Здравствуйте! Я Алексей Триголос и сегодня я защищаю свою выпускную квалификационную работу на тему «Прогнозирование стоимости акций на российском рынке с использованием машинного обучения и анализа текстовых данных».   
  
Моими задачами было следующее. Изучить опыт предыдущих работ по данной теме и продумать возможные пути решения. Получить тикеры бумаг и информацию об их стоимости, новости по этим эмитентам. Также мне было необходимо предобработать эти данные для построения моделей, обучить те самые модели и проанализировать насколько возможно прогнозировать цены российских бумаг и как на это влияет использование новостей.  
  
При анализе предыдущих работ, я видел использование разных моделей (линейных, лес, LSTM, ARIMA и многих других), метрики так же использовались различные, но в основном были MAPE и RMSE. Сами прогнозы делали на разные рынки, такие как Америки, Португалии, России, а некоторые были основаны под конкретные бумаги. Обычно в лидерах по прогнозированию была модель LSTM. Наилучший результат точности, что я находил был 96%. Использование новостей в прогнозировании улучшало показатели модели. Также были обнаружены корреляции цен на акции кампаний с ценами на драгоценные металлы, нефть, газ. И ещё было обнаружено, что использование большего количества новостей из разных категорий повышает точность предсказаний.  
  
Для получения данных именно по акциям российских компаний я сначала нашел известные тикеты, в общей сложности их получилось больше трех сотен, после чего используя библиотеку apimoex получил информацию по каждой из бумаг с дневными таймфреймами. Для новостей я выбрал РИА новости, по той причине, что там они присутствуют начиная с 10 года, от 200 до 900 новостей в день и идут на все темы. Новостей в общей сложности получилось более миллиона на всевозможные тематики.  
  
При обработке данных по ценам я обнаружил, что в некоторых бумагах могли быть пропуски как в начале, в конце, так и в середине, такие данные мне приходилось удалять, есть результаты на левом верхнем рисунке, или дозаполнять линейной интерполяцией, на правом верхнем. Только 90 бумаг из 330 были без пропусков, а в нескольких не было ни одного значения. Почти половина бумаг имело 2624 дня торгов, притом, что там присутствовали пропуски. При обычном удалении пропусков из начала для некоторых бумаг удалось сократить на 85%. И для обучения были выбраны только те акции, у которых процент пропуска не превышал 10 что привело к итоговым 216. Ещё на левом нижнем графике можно увидеть, что в 2014 году произошел резкий скачек ценных бумаг, скорее всего, для многих компаний просто не сохранились значения до этой даты. А на правом нижнем графике видно, что объем торгов в большей части был до 100 миллиардов в день.

Для получения новостей я посчитал имеющиеся на Kaggle и другие найденные не подходящими, поэтому собирал их сам. В большинстве случаев новости я считал полными как в примере снизу. Но иногда попадались опросники как слева или кастомные новости со своими полями как справа. Так же некоторые новости представляли из себя просто аннотации, которые я тоже считал не полными, но использовал для обучения. В общей сложности у меня получилось более 14 лет новостей, из которых чуть меньше 22 тысяч я принимал за не подходящие, что не должно сильно повлиять, ведь это менее одного процента. Для обучения я прогонял новости через Llama 3.3 70B с написанным мною промтом, для получения важности каждой новости на каждую компанию и краткого объяснения почему именно так. В результате я получал примерно следующее. Важность 0.542, «Новость о создании газового хаба в Турции может увеличить поставки газа в Европу, что положительно скажется на компании Газпром». Или -0.235, «Снижение цен на газ в Европе может привести к снижению доходов Газпрома».

Далее я бы хотел рассказать про выбранные модели. У меня уже имеются обученные Ridge и Random Forest Regression и сейчас я работаю над остальными. Например, я дорабатываю LSTM и собираюсь так же обучить модели используя полученные и обработанные мною новости. А для оценки я использую такие метрики как MAPE и RMSE, они лучше зарекомендовали себя в других работах и имеют удобную интерпретацию. MAPE считает средний процент отклонения прогнозирования от реального значения, а RMSE показывает среднее отклонение от реального значения в тех же единицах. Но так как у меня временные ряды, которые, очевидно лучше предсказывают ближайшие дни, я считаю метрики для каждого дня отдельно, в таком случае можно считать, что в формулах n равняется одному. Такой подход делает оценки метрик нагляднее и оставляет возможность считать среднее на желаемом горизонте. С ценными бумагами тяжело работать, потому что это временные ряды с ненормированными отсечками данных и большим влиянием новостной информации. Для своих моделей я добавлял лаги по дням, неделям, месяцам и нескольким годам, использовал GridSearchCV с TimeSeriesSplit, проверял, чтобы предсказания строились только на данных из уже известного прошлого и при дальнейших предсказаниях использовал полученные ранее моделью значения.

