}\sum_{t=1}^{H}(y_t - \hat{y_t})^2}$

Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error, MAPE):

$$MAPE = 100\% \cdot \frac{1}{H}\sum_{t=1}^{H}|\frac{y_t - \hat{y_t}}{y_t}|$$

Средняя абсолютная шкалированная ошибка (Mean Absolute Scaled Error, MASE):

$$MASE = \frac{\frac{1}{T}\sum_{t=1}^{H}(y_t - \hat{y_t})^2}{\frac{1}{T-1}\sum_{t=2}^{H}(y_t - \hat{y_{t-1}})^2}$$

Appendix C Results

Table C.1: Меры качества прогнозов

ticker	mape	rmse_lstm	rmse_fe	rmse_rf	rmse_gb	rmse_combined
GAZP	107.77%	6.46e+06	5.98e+06	5.84e+06	5.80e+06	5.41e+06
GMKN	38.22%	1.44e+06	1.50e+06	1.75e+06	1.75e+06	1.45e+06
LKOH	60.52%	4.59e+04	5.36e+04	5.40e+04	5.58e+04	2.79e+04
NVTK	111.14%	2.32e+05	2.53e+05	1.78e+05	1.90e+05	8.77e+04
PLZL	76.97%	3.55e+04	3.50e+04	4.28e+04	2.92e+04	2.84e+04
ROSN	63.73%	2.30e+05	2.23e+05	2.50e+05	2.51e+05	1.98e+05
SBER	112.74%	2.78e+06	2.52e+06	1.96e+06	2.02e+06	1.28e+06
SIBN	39.07%	1.23e+05	1.23e+05	1.68e+05	1.41e+05	9.93e+04
TATN	64.38%	1.93e+05	1.91e+05	1.42e+05	1.52e+05	1.12e+05

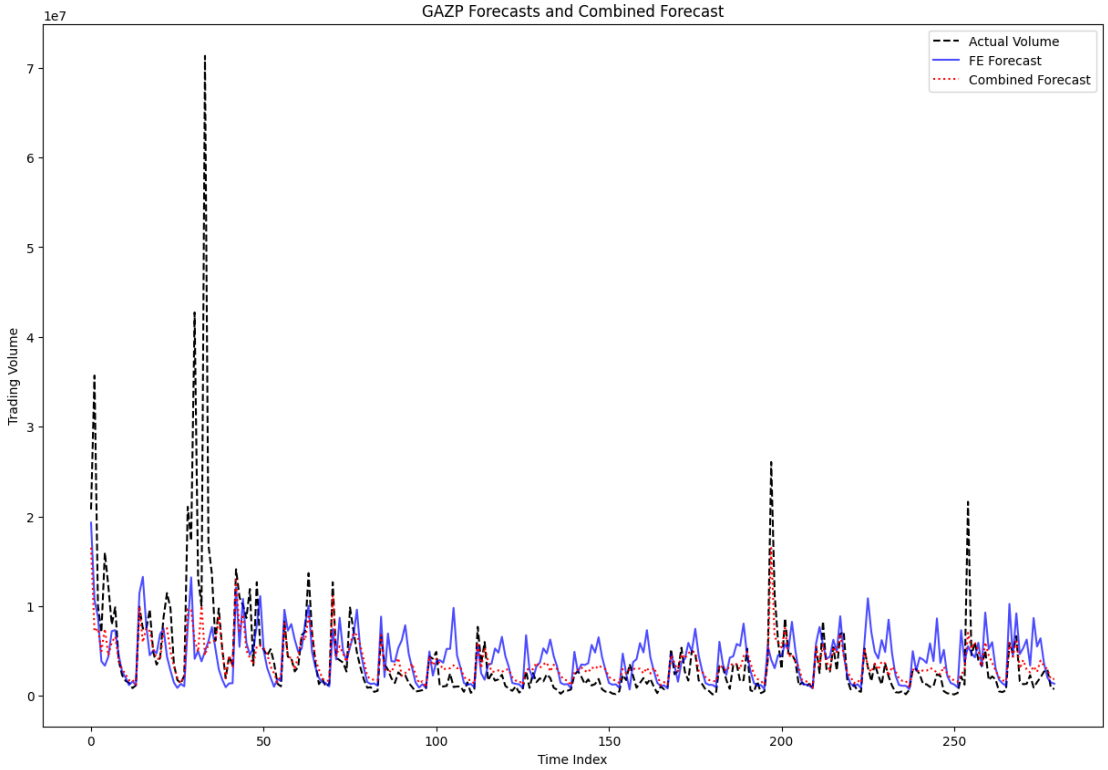
GAZP Forecasts and Combined Forecast
1e7
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
Time Index
0
1
2
3
4
5
6
7
0
50
100
150
200
250

