Прогнозирование волатильности на криптовалюту с помощью GARCH и многомерного LSTM

С момента первого появления Биткойна в 2009 году он существенно изменил мировой финансовый ландшафт. Децентрализованная криптовалюта зарекомендовала себя как класс активов, признанный многими управляющими активами, крупными инвестиционными банками и хедж-фондами. Поскольку скорость массового внедрения продолжает расти, это также побуждает инвесторов изучать новые инструмент хеджирования, такие как крипто-опционы и фьючерсы.

> Исторически известно, что Биткойн более волатильный, чем регулируемые акции и товары. Его последний всплеск в конце декабря 2020 года, начале января 2021 года вызвал много вопросов и неуверенности в отношении будущего финансового ландшафта. На момент написания этого отчета (конец августа 2021 года) биткойн торгуется по цене чуть ниже 50 000 долларов США, что немаловажно, учитывая, что в 2020 году он вошел на уровне около 7 200 долларов США.

> Цель этого проекта — заглянуть в будущее, **спрогнозировав среднюю дневную реализованную волатильность (RV) BTC-USD на следующие 7 дней** с использованием двух различных подходов — традиционного эконометрического подхода к прогнозированию волатильности финансовых временных рядов **GARCH** и современные **нейронные сети LSTM** .

Бизнес-проблема

Волатильность пытается измерить величину движения цены финансового инструмента в течение определенного периода времени. Чем более резкие колебания цены в этом инструменте, тем выше уровень волатильности, и наоборот.

Волатильность обычно считается лучшей мерой рыночного риска, и прогнозирование волатильности используется во многих различных приложениях в отрасли. **Реализованные** модели прогнозирования волатильности обычно используются в управлении рисками, создании рынка, оптимизации портфеля и торговле опционами. В частности, согласно Sinclair (2020), ряд торговых стратегий вращается вокруг выявления ситуаций, в которых возникает это несоответствие волатильности:

$$P / L = Vega|\sigma_{implied} - \sigma_{realized}|$$

в котором Вега — это измерение чувствительности цены опциона к изменениям волатильности базового актива, а σ— волатильность. Поскольку подразумеваемая волатильность (IV) может быть получена из цен опционов с использованием таких моделей, как модель Блэка-Шоулза, прогнозирование реализованной волатильности даст нам ключ ко второй части уравнения.

Хотя прогнозирование и моделирование волатильности находится в центре внимания многих эмпирических и теоретических исследований в академических кругах, точное прогнозирование волатильности остается серьезной проблемой для ученых. Вдобавок ко всему, поскольку торговля крипто-опционами является относительно новой, не было проведено столько исследований по прогнозированию волатильности биткойнов. Кроме того, криптовалюты имеют определенные нюансы, которые сами по себе отличаются от традиционных регулируемых акций и товаров, которые также необходимо учитывать.

Набор данных

Исторический набор данных о ценах открытия/закрытия/максимума/минимума биткойнов был получен с использованием Yahoo Finance API **`yfinance`** . Этот API бесплатный, его очень легко настроить, и он содержит широкий спектр данных и предложений.

Я буду загружать цены BTC-USD, используя тикер `BTC-USD` с интервалом в 1 день. Yahoo не добавляла Биткойн до 2014 года; и поэтому, хотя он впервые был продан в 2009 году, он **`yfinance`** содержит данные только с сентября 2014 года по настоящее время (декабрь 2022 года). Поэтому я буду работать с ок. 2500 точек данных, охватывающих около 7 лет торговых дней.

Структура набора данных

Набор данных содержит ежедневные цены BTC-USD, включая:

- `Open`
- `High`
- `Low`
- `Close`

Цель этого проекта — сравнить классический подход с моделью на нейронных сетях, чтобы спрогнозировать среднюю дневную **Реализованную волатильность** BTC-USD на 7 дней вперед, используя 30-дневный интервал.

Измерение волатильности

Волатильность измеряет **не** направление изменения цены финансового инструмента, а только ее дисперсию за определенный период времени. Высокая волатильность связана с более высоким риском, а низкая волатильность с меньшим риском. Существует 2 основных типа волатильности:

- **Историческая волатильность** или **реализованная волатильность** (RV) — это фактическая волатильность, продемонстрированная базовым активом в течение определенного периода времени. Реализованная волатильность обычно рассчитывается как стандартное отклонение доходности цены, то есть изменение цены в долларах в процентах от цены предыдущего дня.
- **С другой стороны, подразумеваемая волатильность** (IV) — это уровень волатильности базового актива, который подразумевается текущей ценой опциона.

(Основное внимание в этом проекте уделяется **НЕ подразумеваемой волатильности** , которая может быть получена из моделей ценообразования опционов, таких как модель Блэка-Шоулза).

