ML Live

Предсказание доходности ценных бумаг (crowdAI IEEE Investment Ranking Challenge).

4-е место

Кирилл Романов

Москва, 16.06.2018

План выступления

1. Мой опыт и мотивация для участия
2. Обзор соревнования и правил
3. Общий обзор решения
4. Ключевые идеи:
    - Подготовка данных и генерация признаков
    - Моделирование, стекинг
5. Анализ лучших моделей
6. Интересные идеи от первого и второго места
7. Общие выводы по итогам участия

1. Мой опыт и мотивация

- Специалист по экономике (Финэк, специализация «Международные финансы»)
- Основная деятельность – управленческий консалтинг (стратегия, supply chain management, финансы)
- Когда то давно писал торговые стратегии как хобби
- Сейчас много времени уделяю DS и ML и было интересно попробовать применить разные модели к вроде бы понятной области…

2. Обзор соревнования и правил. Вступление

… Но в итоге, данные были так переработаны, а правила выстроены так, что все свелось к стандартному алгоритму:

2. Обзор соревнования и правил. Данные

Дана финансовая информация с 1996 по 2016 год, один период – полугодие. Всего 42 периода. Внутри примерно по 900 акций компаний с наибольшей капитализацией. Задача проранжировать акции по доходности

1 2 3 4 5

	time_period	index	Train	Norm_Ret_F6M	Rank_F6M	X1_1	X1_2	X1_3	X1_4	X1_5	...	X69_3	X69_4	X69_5	X69_6	X70_1	X70_2	X70_3	X70_4	X70_5	X70_6
0	1996_2	1996_2_lo2py80q	1	-0.164343	563	-0.581405	0.324594	0.606611	0.508565	0.304816	...	0.00218	0.002179	0.00218	0.00218	NaN	NaN	NaN	NaN	NaN	NaN
1	1996_2	1996_2_c0lbkx5l	1	0.159314	402	0.355752	0.483632	0.885929	-0.091519	2.189224	...	0.002179	0.00218	0.00218	0.00218	NaN	NaN	NaN	NaN	NaN	NaN
2	1996_2	1996_2_awxeoifz	1	0.931337	131	-1.232963	0.032044	0.212912	-1.001768	0.849036	...	0.002178	0.002178	0.002179	0.002179	NaN	NaN	NaN	NaN	NaN	NaN
3	1996_2	1996_2_4s31wr2v	1	0.520933	254	-2.807413	1.122146	0.145536	-1.160902	1.201142	...	0.002177	0.002177	0.002179	0.002179	NaN	NaN	NaN	NaN	NaN	NaN
4	1996_2	1996_2_d70wuvm	1	-0.75041	772	-1.009417	0.748711	0.050893	-0.017163	0.202561	...	0.002138	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

1. Временной период – сначала год, потом номер полугодия. До 2002 года предоставлены все данные. Начиная с первого полугодия 2002, от периода «откусывалось» 40% данных и их надо было предсказывать. По правилам, при создании модели для предсказания, например, периода 2002_1, нельзя было использовать периоды и данные позднее 2001_2, но можно было **валидироваться** на существующих 60% датасета 2002_1

2. Названия акций были анонимизированы (что логично). Но при этом, настолько, что даже если из кода «отсечь» период, код для каждой строчки уникален (нельзя проследить динамику по любой из акций)

3. Признак, относится ли строка к train или test

4. Целевой показатель (прогнозная доходность акции в следующем полугодии) был дан как в виде значения (нормализованного), так и в виде ранга, который надо предсказать. Топ-6 решений по факту решало задачу регрессии и дальше ранжировала эти значения

5. Признаки. Всего 70 анонизированных и нормализованных признака (далее я их называю базовыми). Даны значения за каждый месяц периода (префиксы _1 - _6). То есть 70*6 = 420 значений на каждую строку.