Figure C.1

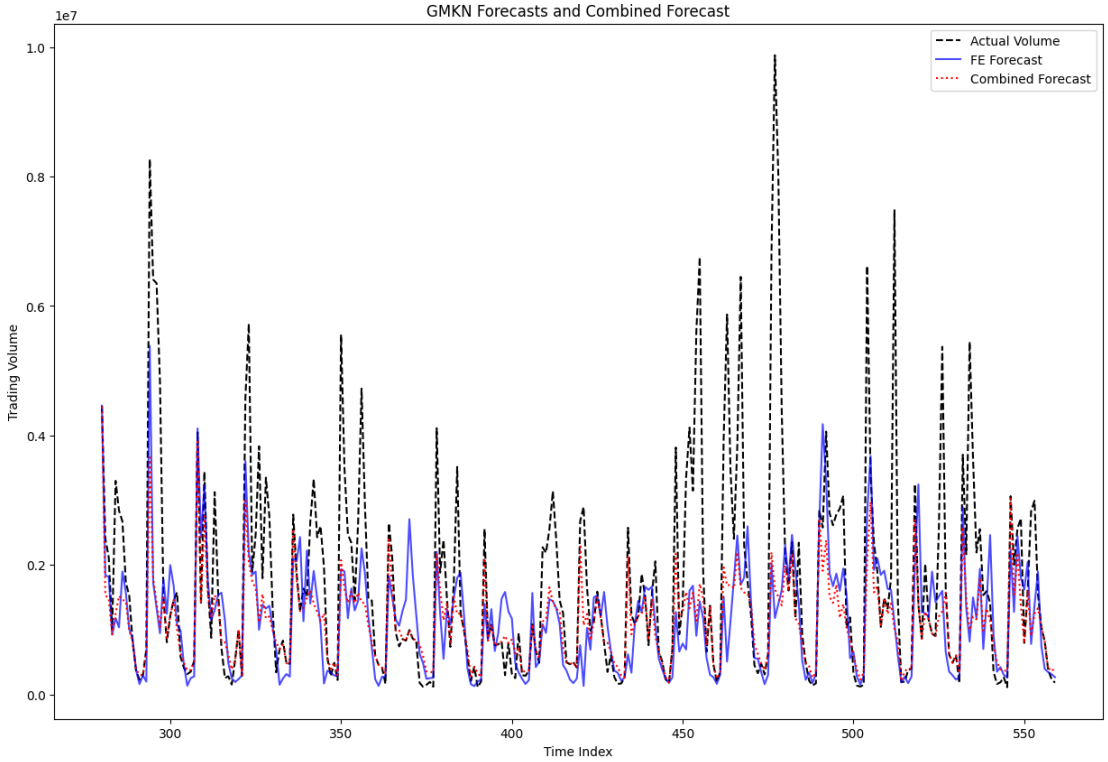
GMKN Forecasts and Combined Forecast
1e7
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
Time Index
0.0
0.2
0.4
0.6
0.8
1.0
300
350
400
450
500
550