Традиционно реализованная волатильность определяется как **стандартное отклонение дневной доходности за определенный период времени** .
Математически **ежедневный доход** можно представить как:
$R_{t,\ t+i} = P_{t+i} / P_t * 100$

Однако с практической точки зрения обычно предпочтительнее использовать **Log Returns** , особенно в математическом моделировании, поскольку это помогает устранить нестационарные свойства данных временных рядов и делает их более стабильными:

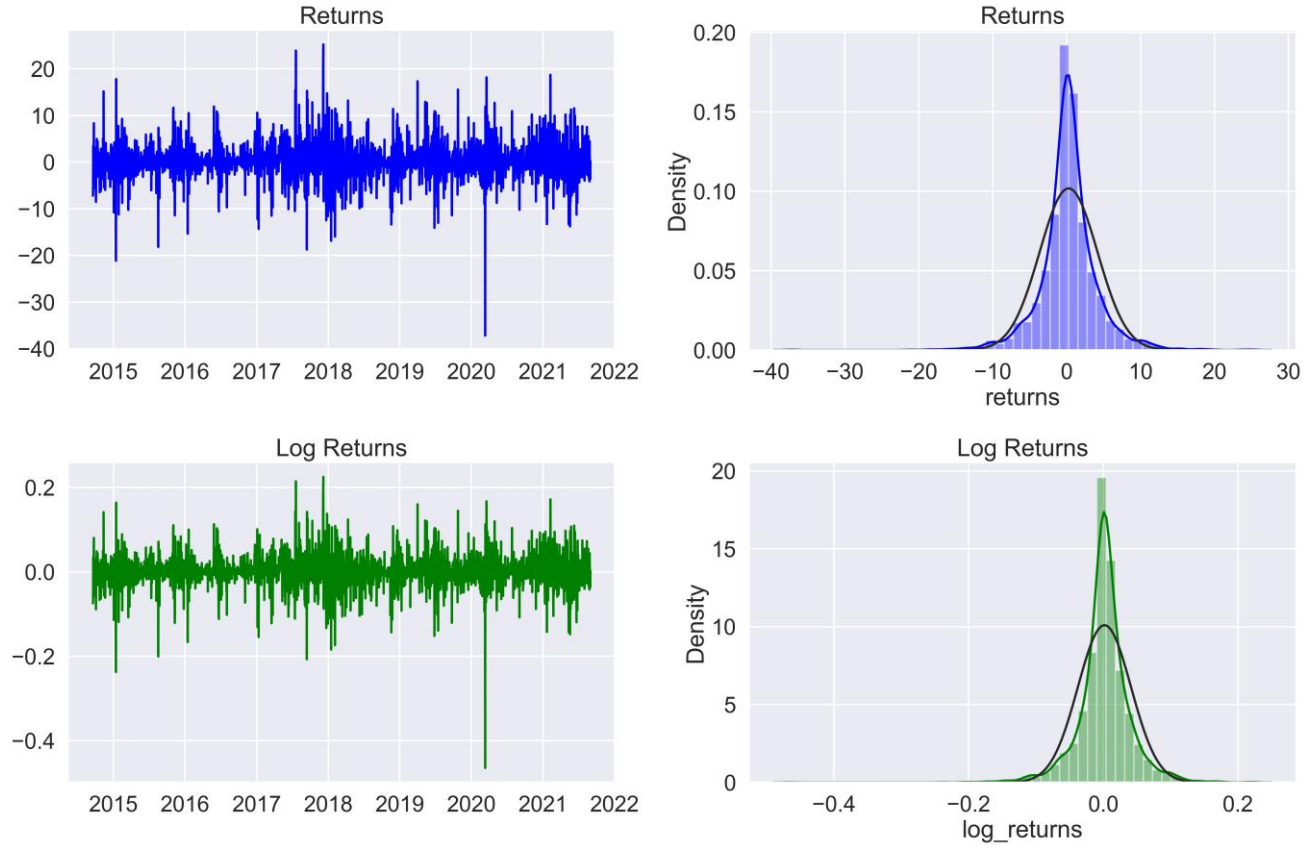
Формула **logarithmic return :**

$$r_{t,\ t+i} = log(P_{t+i} / P_t)$$

(В обеих формулах P_tпредставляет цену на временном шаге t)

Есть еще одно преимущество журналирования возвратов, заключающееся в том, что они складываются во времени:

