**Введение**

Вашему вниманию представляется работа на тему «СРАВНЕНИЕ ПРОГНОЗНЫХ СВОЙСТВ ЭКОНОМЕТРИЧЕСКИХ МОДЕЛЕЙ И МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ПРИМЕРЕ ЦЕН КРИПТОВАЛЮТ».

Актуальность исследования обусловлена растущим интересом к криптовалютам. Криптовалютный рынок характеризуется высокой волатильностью и сложными нелинейными зависимостями, которые делают традиционные методы прогнозирования недостаточно эффективными и надежными. Точные прогнозы цены важны для участников рынка для оптимизации торговых стратегий и управления рисками.

**Анализ предметной области и цель исследования**

Анализ предметной области основывался на изучении классических финансовых теорий, таких как портфельная теория Марковица, модель оценки капитальных активов (CAPM), арбитражная теория ценообразования (APT) и многофакторные модели Фамы-Френча, а также специализированных теорий и подходов (например, теория курсообразования и теории внутренней стоимости), адаптированных к особенностям криптовалютного рынка. Источники по данным темам были необходимы для понимания фундаментальных и теоретических основ ценообразования криптовалют, для выявления ограничений традиционных подходов к прогнозированию и перспектив применения современных методов.

Рассматривались актуальные эмпирические исследования на тему использования эконометрических моделей и моделей машинного обучения для прогнозирования цен криптовалют. Основываясь на результатах таких работ, были отобраны самые перспективные эконометрические модели и модели машинного обучения для проведения собственного исследования. Так, например, во всех работах самые точные результаты прогнозов показывали нейросети и ансамблевые модели машинного обучения, включая градиентный бустинг, в то время как эконометрические модели имели не такие точные прогнозы, но использовались как бенчмарк для других моделей. Также в разных исследованиях предлагалось использование различных внешних факторов, для улучшения прогнозных свойств моделей. В разных работах в модели добавляли макроэкономические факторы, индикаторы уровня активности в блокчейне, рыночного сантимента.

Цель исследования заключается в том, чтобы выявить подходы, обеспечивающие наиболее точные и стабильные прогнозы цен на высоко волатильном и трудно прогнозируемом криптовалютном рынке, а также определить значимость различных групп факторов, используемых в прогнозных моделях.

Для достижения указанной цели были поставлены и выполнены задачи:

* Провести обзор экономической теории и предпосылок эконометрических и машинных моделей;
* Провести обзор актуальных эмпирических исследований на тему использования эконометрических моделей и моделей машинного обучения для прогнозирования цен на криптовалюты;
* Отобрать потенциально лучшие модели и метрики оценивания их точности для проведения собственного исследования;
* Выбрать дополнительные индикаторы и факторы для потенциального улучшения прогностических способностей моделей;
* Провести собственное исследование с использованием отобранных методов на реальных исторических данных;
* Оценить точность прогнозов и выбрать лучшую модель для задачи прогнозирования абсолютных цен на криптовалюту.

**Дизайн Исследования**

На основе изученной литературы была составлена селекция моделей, которые использовались в моем исследовании. Были оценены эконометрические модели SARIMAX(2;1;2)(0;0;1)[7] для прогнозирования среднего уровня цены, GARCH(1;1) с распределением Стьюдента для моделирования и прогнозирования волатильности (именно такими оказались лучшие спецификации этих моделей). Также были обучены такие ML модели как RandomForestRegressor, а также регрессоры представителей градиентного бустинга XGBM, LGBM, CatBoost.

Для всех моделей были построены динамический и Рекурсивные прогнозы, а для оценки точности прогнозов и сравнения моделей использовались метрики RMSE, MAE, MAPE и вместе с визуальным анализом прогнозов.