Figure C.2

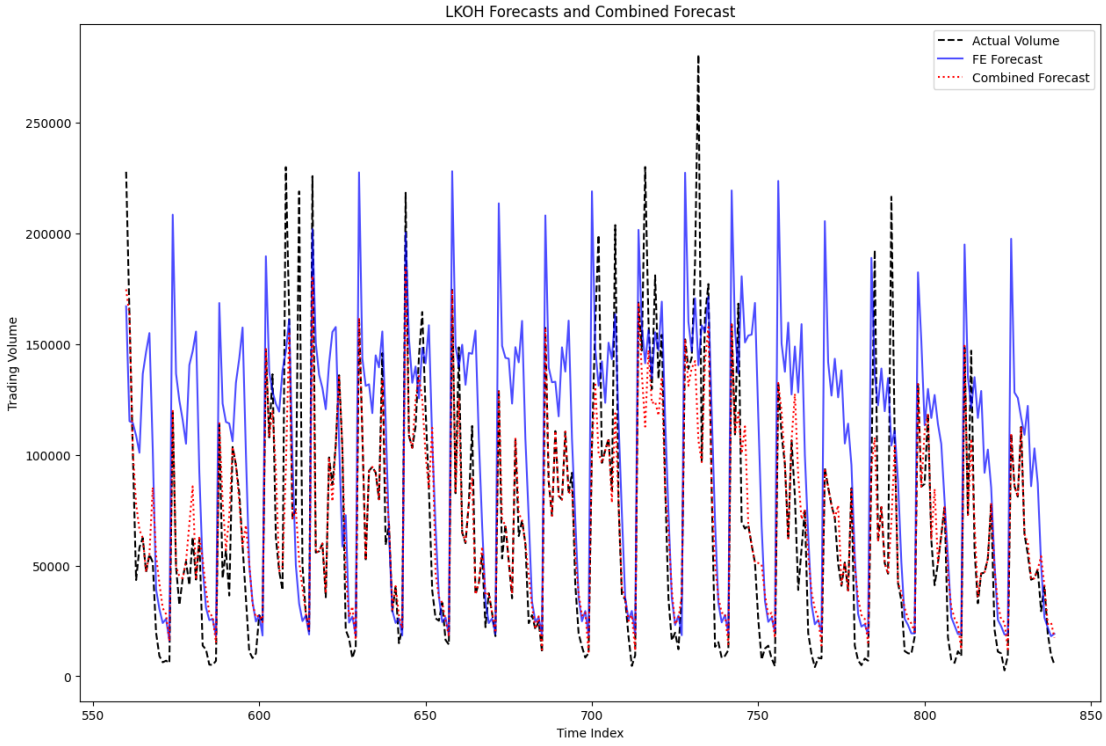
LKOH Forecasts and Combined Forecast
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
Time Index
0
50000
100000
150000
200000
250000
550
600
650
700
750
800
850

Figure C.3

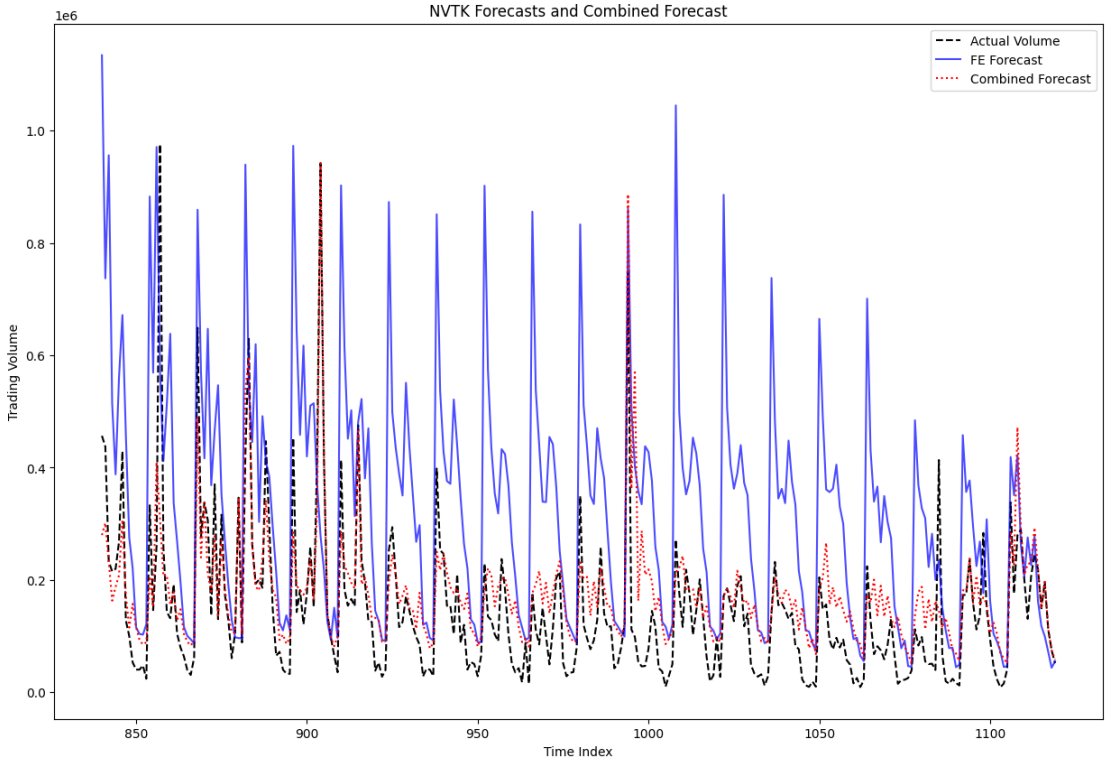
NVTK Forecasts and Combined Forecast
1e6
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
0.0
0.2
0.4
0.6
0.8
1.0
850
900
950
1000
1050
1100
Time Index