$$r_{t1,\ t2} + r_{t2,\ t3} = r_{t1,\ t3}$$

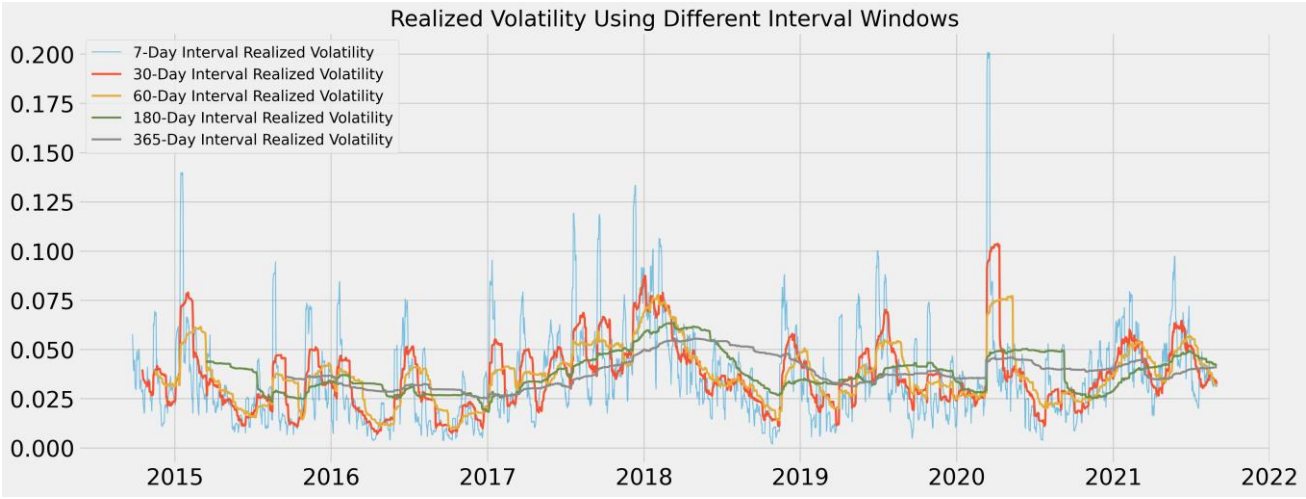


Для этого конкретного проекта **ЕЖЕДНЕВНАЯ РЕАЛИЗОВАННАЯ ВОЛАТИЛЬНОСТЬ** рассчитывается с использованием **окна интервала** в **30 дней** следующим образом:

$$\sigma_{daily} = \sqrt{\sum_t r^2_{t-1,\ t}} * \sqrt{\frac{1}{interval - 1}}$$

Причина, по которой я выбрал 30 дней, заключается в том, что 7 дней кажутся слишком шумными для наблюдения значимых моделей, в то время как более длинные интервалы, по-видимому, значительно сглаживают волатильность и имеют тенденцию возвращаться к среднему значению.

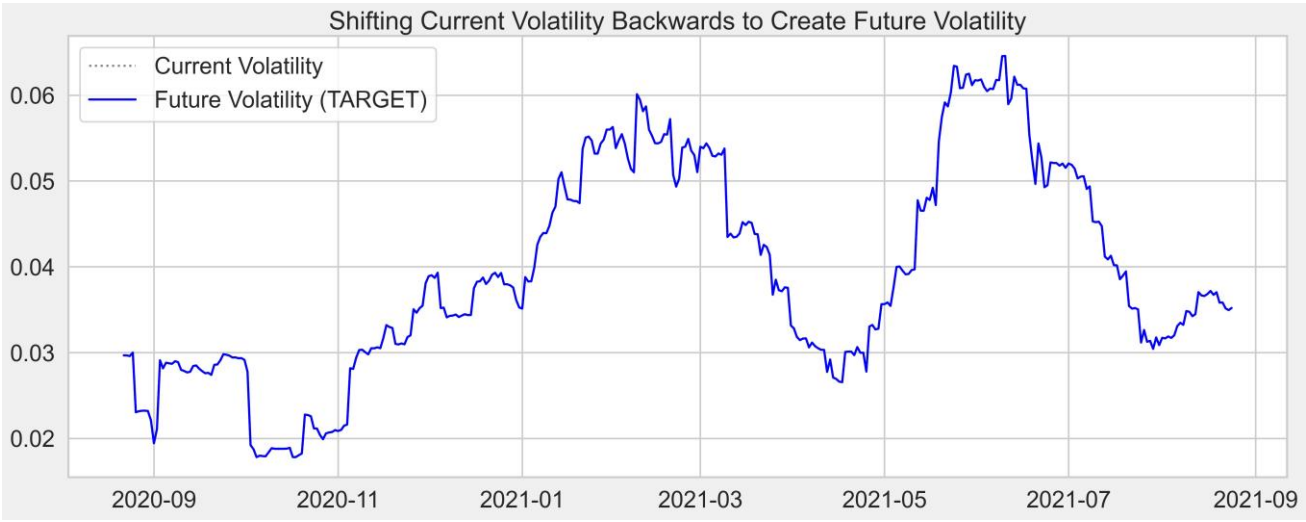
Использование окна интервала в 30 дней также поможет избежать потери слишком большого количества точек данных в начале набора данных.



Модели прогнозирования временных рядов — это модели, способные предсказывать **будущие** значения на основе ранее наблюдаемых значений. Целевые « **будущие** » данные в этом случае получаются путем **сдвига текущей волатильности назад** на количество `n_future` лагов.

Например, по отношению к понедельнику прошлой недели понедельник этой недели является « **будущим** »; поэтому мне просто нужно сдвинуть волатильность на этой неделе назад на 7 дней и использовать ее в качестве желаемого « **будущего** » результата прошлой недели, который я затем буду использовать для обучения нейронных сетей и оценки производительности модели.

Это визуализация того, как текущая волатильность смещается назад, чтобы стать будущими значениями, к которым я хочу в конечном итоге стремиться.



На приведенном выше графике **синяя линия** указывает на **целевое будущее** значение, которому я в конечном итоге пытаюсь соответствовать. А пунктирная **серая линия** представляет **текущую волатильность** на этом временном шаге.