2. Обзор соревнования и правил. Метрики

Основными метриками для расчета качества решения были две:

1. Корреляция Спирмена (далее КС)
2. Normalized Discounted Cumulative Gain of Top 20% (топ 20% предсказанных моделью). Далее NDCG

Конечная формула расчета призовых мест была следующей:

Итоговый балл = (
место по КС (первое место 10 баллов, 10-е один балл) для предсказаний до 2017
+ место по NDCG для предсказаний до 2017
+ место по КС для периода 2017_1*
+ место по NDCG для периода 2017_1*) /4

*Нюанс: для периодов 2002_1 -2016_2 была выборка для валидации и показывался результат на лидерборде.
Для периода 2017_1 выборки для валидации не было вообще (что нормально), но лидерборд обещали открыть только после окончания соревнования (но как оказалось потом, был еще один нюанс ☺) …

3. Общий обзор решения (1/2)

1. Анализ ситуации

- Признаки предобработаны и анонимизированы - **ок, знание предметной области в части финансов не применить**
- Целевые значения выглядят нормально распределенными, **не надо возиться с их нормализацией и вроде неплохо для применения линейных моделей**
- Все признаки выглядят как continuous –светлую идею применить categorical features embedding и разные mean encoding тоже отбросить. **Не очень хорошая ситуация для tree models?**
- Относительно небольшой набор данных (полный train+test около 39К, при этом для предсказания ранних периодов можно использовать менее 10К). **Не очень хорошо для NN?**
- Есть выбросы, пропущенные значения, несмотря на нормализацию разные scales. **Если использовать линейные модели надо думать как делать preprocessing**
- Для каждого периода с 2002_1 у нас есть только 60% данных, оставленных случайным образом. **Сложности для моделей, которые будут предсказывать поздние периоды (например, 2014_1) не имея 40% истории для каждых из прошлых периодов**