Figure C.4

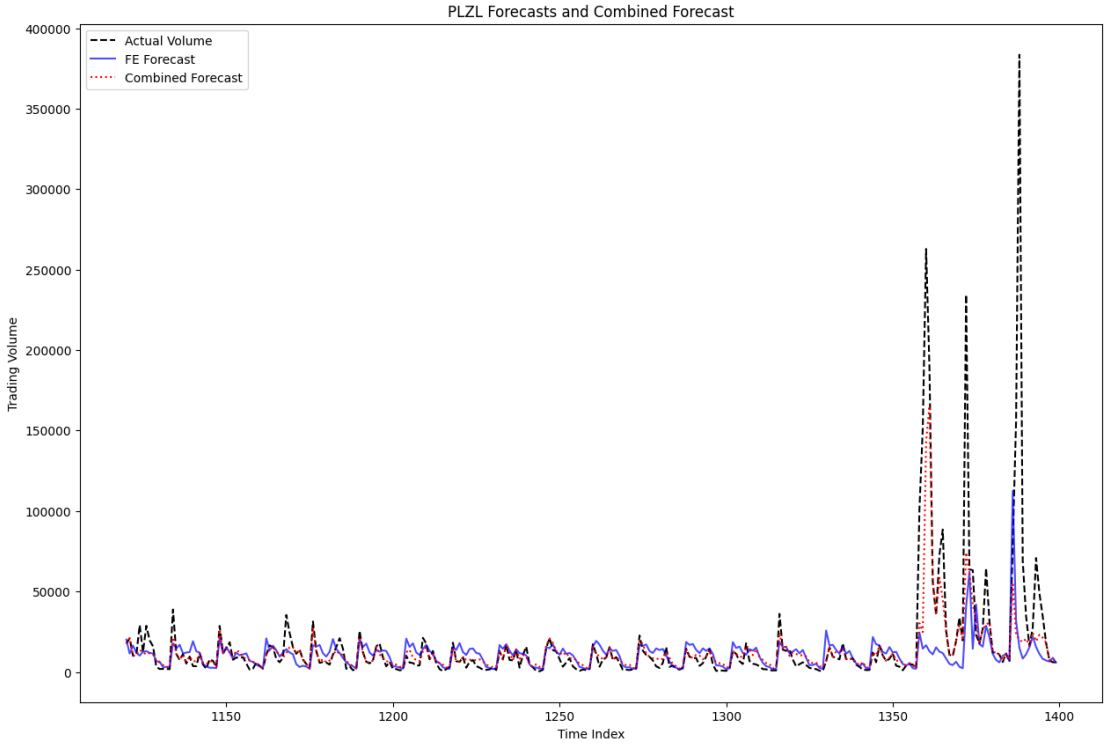
PLZL Forecasts and Combined Forecast
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
Time Index
400000
350000
300000
250000
200000
150000
100000
50000
0
1150
1200
1250
1300
1350
1400

