**UPDATE:** код для получения фичей в jupyter notebook. Поменяла способ подсчета и оформление некоторых фичей. Изменения отмечены курсивом.

**Задача 1.**

Имеется торговая история трейдера на фондовой бирже (формат в таблице 1 ниже),

а также история финансовых, политических, и прочих глобальных событий (формат в таблице 2 ниже).

Необходимо: сформировать фичи описывающие трейдера (придумать)

Например, горизонт инвестирования, склонность к риску, сумма инвестиций, интересы.

**Решение.**

Из указанных данных можно извлечь множество фичей, но я сфокусируюсь на тех, которые представляются наиболее значимыми для описания трейдера.

В дальнейшем буду предполагать, что все записи торговой истории включают в себя данные как об открытии, так и о закрытии позиции (в случае, если закрытия не произошло, всегда можно обработать пропуски). Также буду считать, что все события, представленные в таблице, – систематические, т.е. влияют на весь рынок в целом, а не на конкретный сектор.

Ещё один важный вопрос – какую дополнительную информацию, не представленную в этих таблицах, мы можем использовать. Кажется, что не должно быть никаких препятствий, чтобы использовать информацию из открытых источников. В любом случае, я предложу также альтернативу, полученную исключительно из предоставленных данных (и в этом случае я буду особенно часто делать предположения, основанные на специфике предметной области, что, конечно, лучше делать при консультации специалиста).

Понятно, что в реальной жизни перед feature engineering’ом стоит предварительно пообщаться с профессиональным финансовым консультантом, который мог бы подсказать какие-то неочевидные фичи, выбрать наиболее важные и в целом помочь data scientist’у сориентироваться в области, но у меня такой возможности нет, так что буду руководствоваться своими скромными познаниями и здравым смыслом).

**Трейдер**

|  |  |  |
| --- | --- | --- |
| **Атрибут** | **Пример** | **Комментарий** |
| **Id аккаунта** | 111 | (без изменений) |
| **Сумма инвестиций** | 100000,39 USD | объём \* цена for позиция in торговая\_история |
| ***Коэффициент окупаемости*** | 1.2 | 1 + (цена при закрытии – цена при открытии)/ цена при открытии, рассчитаны по всей торговой история (для записей с закрытой позицией) |
| **Интересы** | **в случае, если используем данные о том, к какому сектору относится каждый инструмент:**  *n фичей вида* 0.01, 0.32, 0, …, где n – количество секторов, каждая фича – доля позиций i-го сектора в портфеле  **если данных о секторах нет:**  *n фичей вида* 0.01, 0.32, 0, …, где n – количество уникальных инструментов, каждая фича – доля позиций i-го инструмента в портфеле | доля в портфеле рассчитывается на как объём \* цена/общий\_объём.  Общий\_объём может высчитываться как по всей истории, так и по финальному периоду. В первом случае мы можем не учесть, что у трейдера изменились интересы, во втором – проигнорировать важные исторические данные. Я бы остановилась на первом варианте.  Предпочтительнее использовать вариант с секторами, потому что так удаётся получить намного более сжатое и интерпретируемое представление, что в результате упростит процесс кластеризации. |
| **Склонность к риску** | **если есть данные о волатильности *секторов*:**  волатильность портфеля: 0.1  **если дополнительных данных нет:**  доля акций в портфеле: 0.5  также можно использовать:  доля коротких сделок, доля маркет-ордеров, доля сделок, совершенных во время негативных событий, средний срок удержания позиций, коэффициент окупаемости | Волатильность портфеля рассчитывается с помощью корреляции между *секторами* как  , – доля *сектора* в портфеле, – ковариация между *секторами*  Если данных о волатильности нет, то о склонности к риску также можно судить по доле акций в портфеле (чем больше акций, тем агрессивней и более склонен к риску трейдер), а также по доле коротких сделок, маркет-ордеров, сделок во время негативных событий, коэффициенту окупаемости (чем больше, тем рискованней) и по сроку удержания позиций (чем меньше, тем рискованней).  При отсутствии консультации специалиста или примеров размеченных данных вручную эти признаки комбинировать не стоит, а лучше рассматривать каждый из них как отдельную фичу, вычислить корреляцию между ними и убрать лишние зависимые признаки. |

При подсчёте всех цен я предполагаю, что они рассчитываются после вычета комиссии.

Есть также фича, которую я не рассматривала – горизонт инвестиций. На данный момент я не вижу способа извлечь эту информацию из предоставленных данных (кроме как посчитать время от первого открытия до последнего закрытия, но это не очень разумно, потому что данные могут быть неполными, инвестор может только начать осуществлении стратегии и т.д.). Косвенно о горизонте можно судить по склонности к риску (как правило трейдеры с долгосрочными целями более склонны к риску, т.к. могут позволить себе кратковременные), но в таком случае создавать дополнительную линейно зависимую фичу не имеет смысла.

Ещё я думала о добавлении фичи с весами компаний различной капитализации (10% маленьких компаний, 20% корпораций и т.д.) и возможное использование этой фичи для оценки склонности к риску, однако вероятно, что включение в портфель бумаг с низкой капитализацией может быть сделано с целью диверсификации. Поэтому я решила, что без консультации специалиста такие фичи выделять не стоит.