2. Мой подход

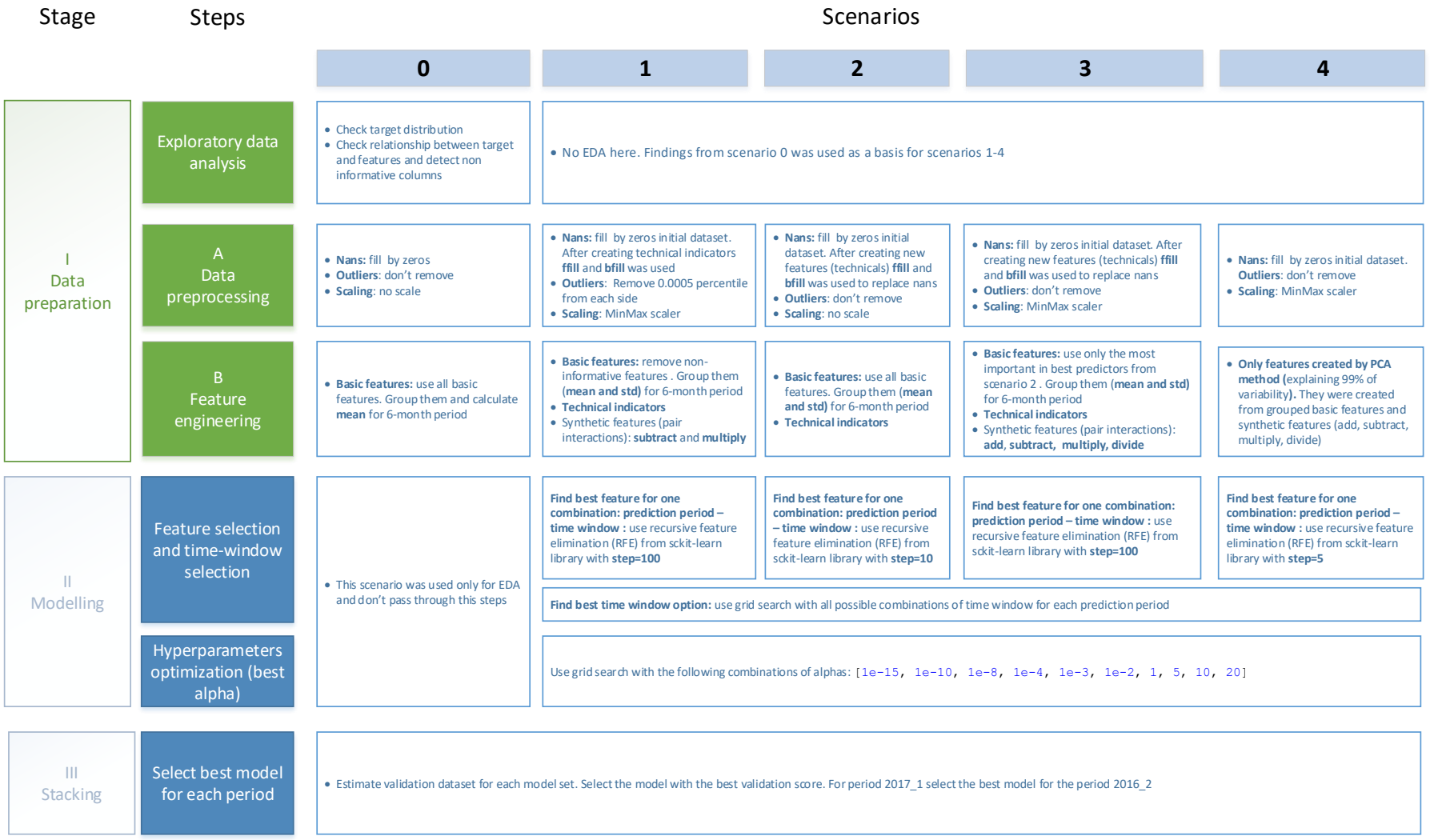
- Если не применить экспертные знания, используем ~~метод Карлссона~~ строгий пайплайн из ML (подготовка данных, оптимизация, валидация, коррекция)
- Я пробовал разные модели, но в итоге линейная регрессия (Ridge) оказалась лучшей и по скорости и по качеству. Как оказалось, в итоге все, кроме 6-го места, использовали и ее тоже
- Раз количество данных мало, попробую создать много разных фич

3. Финальное решение

- Четыре сценария
- По сути, различие только в первом шаге – набор фич
- Модель – Ridge regression из scikit learn
- Параметры оптимизации:
  - Отбор только эффективных фич с помощью RFE из sckit-learn
  - Подбор лучшего временного окна (глубина на которую я смотрю в прошлое)
  - Коэффициент регуляризации в Ridge (alpha)

3. Общий обзор решения (2/2)

Картинка на которой можно залипнуть. Тут показаны все нюансы как и что я делал в каждом из сценарией. Есть на GitHub, сейчас не буду на ней концентрироваться, расскажу о ключевых идеях



4. Ключевые идеи. Подготовка признаков

A. Группировка базовых признаков

- Подумал, что иметь значение базовых признаков за каждый месяц не нужно и надо их просто правильно сагрегировать
- Посчитал среднюю, СКО (std) и медиану (последнюю в итоговом решении исключил, т.к не приносила пользы)
- В итоге агрегации «потерял» как минимум один важный признак, который сыграл у лучших решений - значения признака в последнем месяце (возможно помимо интервальных показателей были показатели на дату и иметь свежее значение иногда было полезнее, чем среднее за период)

C. Синтетические признаки

- Общую идею подсмотрел на хабре из лучших решений ML Bootcamp от mail.ru
- Взял за базу для генерации среднее каждого базового признака
- Далее, для каждой пары признак 1, признак 2 генерится до 4-х новых признаков:
  - Новый признак 1 = Признак 1 - Признак 2
  - Новый признак 2 = Признак 1 + Признак 2
  - Новый признак 3 = Признак 1 / (Признак 2 +0.01)
  - Новый признак 4 = Признак 1 * Признак 2
- Если генерить все комбинации, то получится более 3К признаков, это было много для шага отбора признаков. Поэтому в разных сценариях я старался по умному отбирать базу и использовал не все варианты для генерации