Ridge Regression. Это модель множественной линейной регрессии. Здесь я привел несколько графиков, наиболее интересные правые и левый нижний. Так вот правые картинки как будто получили очень плохие предсказания, MAPE доходит до 80 и видно, что предсказания очень отдаляются. Но хотел бы заметить, что такой бурный рост был просто непредсказуем, но несмотря на это первые два месяца прогнозирования были с ошибкой меньше 5. А слева снизу представлен средний график по регрессионным предсказаниям для всех эмитентов. Результат мог бы быть лучше, но не все можно предсказать, например тот же сплит акции 1 к 100, поэтому подобные биржевые действия могут очень сильно подпортить результат предсказаний.

(помогает справиться с проблемой мультиколлинеарности и переобучения модели. Этот метод добавляет штраф к функции потерь, что позволяет улучшить обобщающую способность модели.

Мультиколлинеарность: Это ситуация, когда независимые переменные в модели линейной регрессии сильно коррелируют друг с другом. Это может привести к нестабильным оценкам коэффициентов и затруднить интерпретацию модели.

Регуляризация: Это техника, используемая для уменьшения переобучения модели путем добавления штрафа к функции потерь. В ридж-регрессии используется L2-регуляризация, которая добавляет квадрат нормы коэффициентов к функции потерь.)

Также я обучил случайный лес. Это метод машинного обучения, который основан на ансамблевом подходе, который объединяет множество деревьев решений для улучшения точности и устойчивости модели. Для обучения модели производились те же самые действия подготовки. А результаты получились очень даже различными, например на верхнем слева результат получился достаточно точным, а верхний справа выглядит странным из-за таких перепадов. Вероятно, проблема в том, что я для каждой бумаги проставлял одинаковые возможные параметры, что могло повлиять на результаты. Но несмотря на это видно, что средняя ошибка получилась меньше.

(Деревья решений: В основе Random Forest лежат деревья решений, которые представляют собой модели, принимающие решения на основе последовательности вопросов. Каждое дерево в лесу обучается на случайной подвыборке данных, что помогает избежать переобучения.

Ансамблевый метод: Random Forest использует ансамблевый подход, что означает, что он объединяет предсказания нескольких деревьев для получения более точного результата. Это достигается путем усреднения предсказаний всех деревьев, что снижает вероятность ошибок.

Случайная подвыборка: При обучении каждого дерева используется случайная подвыборка как данных, так и признаков. Это позволяет деревьям быть менее коррелированными и, следовательно, более разнообразными.)

Хотел бы показать промежуточные результаты. Здесь приведена таблица с средними ошибками по всем обученным бумагам. Можно видеть, что в среднем Ridge Regression показывает результат на первый день чуть лучше, чем случайный лес. Но если смотреть на более дальний срок, то Random Forest Regression показывает более устойчивый результат. В будущем я добавлю сюда результаты LSTM модели и гибридных тоже.

Архитектура приложения состоит из следующих элементов. Принимается запрос на https на балансировщик нагрузки Nginx, добавлен удобный доступ есть через Streamlit. Он служит простым интерфейсом для получения информации, выполняя запросы на апишку, которая так же доступна для пользователей, но не столь удобна. Получая запрос FastAPI его обрабатывает и делает обращение в облачное хранилище S3 Яндекса, из которого возвращаются необходимые данные. После чего они могут быть немного обработаны и переданы на отображение в Streamlit. Ну и, конечно, всё это разворачивается с использованием Docker Compose.

Здесь я демонстрирую как раз саму работу сервиса. На данный момент есть множество тикетов компаний, возможность выбрать модель, все данные будут обновляться сразу же. Будут предоставлены в виде таблицы сами предсказания и посчитанные ошибки. И есть множество графиков, с обучаемыми данными, только валидационный с предсказанием, и по каждой из ошибок, потому что размерности могут быть очень разными, и для всех графиков кроме первого по возможности добавлены временные отрезки на неделю, месяц, год и на весь период предсказания.

К защите я хочу закончить обучение модели, основанной на новостях, кроме этого, есть желание сделать несколько гибридных моделей из уже имеющихся и новостной. Так же добавлю анализ, по полученным прогнозам, и сравню результативности для определения лучшей модели. И хочу еще немного доработать сервис для возможности выбора отображаемых данных, а не всех найденных. На данном этапе можно видеть, что модели неплохо прогнозируют ближайшие изменения цен, но хотелось бы улучшить этот результат.

Спасибо за внимание.