NIS Homework 1

Евланов Максим, Семенов Владислав

Ноябрь 2023

Содержание

1 Вступление

Выбирая датасет для анализа, мы искали что будет совмещать интересную тему и написание интересной модели, а поскольку фондовая биржа и торговля акциями одна из самых интересных и сложно прогнозируемых вещей, то наш выбор пал на датасет, основанный на индексе Доу-Джонса (DJIA), включающем в себя акции тридцати крупнейших американских компаний, торгуемых на бирже.

Цель нашего проекта - провести анализ датасета: выполнить предобработку данных, разведочный анализ, а также использовать модели для прогнозирования будущих цен акций. Мы будем использовать различные библиотеки (scikit-learn, catboost, TensorFlow), модели (LinearRegression, Polynomial, Preeprocessing, LSTM), метрики (MSE, MAE, RMSE, R^2).

Данные для анализа будут включать значения DJIA за первые два квартала 2011 года, вместе со всеми стандартными для акций данными -ценя закрытия и открытия, высшей и нисшей точек дневных торгов, объема торгов, данный по дням до выплаты дивидендов и другие.

2 Предобработка данных

Самый первый шаг - считывание данных. Для этого используем библиотеку pandas, позволяющая удобно работать с таблица и датасетами. Для считывания данных из файла используем функцию фильтр на строки с нужными кварталами. Возьмём заранее подготовленные цвета

```
pd.read_csv
```

и распечатываем считанную таблицу.

Из распечатанной таблицы видно множество характеристик, такие как цена на открытии, закрытии, самая высокая и самая низкая за неделю, объём, дивиденды и другие. Все они разбиты по компаниям и по неделям. Все эти данные будут использованы для предсказания процента изменения цена акции на следующей недели.

Однако для удобной работы необходимо обработать данные. Из таблицы явно видно, что присутствуют пустые значения NaN, которые лишь помешают настроить модель. Удаляем такие строки при помощи

```
.dropna()
```

.

Также перед ценами стоит валюта, что также мешает корректно работать с данными, так что очищаем данные от знака "$"при помощи

```
.str.replace('$', '', regex=True).astype(float)
```

.

Следующим шагом определяем столбцы, что пригодятся для прогнозирования процента изменения цена акции на следующей недели при помощи линейной регрессии. Откидываем все столбцы, что не содержат полезной информации - название тикера акции, квартал, дата - они не несут полезной информации при прогнозировании.

Теперь разделим данные на две части: X и y.

3 Разведочный анализ данных

Exploratory data analysis (EDA) - разведочный анализ данных для обнаружения и анализа основных свойств данных. Для начала отобразим гистограмму процента изменения цена акции на следующей недели при помощи библиотеки matplotlib.pyplot. Заметим, что распределение близко к нормальному (с небольшим перевесом в положительных значениях), что соответвует примерному положению дел на бирже.

А теперь взглянем на основные метрики для наших данных.

Теперь к визуализации. Будем использовать библиотеку plotly. Для начала отобразим изменение цены акции в первом и во втором квартале. Подготовим две таблицы, применив фильтр на строки с нужными кварталами. Возьмём заранее подготовленные цвета

```
px.colors.qualitative.Plotly.
```

Далее отобразим графики при помощи функции, используя

```
go.Figure()
```

. Данное выражение позволяет добавлять подграфики, которые будут визуализированы при помощи функции for и функции

```
.add_trace(go.Scatter())
```

. Внутри задаём x, y, имя, а также

```
mode="markers+lines"
```

, что позволяет отображать линию и задать параметры marker и line. Эти параметры позволяют отделить каждую линию своим цветом.

Далее, реализуем интересную опцию по анализу котировок акций: визуализируем одну из самых известных биржевых терминов - Свечи. Для самых дорогой акции - IBM, самой дешевой - BAC, самый выросшей акции (конец-начало) и самой упавшей (найдем их также).

Для начала подготовим данные для визуализации. Для этого найдем самую выросшую в процентах акцию и самую упавшую.

Покажем на двух графиках рост или падение акции за весь период, данный в датасете. Из полученного графика видно, что сильнее всего выросла компания IBM, а упала - компания CSCO.