B. Технические индикаторы

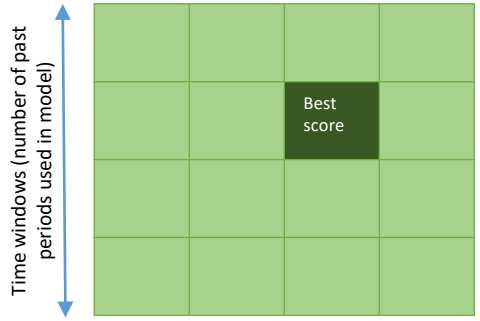
- **Основная идея –** сделать из целевого индикатора признак и на его основе посчитать разные технические индикаторы фин.рынка. Например, в периоде 2002_1 у нас есть прогноз средней доходности на период 2002_2. Но к концу периода 2002_2 это уже признак – историческая доходность за период.
- Так как для каждой строки с признаками акции у нас уникальный код, можно для целого периода рассчитать среднюю доходность, сделать так для всех периодов и от этого показателя считать другие (скользящее среднее, экспоненциальное среднее, моментум, ROC итд)
- В концу просто добавить данные показатели для каждой строки периода
- К сожалению в итоге эта идея вообще не сработала

D. PCA

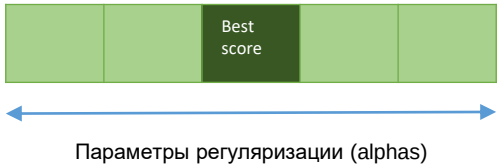
- На первом шаге создал все возможные комбинации синтетических признаков
- Затем с помощью PCA создал новые признаки, которые объясняли 99% вариации (их получилось более 300)

4. Ключевые идеи. Моделирование и оптимизации

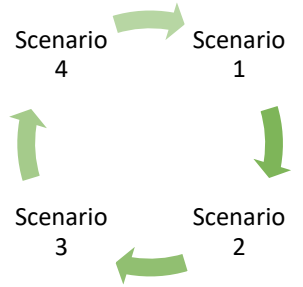
1. Сформировать все комбинации временного окна (TW) для периода на котором делать предсказания (PP), для каждого TW запустить RFE с валидацией на доступных данных PP по показателю корреляции спирмена (SC). Повторить для каждого TW. В конце шага выбрать модель с наивысшим SC



2. Лучшую модель с шага 1 потренировать с различными показателями alpha. Взять лучшую модель по скору валидации



3. Повторить для каждого из сценариев



4. На этом шаге для каждого периода, на который нужно делать предсказания есть по лучшей модели каждого сценария. Тут я просто взял модель, которая давала лучший результат на валидации по корреляции спирмена и NDCG (среднее). На период 2017_1 данных для валидации и паблика не было вообще, поэтому просто взял лучшую модель самого свежего доступного периода (2016_2. В результате она оказалось весьма неплохой

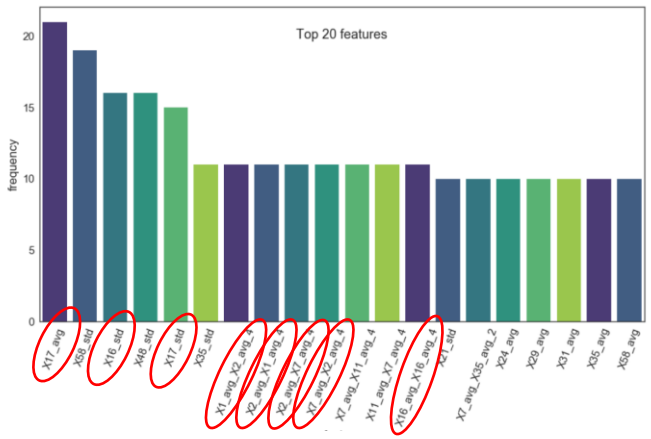
Period	Scenario 1	Scenario 2	Scenario 3	Scenario 4
2001_2				
2002_1				
2002_2				
……				
2017_!				