Figure C.5

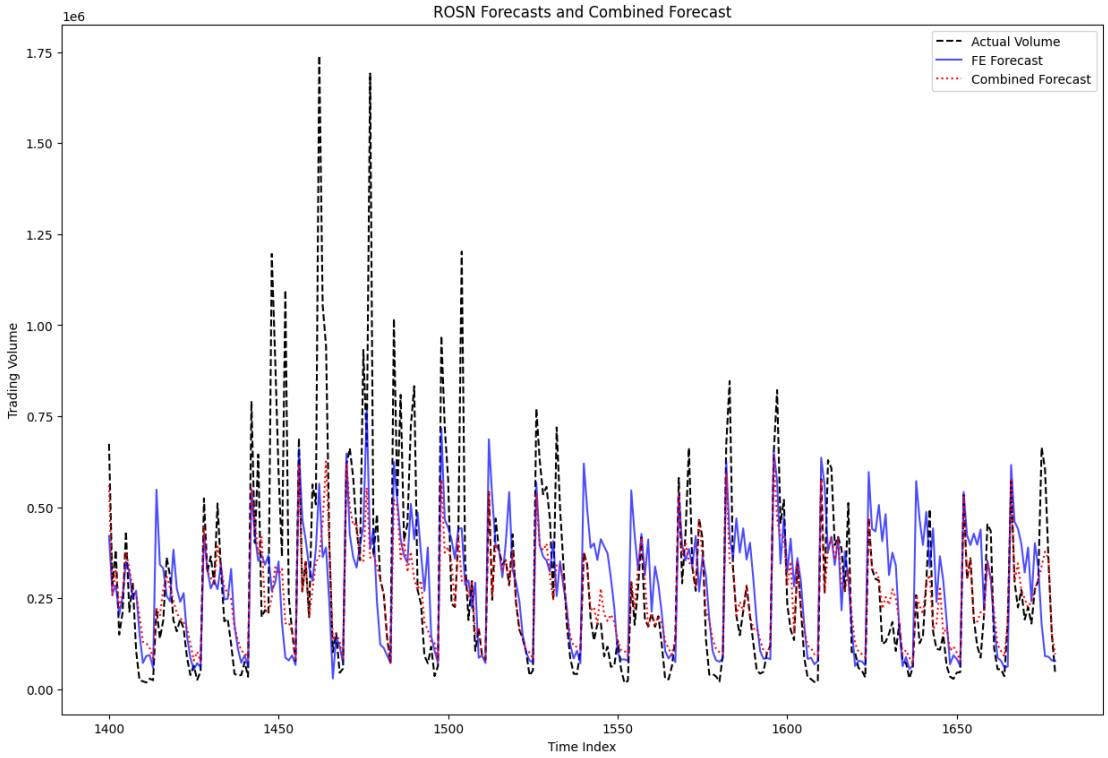
ROSN Forecasts and Combined Forecast
1e6
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
Time Index
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1400
1450
1500
1550
1600
1650

