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 пропуски на 85%. И для обучения были выбраны только те акции, у которых процент пропуска не превышал 10 что привело к итоговым 215. На нижних графиках виден объем торгов, и что общая цена бумаг за сутки в большей части был близка к сотне 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. Я ориентировался на РИА новости по причине того, что там каждый день от 200 до 900 новостей начиная с 10 года. Я получил все ссылки на новости, распарсил каждую из них, получив заголовки, текст, цитаты и аннотации. 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. Эти данные я закидывал в Llama 3.3 70B с написанным промтом, для получения важности каждой новости на каждую компанию и краткого объяснения почему именно так. Кроме этих данных я передавал тикеры компаний, сказал обращать внимание на влияние в смежных областях и не учитывать посторонние события. 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».

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.

Далее я бы хотел рассказать про выбранные модели. Я обучил SARIMAX, Ridge, Random Forest Regression, XGBRegressor и LSTM. Так же я обучал эти модели с использованием обработанных мною новостей. 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. В основном я использовал GridSearchCV с TimeSeriesSplit, проверял, чтобы предсказания строились только на данных из уже известного прошлого и при дальнейших предсказаниях использовал полученные ранее моделью значения. Исключениями были только SARIMAX, в котором для обучения я использовал только данные о цене компании и LSTM, который состоял из нескольких LSTM слоёв с нормализациями, dropout’ом, слоем активации и выходным. Для прогнозирования с использованием новостей основной смысл оставался в тех же вещах, но были добавлены максимальная и минимальная важности новостей за день, их общее число, среднее и время новостей. Для этих данных тоже добавлялись лаги. Кроме того, в таких моделях было больше данных для предсказаний, ведь каждый новый день предсказания я использовал для следующих.

На следующем слайде я привожу несколько примеров работы модели. В данном случае выбрал компанию с тикером ZILL. На данных графиках можно видеть прогнозы каждой из моделей. Видно, что каждая из моделей показывает какие-то реальные значения. На верхнем левом и правом нижнем приведены модели с использованием новостей, им даже удается угадывать увеличение цены, но более интересным кажется сравнение правых графиков, ведь тут представлена одна и та же Ridge модель, только сверху вариант обученный без новостей, а снизу тот, который использовал их и в данном случае хорошо видно, как нижняя с новостями имеет гораздо большую правдивость прогнозов.

Что касается итоговых результатов, то я решил привести таблицу средних значений MAPE для каждой из моделей. Я брал промежутки на день, неделю, две, месяц и квартал. Я выделил жирным лучшие средние результаты на каждый из периодов. Здесь можно видеть, что рекордсменом становится Random Forest. Она отстает только от Ridge на первые пару дней, но дальше обгоняет остальных соперников, ближайшей к ней лучшей моделью стал XGBRegressor. В среднем её результат чуть хуже леса. Кроме этого видно, что в среднем модели с новостями показывали результаты хуже, чем без них. Это происходит по той причине, что новости добавляют много признаков, но для малопопулярных акций тяжело понять влияние новостей, ведь они реже обсуждаются. Поэтому модели с использованием новостей лучше использовать на тех бумагах, которые чаще появляются в новостном фоне. Но даже так результат неплохой, лучшей из таких моделей является XGBRegressor на долгий горизонт и Ridge до пары недель.

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. Добавлено множество тикетов компаний, возможность выбрать модель, все данные сразу будут обновляться. Есть предоставление в виде таблицы, так же 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. Кроме этого, предоставлена информация о средних результатах модели и есть топ моделей на определенные горизонты по среднему занимаемому месту.

Результатами моей работы получился сервис и обученные модели с хорошей предсказательной способностью. На ближайшие пару дней лучше выбирать Ridge модель. Дальше более стабильный результат даёт Random Forest, а XGBoost может составлять ему конкуренцию. SARIMAX достаточно простая модель, но и менее точная, поэтому я бы лучше на неё не расчитывал. LSTM показал малую эффективность, для улучшения прогнозов необходимо дорабатывать структуру. Новостные данные полезны для прогнозов, но лучше их использовать на акциях с большим упоминанием в новостях. В таких случаях следует выбирать Ridge или XGBoost модели.

Спасибо за внимание.