Period	Best of breed model
2001_2	
2002_1	
2002_2	
……	
2017_!	

5. Анализ лучших моделей текущего решения

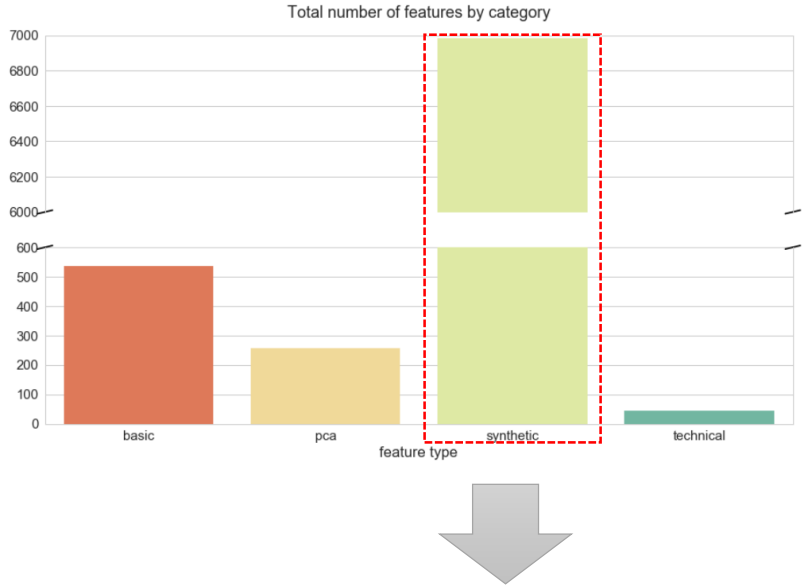
1. В анонимных индикаторах были «киллер-фичи»

- X17, X16 как отдельные и X2 в комбинации с X1 и X7

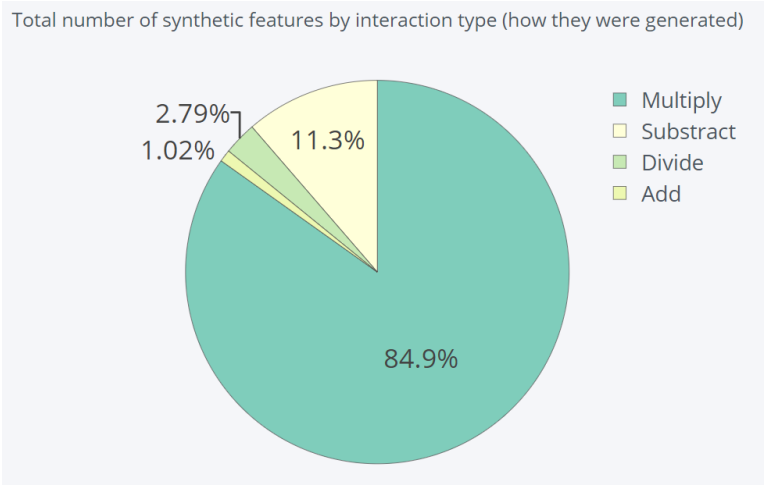


2. Если посмотреть на тип признаков по способу их создания, то можно увидеть что:

- Очень хорошо сыграли агрегированные базовые признаки, надо было их «дожимать» и возможно не агрегировать
- Синтетические признаки тоже сыграли достойно, в подобных соревнованиях надо пробовать их применять
- PCA-признаки немного, но улучшими результат
- А вот квази-технические индикаторы для всего периода не сработали совсем