Figure C.6

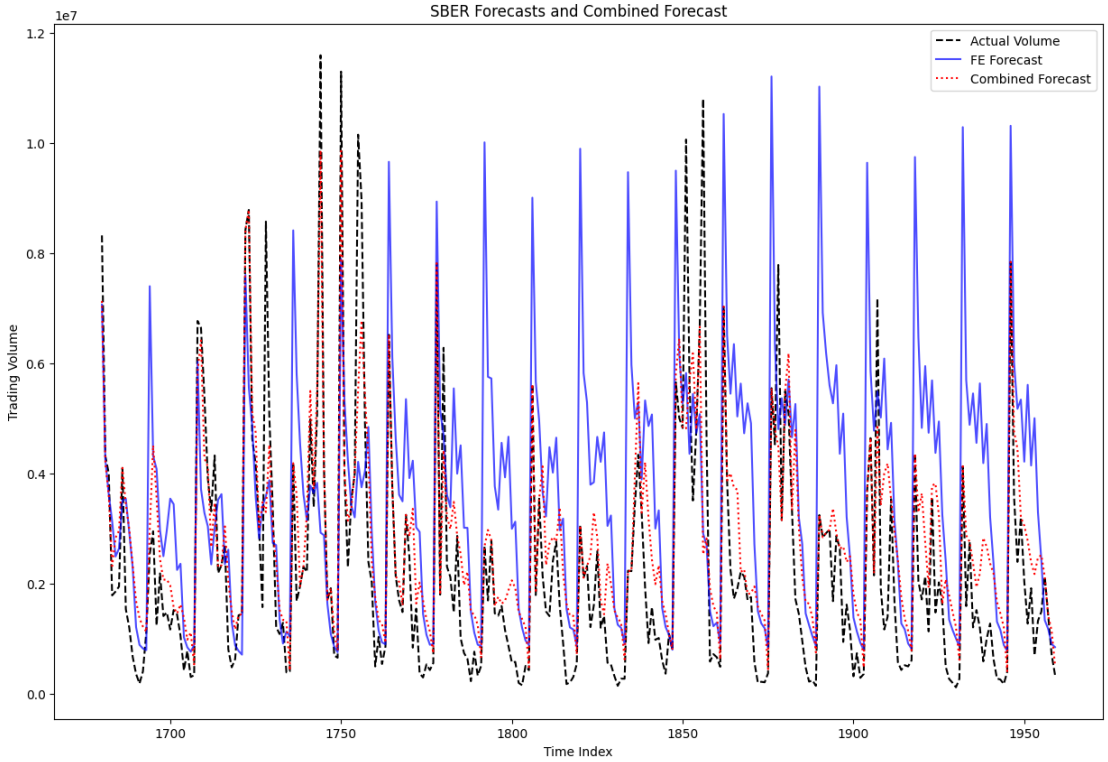
SBER Forecasts and Combined Forecast
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Trading Volume
1700
1750
1800
1850
1900
1950
Time Index
Actual Volume
FE Forecast
Combined Forecast

Figure C.7

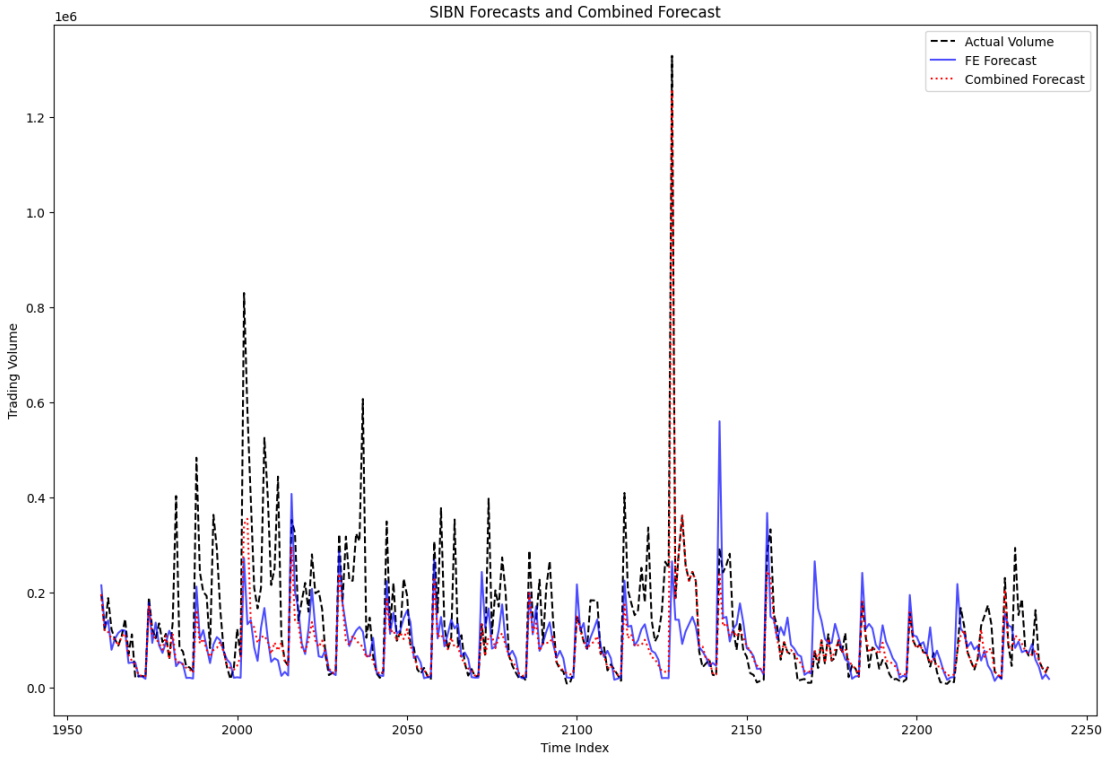
SIBN Forecasts and Combined Forecast
1e6
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
Time Index
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1950
2000
2050
2100
2150
2200
2250