Цель прогнозирования

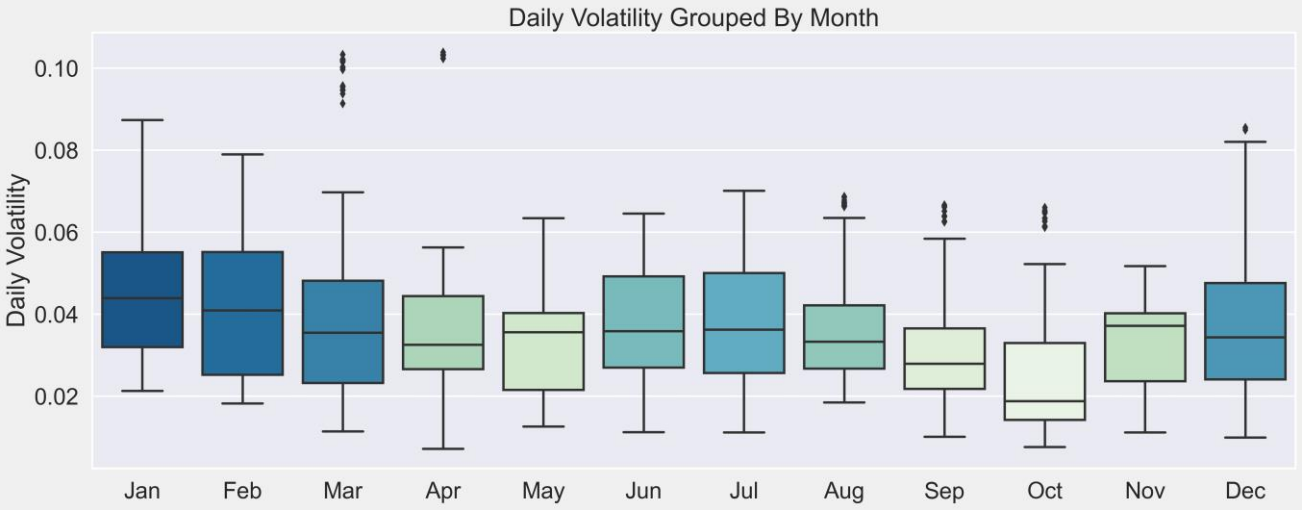
Целью здесь будет то, `vol_future` что представляет ежедневную реализованную волатильность следующих `n_future` дней с сегодняшнего дня (средняя дневная

волатильность от `t + n_future - INTERVAL_WINDOW + 1` до временного шага `t + n_future`).

Например, используя `n_future` значение 7 и значение `INTERVAL_WINDOW` 30, значение, которое я хочу предсказать на временном шаге `t` , будет средней ежедневной реализованной волатильностью от временного шага `t-22` к временному шагу `t+7` .

Анализ данных

Ежедневная волатильность, сгруппированная по месяцам



Можно заметить, что:

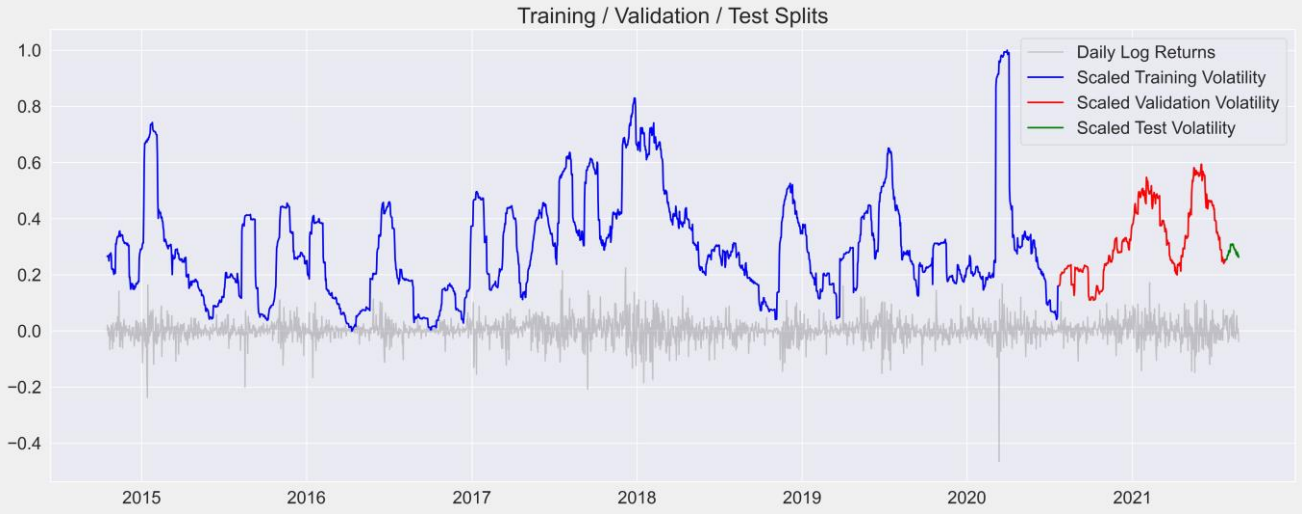
- волатильность постоянно достигала некоторых из своих самых высоких точек в декабре/январе за исторические периоды.
- Март и апрель имеют наибольшее количество крупных выбросов в то время
- , как август и сентябрь (которые являются предстоящими месяцами окончательного прогноза тестирования) исторически были относительно спокойными

Сплиты Train-Validation-Test

Всего в этом наборе данных 2500 пригодных для использования точек данных, которые охватывают период почти 7 лет с октября 2014 года по сегодняшний день (конец августа 2021 года). Поскольку криптовалюты не торгуются на регулируемой бирже, рынок биткойнов открыт круглосуточно и без выходных, 1 год охватывает целых 365 торговых дней вместо 252 дней в году, как в случае с другими акциями и товарами.