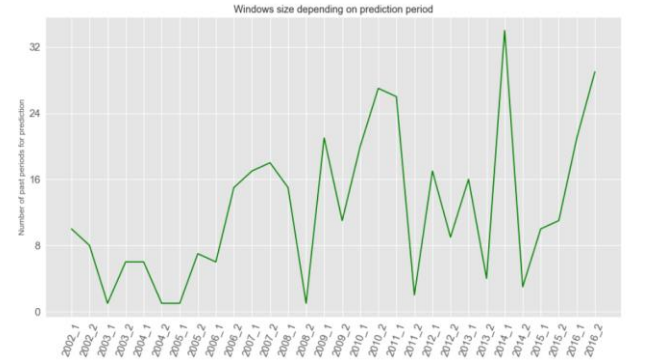


3. При анализе способа создания признака оказалось, что большинство хороших признаков получилось через перемножение агрегированных базовых признаков



4. При анализе размера временного окна у лучших моделей, какой то закономерности не видно

Возможно причина как раз в том, что часть данных была недоступна при временном окне



6. Интересные моменты в решении 1-го и 2-го места

- Оба решения более плотно работали со всеми значениями базовых признаков (участник, занявший второе место показал полезность значения базового признака за последний месяц периода)
- Участник, занявший второе места разбивал выборку для валидации две части: валидацию и тест. По валидации решения отбирались, по тесту проверялось, что этот выбор не случаен.
- Победитель использовал bayesian ridge regression
- Победитель не акцентировался на подборе уникальных признаков (даже упомянул, что это особо и не требовалось), но вместо подбора размера временного окна, **рассматривал каждый временной период, как признак и оптимизировал комбинацию прошлых периодов**, **которые надо включить в предсказательную модель.** Алгоритм подбора какие периоды включать, а какие нет в целом был похож на RFE (последовательное исключение периодов и если при исключении периода результат на валидации растет, убирать, если падает, оставлять).
- Для периода 2017_1, все участники, кроме первого места, подбирали модель наугад. Как итог, участник занявший второе место отправил на этот период нейронки (которые давали хороший результат по прошлым периодам), но получил отрицательный NDCG и корреляцию около 0
- Победитель (единственный) обнаружил, что в конце соревнования (за 10 дней до конца) в кабинете участника появилась неаносированная закладка, на которой был скор сабмита. Оказалось это был скор как раз за модель 2017_1. В итоге за 10 дней победитель отправил на сабмит 50 разных моделей, сформированных вручную, и выбрал лучшую ☺

7. Общие выводы по итогам участия

- Как итог соревнования, победители для каждого из периодов построили по модели, которая по сути заточена на данные валидации будущих периодов
- Однако польза от этого может быть и для организаторов и участников: модели показали лучшие признаки, многие из которых повторялись из периода в период.
- Для участников соревнований решения победителей (особенно второе-пятое места) могут быть полезны в части построения пайплайна и идей, которые применимы и на других соревнованиях
- Какие моменты с моей точки зрения можно было улучшить:
  - Запретить использование отдельной модели на каждый период. Это должно было бы подстегнуть мотивацию создавать более универсальную модель, делать ансамбли и стекинг
  - Второй индикатор тяжело оптимизировать и для линейных моделей он часто сильно коррелировал со спирменом. Можно было оставить одну метрику
  - Не делать уникальный кодинг акций для каждого периода, а сделать как, например, на Kaggle two sigma где, несмотря на то, что индивидуальные акции были закодированы, но код повторялся отслеживался между периодами. Это повысило бы качество моделей, так как позволило учитывать динамику акций и применять трюки как для категориальных признаков
  - Сделать хотя бы минимальный призовой фонд ☺ Это должно было бы подстегнуть конкуренцию

- Полное решение можно найти на GitHub https://github.com/kvr777/ieee-challenge
- В случае вопросом можно писать мне в слаке на ODS или на почту: kirill.v.romanov@gmail.com