Исследование проводилось на основе исторических дневных данных за период с 1 января 2020 по 16 апреля 2025 года. В качестве переменных для моделей первоначально были загружены различные криптовалютные и макроэкономические данные:

* **Криптовалютные данные**:
  + Цены Ethereum и Bitcoin, индексы доминации биткоина и USDT (BTC.D и USDT.D), общая капитализация крипторынка, индексы страха и жадности, цитируемости Ethereum в Google, среднедневная стоимость газа (комиссии за транзакции) и TVL в сети Ethereum.
  + Цена эфира на следующий день является целевой переменной прогнозирования.
* **Макроэкономические факторы**:
  + Доходности казначейских облигаций и векселей США, индексы S&P 500 и волатильности VIX, цены золота и нефти, индекс доллара США, а точнее его ETF-прокси.

Также были дополнительно сгенерированы некоторые переменные из этих данных. Например, дневной процентный показатель волатильности и логарифмическая дневная волатильность Эфира. В разделе эконометрического исследования были созданы лаговые значения для некоторых экзогенных переменных, а в блоке моделей машинного обучения – лаговые значения цены и скользящие средние цены Эфира, некоторые технические индикаторы и временные признаки.

Исследование построено следующим образом:

1. Загрузка Данных;
2. Предобработка и визуализация;
3. Эконометрическое исследование;
4. Исследование с использованием ML моделей;
5. Сравнение подходов и результаты исследования.

**Эконометрическое исследование.**

На первом этапе эконометрического исследования были выполнены проверки рядов на стационарность и сезонность, а также тесты на причинность по Грейнджеру для выбора лаговых значений экзогенных переменных.

Для соответствия предпосылкам модели SARIMAX, потребовалось сделать преобразования рядов согласно таблице (ознакомиться можно в презентации), чтобы гарантировать стационарность и отсутствие сезонных эффектов в данных. Использовалось дифференцирование рядов и сезонная декомпозиция.

В результате теста по Грейнджеру были выявлены значимые лаги некоторых экзогенных переменных, которые были включены в окончательные модели.

Данные были разделены на обучающую и тестовую выборку, после чего с учетом всех преобразований была оценена пробная модель SARIMAX с подбором оптимальной спецификации с помощью auto\_arima, после чего использовался пошаговый подход удаления незначимых переменных до момента, когда значение AIC не перестало снижаться.

Финальным шагом подготовки данных стало избавление от мультиколлинеарности с помощью использования VIF и корреляционной матрицы.

После этих шагов была оценена финальная модель с лучшей спецификацией SARIMAX(2;1;2)(0;0;1)[7]. Остатки получились неавтокоррелированы, распределение несимметрично. Результаты диагностики модели свидетельствуют, что модель адекватно описывает основные свойства остатков, но требуется формальная проверка на гетероскедастичность.

Результат Arch-Теста Лююнг-Бокса (ARCH test) (LM-stat: 217.22206699017 p-value: 4.076175005549618e-41) подтверждает наличие ARCH-эффектов, что обосновывает дальнейший переход к GARCH модели.

**Переход к GARCH модели.**

При построении модели GARCH(1;1) на остатках модели SARIMAX лучшее объяснение распределения случайных шоков продемонстрировала модель с распределением Стьюдента. Для окончательной спецификации было выбрано именно оно, по наименьшему показателю критерия Акаике, хотя сравнивались различные распределения.

Сумма alpha[1] + beta[1] ≈ 1.000 (0.0867 + 0.9133) в результатах диагностики говорит о медленно убывающей волатильности: шок одного дня «рассеивается» очень медленно. Параметр степеней свободы nu = 4,6334 демонстрирует выраженные тяжелые хвосты распределения остатков.

**Прогнозы.**

После этого финальные модели с лучшими спецификациями были оценены на тренировочной выборке. Были построены динамический прогноз модели SARIMAX + GARCH сразу на всю тестовую выборку в 30 дней и рекурсивный одношаговый прогноз с переоценкой модели на каждом шаге на объединении данных. Как не удивительно, рекурсивный прогноз прогноз обладает лучшей предсказательной силой, в чем можно убедиться проанализировав визуализацию и метрики точности прогнозов.