Я бы разделил набор данных на 3 части следующим образом:

- самые последние 30 пригодных для использования точек данных будут использоваться для **окончательного тестирования модели - прибл. 1,2%**
- 1 полный год (365 дней) для **проверки и настройки модели во время обучения — прибл. 14,7%**
- а остальные на **обучение -84,1 %**



Моделирование

Показатели эффективности

Обычно с финансовыми временными рядами, если мы просто просматриваем исторические данные, пробуя разные методы, параметры и временные масштабы, почти наверняка в какой-то момент будет найдена какая-то стратегия с прибыльностью в выборке. Однако вся цель «прогнозирования» состоит в том, чтобы предсказать будущее на основе доступной в настоящее время информации, и модель, которая лучше всего работает на обучающих данных, может быть не лучшей, когда речь идет об обобщении вне выборки (или **переоснащении**). Предотвращение/минимизация переобучения еще более важно на постоянно развивающихся финансовых рынках, где ставки высоки.

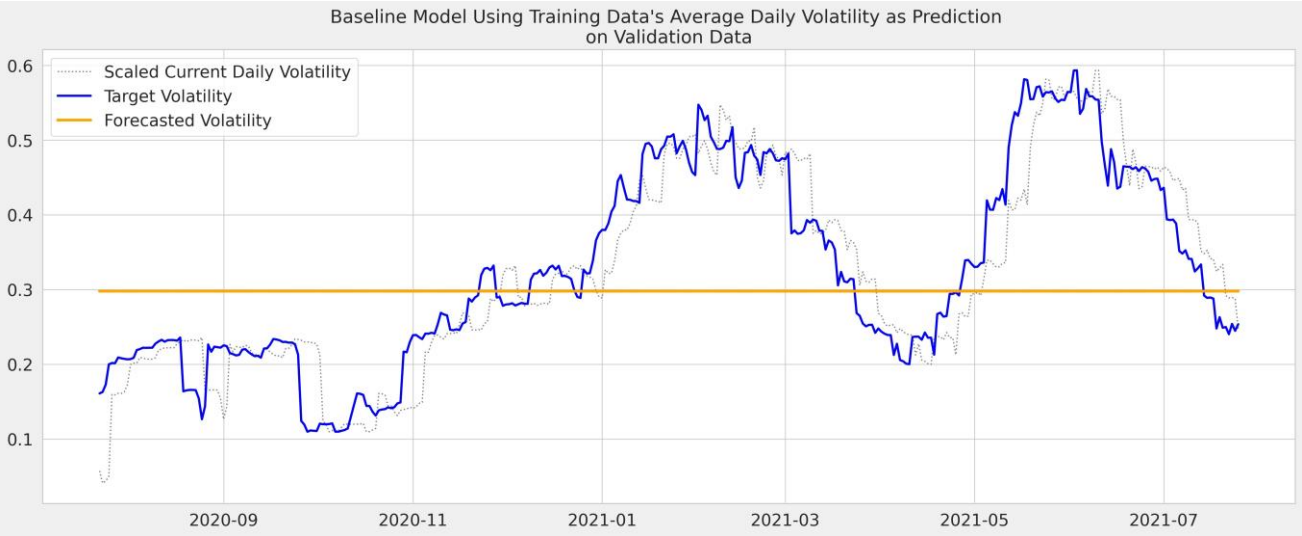
Я бы использовал две основные метрики: **RMSPE (среднеквадратичная ошибка в процентах)** и **RMSE (среднеквадратичные ошибки)** с приоритетом RMSPE. Масштабирование по времени играет решающую роль в расчете волатильности из-за уровня свободы выбора окна частоты/интервала. Таким образом, RMSPE поможет фиксировать степень ошибок по сравнению с желаемыми целевыми значениями лучше, чем другие показатели. Кроме того, RMSPE наказывает за большие ошибки больше, чем обычный MAPE (средняя абсолютная ошибка в процентах).

RMSE и RMSPE будут отслеживаться по производительности различных моделей при прогнозировании проверочного набора, чтобы показать их способность обобщать данные вне выборки. Поскольку обе эти метрики указывают уровень ошибки, цель состоит в том, чтобы постепенно уменьшать их значения с помощью различных структур модели и итераций.