Теперь вернёмся к свечам. Стоит отметить, что IBM - это одновременно самая дорогая и самая прибыльная акция за весь период. Именно по этой причине график для неё был выведен только один раз.

4 Анализ данных

4.1 Линейная регрессия

Перейдём к самой самой модели. Здесь мы используем ленийную регрессию из библиотеки sklearn.

Принцип работы линейной регрессии: Модель представляет линейную зависимость одной переменной (y) от нескольких других факторов (переменных) X. Иначе это функция от нескольких переменных, каждая умноженная на свой вес, вида:

$$f(x) = w_0 + w_1 * X$$

Где w_0 - отдельный "линейный"коэффициент, а w_1 - вектор весов переменных. Задача сводится к тому, чтобы подобрать наилучшие коэффициенты (веса) для наилучшего предсказания y.

Любую модель необходимо для начала обучить. Для этого разбиваем данные, что были предоставлены на обучающую и тестовую выборку. Идея данного разбития состоит в том, чтобы правильно оценить эффективность нашей модели, так как при обучении на всех данных модель подберёт данные под них. А задача поставленная перед моделью - найти зависимости для всех данных, а не только доступных нам.

При помощи

```
.fit
.predict
```

обучаем модель на обучающей выборке и получаем предположительный y, полученный из нашей модели.

Для оценки точности нашей модели используем среднеквадратичную ошибку и среднюю абсолютную ошибку.

Обе метрики чаще всего используются для оценки моделей линейно регрессии. Главное отличие двух метрик - в том, что MSE лучше показывает ошибку при разнице более одного, а MAE удобнее использовать в задачах, где разница меньше единицы.

Результаты получились довольно хорошие: MSE около 2.635975013777764, MAE 1.0422403129819309. (Данные могут обновиться в силу случайности разбития выборки).Теперь визуализируем полученные результаты на примере акции компании Intel. Из графика видно, что полученное предсказание процента изменения цены акции имеет погрешность, но при этом имеет схожую линию графика.

4.2 Catboost

В качестве второй модели будем использовать регрессию из библиотеки catboost. Принцип работы подобен прошлой модели, однако она более гибкая для настройки. Простая оценка модели дала более худший результат, чем в прошлой.

Далее пробуем использовать настройки модели. Используем grid search. Данная функция находит оптимальные веса при помощи перебора. При этом функция тратит много времени на работу (около 7 минут в данном случае).

Полученный результат уже лучше, но хуже самой первой модели. Вероятнее всего, данную модель можно улучшить и получить лучший результат.

4.3 TensorFlow

В данном случае будем использовать модель Долгой краткосрочной памяти (Long Short Term Model — LSTM) из библиотеки TensorFlow. Данная модель является видом рекуррентной нейросетью (Recurrent Neural Networks — RNN). Здесь идея в том, что данный вид нейтосетей сохраняет информацию о прошлых событиях.

Однако из-за подобного некоторые коэффциенты могут "исчезать ушудшая результат обучения. Именно для этого и создана данная модель, так как решает данную проблему при помощи учитывания "короткой"памяти. При этом для оптимизации используем Adam, что представляет собой объединение AdaGrad (для улучшение результатов при расхождении весов) и RMSProp (убирает шумы и иные выбросы).

При обучении добавляем слои, чтобы модель точнее устанавливала зависимости. Данная технология помогает все входящие данные обработать сразу много раз весами, при чём комбинирую их. Таким образом, мы получаем более точные зависимости. Технология эпох помогает найти более оптимальный результат при подборе весов.

Как видно из ошибки на тестовых данных, данная модель также не даёт хорошего результата для нашей задачи. Ошибка оказалась больше, чем в первой модели. Вероятнее всего при лучшей оптимизации можно добиться лучшего результата.

Также, возможно, стоило иначе подавать данные в данную модель для лучшего результата.

5 Заключение

В завершение нашего проекта по анализу данных с использованием датасета Dow Jones, мы сталкиваемся с тем, что точное предсказание движения рынка акций - задача исключительно сложная. Финансовые рынки подвержены множеству влияющих факторов, включая экономические показатели, политическую обстановку и даже психологию инвесторов. Однако, мы нарисовали много красивых графиков и хорошо провели время.