**Модели машинного обучения.**

В разделе моделей машинного обучения данные были разделены на обучающую, валидационную и тестовую выборки. Тестовая выборка составила те же 30 последних дней, что и в эконометрическом исследовании, а оставшиеся данные для тренировочной и валидационной выборок были поделены в соотношении 80 на 20 в хронологическом порядке .

Для отбора значимых признаков применялись 5 различных методов, среди которых RFE на Random Forest, LassoCV, Permutation Importance на Random Forest, XGBoost Importance и LightGBM Importance. Каждым методом отбиралось 20 самых важных признаков, а в итоговую модель вошли признаки, которые были отобраны 3 или более раз. Таким образом был отобран 21 признак.

Лучшие гиперпараметры моделей подбирались с помощью процедуры GridSearchCV с применением временных скользящих окон (TimeSeriesSplit) на 4 скользящих фолдах.

Далее модели были обучены с лучшими гиперпараметрами, для каждой построен out-of sample прогноз на валидационную выборку. Самые точные прогнозы показали модели RandomForestRegressor и LGBMRegressor, худший результат показала модель CatBoostRegressor, а вот результаты модели XGBMRegressor вызвали интерес, благодаря чему в исследовательских целях, она единожды использовалась и дальше, хотя в реальной ситуации была бы отброшена.

**Out-of-sample и рекурсивный прогнозы на тесте**

После этого лучшие модели были обучены на объединении тренировочной и валидационной выборок. Были построены классический для моделей машинного обучения out-of-sample прогноз, когда модель обучается один раз, а также рекурсивный одношаговый прогноз с дообучением моделей по мере получения новых данных на их объединении. Модели продемонстрировали высокую точность прогнозов. Однако, при рекурсивном методе точность прогноза случайным лесом выросла, а вот у LightGBM снизилась, что может сигнализировать о том, что модель начала переобучаться. Дообучение моделей призвано повысить стабильность и адаптивность прогнозов к изменяющимся рыночным условиям.

Интересным является то, что out-of-sample прогноз на тесте для XGBMRegressor имеет самые низкие значения метрик ошибок. Но, как уже обсуждалось ранее, результаты этой модели не рассматриваются из-за недостаточно точного прогноза на валидации, чтобы имитировать реальные условия.

**Динамический рекурсивный прогноз**

Также был построен динамический рекурсивный прогноз для более корректного сравнения с эконометрическими моделями. Здесь был реализован Pipeline посредством написания собственного класса, чтобы обеспечить обработку данных и подставление в модель предсказанных ей значений для прогноза на следующий период. Визуализация и метрики точности прогноза представлены в презентации. В этом случае результаты моделей машинного обучения оказались не эффективными и уступают нашей гибридной моделе SARIMAX + GARCH.

**Сравнение результатов**

Сравнивая результаты прогнозов эконометрических моделей и моделей машинного обучения, мы пришли к следующим выводам:

* В случае рекурсивного прогноза модели машинного обучения (RandomForest и LightGBM) превосходят эконометрическую модель SARIMAX+GARCH по точности прогнозов.
* Для динамического прогнозирования на среднесрочном горизонте преимущества эконометрических моделей становятся очевидными благодаря их устойчивости и интерпретируемости. Несмотря на более низкие значения метрик точности, визуальный анализ прогнозов моделей машинного обучения делает невозможным полагаться на такие прогнозы, особенно при большем горизонте прогнозирования.

**Результат исследования и выводы**

В результате проведенного исследования были получены следующие научные и практические выводы:

* Включение лаговых значений цены, макроэкономических и поведенческих факторов в прогнозные модели оказалось значимым.
* Модели машинного обучения обеспечивают более высокую точность прогнозов по сравнению с эконометрическими моделями при рекурсивном подходе.
* Модели машинного обучения не подходят для динамических прогнозов на длительный срок.
* Эконометрические модели остаются предпочтительным выбором для среднесрочных динамических прогнозов благодаря лучшей интерпретируемости и устойчивости.