Базовые модели

Были созданы две разные простые базовые модели для сравнения с более поздними моделями. Эти две простые модели основаны на двух основных характеристиках волатильности:

- **Средняя базовая модель** : волатильность в долгосрочной перспективе, вероятно, будет **означать возврат** (это означает, что она будет близка к историческим долгосрочным средним значениям)



- **Наивное прогнозирование случайных блужданий**волатильность завтра будет близка к сегодняшней (**кластеризация**)

GARCH Модели

GARCH означает **обобщенную авторегрессионную условную гетероскедастичность** , которая является расширением модели ARCH (авторегрессионная условная гетероскедастичность).

GARCH включает термины дисперсии запаздывания с остаточными ошибками запаздывания от среднего процесса и представляет **собой традиционный**

эконометрический подход к прогнозированию волатильности финансовых временных рядов .

Математически GARCH можно представить следующим образом:

$$\sigma_t^2 = \omega + \sum_i^q \alpha_i \epsilon_{t-i}^2 + \sum_1^p \beta_i \sigma_{t-i}^2$$

в котором σ_t^2 дисперсия на временном шаге t ϵ_{t-i}^2 и остатки модели на временном шаге $t-1$

GARCH(1,1) содержит только запаздывающие члены первого порядка, и математическое уравнение для него:

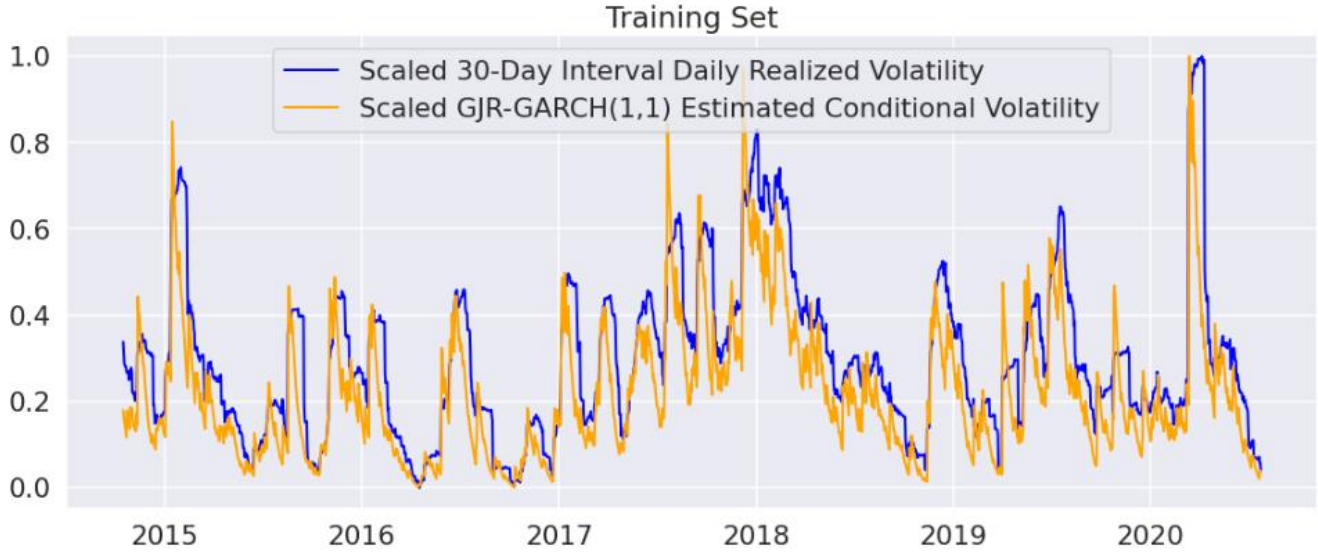
$$\sigma_t^2 = \omega + \alpha \epsilon_{(t-1)}^2 + \beta \sigma_{(t-1)}^2$$

где α β ω ω, и сумма до 1, и долгосрочная дисперсия.

(Синклер (2020))

GARCH обычно считается проницательным улучшением наивного предположения, что будущая волатильность будет такой же, как в прошлом, но также считается, что некоторые эксперты в области волатильности сильно переоценивают его как предиктор. Модели GARCH фиксируют основные характеристики волатильности: кластеризация и возврат к среднему.

Среди всех вариантов семейства GARCH, которые я создал, **TARCH(1,2)** с методом прогнозирования **Bootstrap** смог достичь самых низких значений RMSPE и RMSE в проверочном наборе.



Нейронные сети

Хотя GARCH остается золотым стандартом для прогнозирования волатильности в традиционных финансовых учреждениях, в последние годы все больше профессионалов и исследователей обращаются к машинному обучению, особенно нейронным сетям, чтобы получить представление о финансовых рынках.