Figure C.8

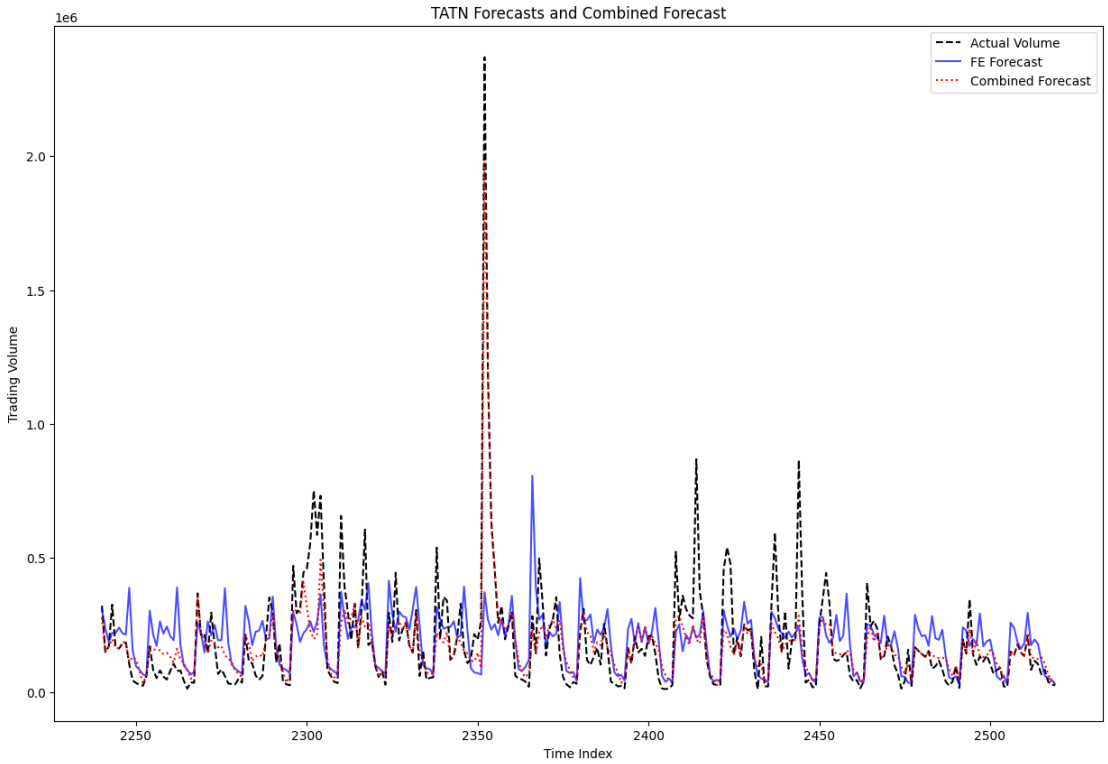
TATN Forecasts and Combined Forecast
1e6
Actual Volume
FE Forecast
Combined Forecast
Trading Volume
2.0
1.5
1.0
0.5
0.0
2250
2300
2350
2400
2450
2500
Time Index

Figure C.9