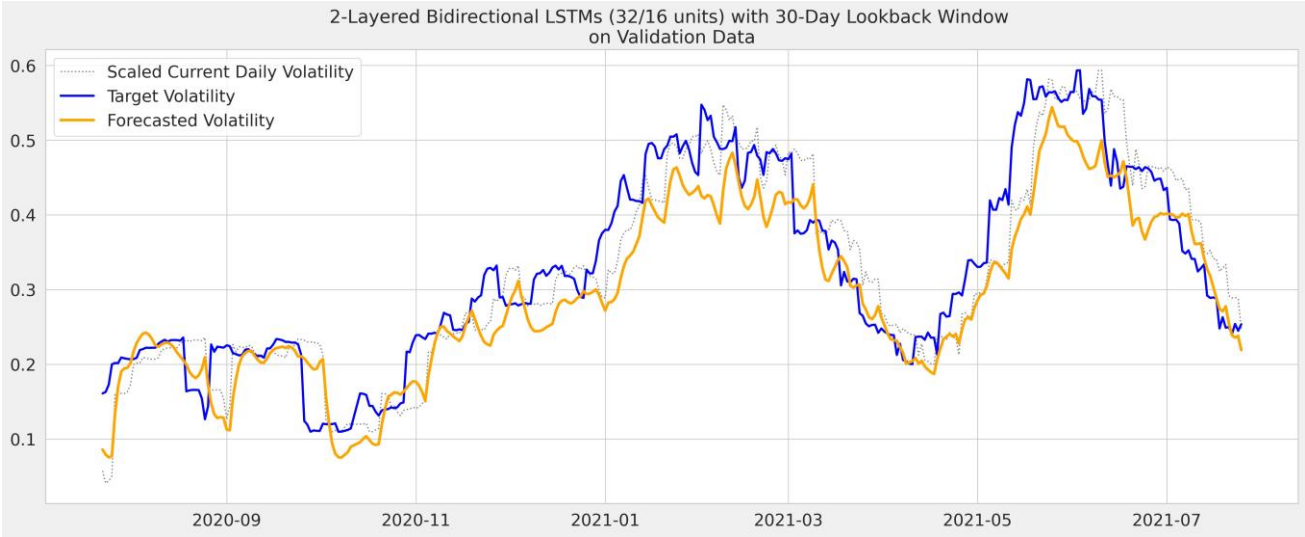
Одномерный двунаправленный LSTM

Двунаправленный LSTM является расширением обычного LSTM. Поскольку все временные интервалы входной последовательности уже доступны, двунаправленный LSTM может обучать 2 вместо 1 LSTM на одной и той же входной последовательности:

- 1-й на входах как есть
- 2-й на обратной копии входов

Это могло бы помочь обеспечить дополнительный контекст для сетей и обычно приводит к более быстрому и полному изучению проблемы.

После экспериментов с различными архитектурами нейронных сетей я обнаружил, что простая двухуровневая двунаправленная модель LSTM с 32 и 16 единицами превосходит все остальные, включая лучшую найденную модель GARCH.



Окончательная модель

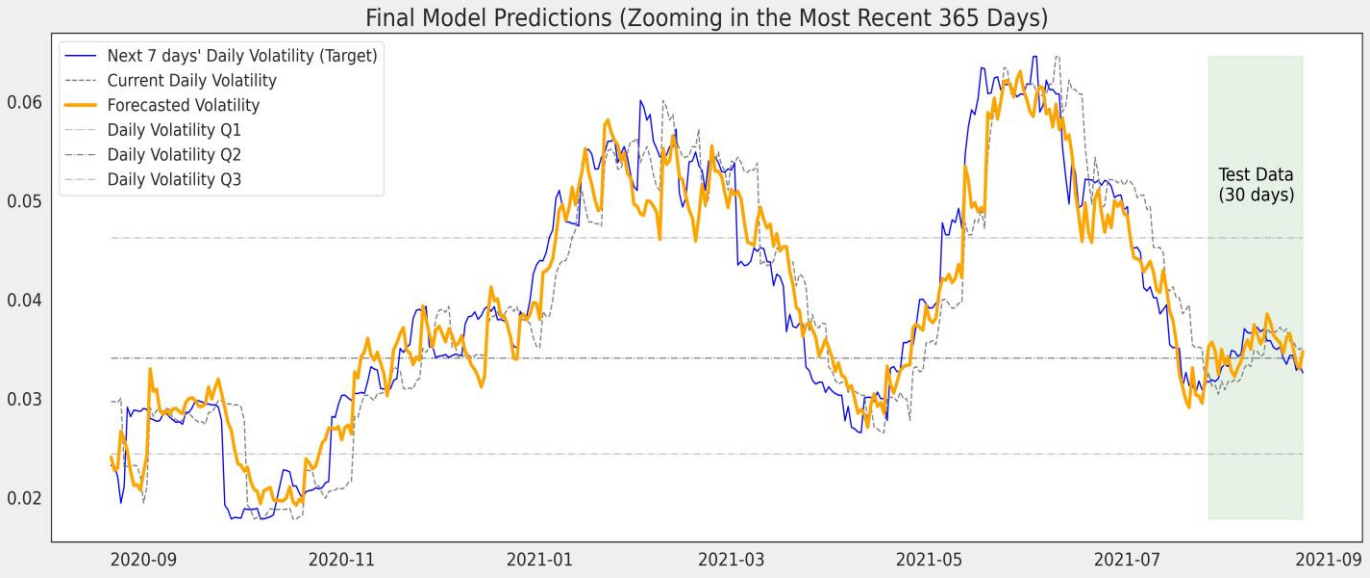
Многомерный LSTM

Для финансовых данных использование только одномерных данных, вероятно, недостаточно. Это может быть причиной того, что большинство из вышеперечисленных моделей не дали лучшего результата, чем наивное прогнозирование. Неважно, сколько нейронов или скрытых слоев используется, или насколько сложна архитектура модели, неадекватные данные не приведут к лучшим результатам. Поэтому я решил создать еще один набор моделей LSTM, но

многовариантных (это означает, что они могут обрабатывать другие функции, кроме самой волатильности).

Окончательная архитектура модели

Самая эффективная многомерная модель — это простые двухуровневые двунаправленные LSTM с 32 и 16 единицами с использованием окна ретроспективного анализа `n_past` в 30 дней и `batch_size = 64` . Кроме того, после каждого скрытого слоя LSTM есть 2 слоя Dropout с 0,1.



Заключение

	Model	Validation RMSPE
0	Mean Baseline	0.507040
1	Random Walk Naive Forecasting	0.224657
2	GARCH(1,1), Constant Mean, Normal Dist	0.530965
3	Analytical GJR-GARCH(1,1,1), Constant Mean, Skewt Dist	0.276680
4	Bootstrap TARCH(1,1), Constant Mean, Skewt Dist	0.209534
5	Simulation TARCH(1,1), Constant Mean, Skewt Dist	0.215768
6	Bootstrap TARCH(1, 2, 0), Constant Mean, Skewt Dist	0.201579
7	Simple LR Fully Connected NN, n_past=14	0.230476
8	LSTM 1 layer 20 units, n_past=14	0.218641
9	2 layers Bidirect LSTM (32/16 units), n_past=30	0.201927
10	1 Conv1D 2 Bidirect LSTM layers (32/16), n_past=60, batch=64	0.221937
11	2 Bidirect LSTMs (32/16), n_past=30, batch=64, SGD lr=5.9e-05	0.452836
12	Multivariate Bidirect LSTM 1 layer (20 units), n_past=30	0.239090
13	Multivariate Bidirect LSTM 3 layers (64/32/16 units), n_past=30	0.200929
14	Multivariate Bidirect LSTM 3 layers (64/32/16 units), n_past=30	0.208655
15	Multivariate Bidirect LSTM 3 layers (64/32/16 units), n_past=30	0.186887
16	Multivariate 4 Bidirect LSTM layers (128/64/32/16 units), n_past=30, batch=64	0.163791
17	Multivariate Bidirect LSTM 1 layer (20 units), n_past=30	0.292304
18	Multivariate Bidirect LSTM 3 layers (64/32/16 units), n_past=30	0.204601
19	Multivariate Bidirect LSTM 3 layers (64/32/16 units), n_past=30	0.161375

Что касается производительности на проверочном наборе (23.07.2020 — 25.07.2021), моя последняя модель LSTM имеет RMSPE 0,161, что примерно на 4,42% лучше, чем у наиболее эффективного варианта найденных моделей GARCH — TARCH. (1,2) с RMSPE 0,200954. Трейдеру не нужно делать абсолютно точный прогноз, чтобы иметь положительное ожидание при участии в рынках, ему просто нужно сделать прогноз, который является **одновременно правильным и более правильным, чем общий консенсус** . Поскольку GARCH по-прежнему остается самой популярной моделью прогнозирования волатильности, многомерный LSTM потенциально может дать инвесторам преимущество с точки зрения более высокой точности прогнозирования.

Окончательная модель LSTM имеет RMSPE 0,0534 на тестовом наборе (это самые последние 30 дней, из которых данные о будущей волатильности доступны для

сравнения). Поскольку RMSPE указывает среднюю величину ошибки по отношению к фактическим значениям, RMSPE, равное 0,0534, будет означать точность величины 94,65% при прогнозировании среднесуточной волатильности за 7-дневный горизонт в период с 26.07.2021 по 08.08.2021. /24/2021.

Однако, поскольку данные финансовых временных рядов постоянно развиваются, ни одна модель не сможет постоянно прогнозировать с высоким уровнем точности. Средний срок службы модели составляет от 6 месяцев до 5 лет, и в количественной торговле существует явление, называемое **альфа-распадом** , которое представляет собой потерю прогностической способности альфа-модели с течением времени. Кроме того, согласно Sinclair (2020), исследователи обнаружили, что публикация нового «преимущества» или аномалии на рынках снижает его доходность до 58%.

Поэтому эти модели требуют постоянной корректировки и настройки на основе самой последней доступной информации, чтобы быть уверенными, что они остаются актуальными и учатся развиваться вместе с рынками.

Использованная литература:

1. Жерон, А. (2019). *Практическое машинном обучении с помощью Scikit-Learn и TensorFlow: концепции, инструменты и методы создания интеллектуальных систем*. О'Рейли Медиа, Инк.
2. Синклер, Э. (2020). *Позиционная торговля опционами: расширенное руководство*. Джон Уайли и сыновья.
3. https://algotrading101.com/learn/yfinance-guide/
4. https://www.coursera.org/learn/tensorflow-sequences-time-series-andprediction/supplement/DM4fi/convolutional-neural-networks-course
5. https://insights.deribit.com/options-course/
6. https://arch.readthedocs.io/en/latest/univariate/univariate_volatility_forecasting.html
7. https://www.investopedia.com/terms/v/vix.asp