Get a PDF

Подведение Итогов

И ещё раз громадное всем спасибо!

За полтора месяца работы "вслепую" чётко по расписанию, без ясно обозначенных целей. За доверие и поддержку.

С момента объявления эксперимента прошло четыре месяца. За это время было написано втрое больше кода, и проведено в 30 раз больше расчётов, чем планировалось. В основном для того, чтобы понять, что же я увидел. Впрочем, когда задаёшь природе правильно поставленные вопросы, что-то подобное обычно в ответ и получаешь.

Увы, не всё из того, что задумывалось, удалось получить. Но я многое понял, и я очень рад, что всё-таки попробовал.

Целей было три:

1. Проверить, можно ли с помощью ML предсказывать такое будущее, о котором на момент опроса никто, включая самих участников, не имеет достоверной информации. Это удалось, хотя и с меньшей точностью, чем хотелось. Даже в такой форме это открывает возможность объединяться в группы ради усиления совместного интеллекта -- а не как в современных соцсетях.

2. Научиться отделять "истинные" компоненты моделей от "крокодильих". Представления последних могут идти вразрез с объективной истиной, но всё же вносить полезный предсказательный сигнал. Это удалось полностью, как на тестовых, так и на реальных данных.

3. Понять, можно ли обнаружить общее знание между двумя разнородными моделями, и использовать эту общность для косвенного установления истинности моделей, не поддающихся непосредственной проверке. Это удалось лишь частично. Причём я сейчас вижу, что другая постановка эксперимента могла бы гораздо лучше ответить на этот вопрос. Подробнее в итогах.

Содержание:

0. Об участниках

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Всего в эксперименте участвовало 16 человек из Канады, России, США и Украины (что создало нетривиальную задачу по выбору момента рассылки и отслеживанию разных часовых поясов). Из этих шестнадцати -- 12 мужчин, 4 женщины (хотя приглашения рассылались в соотношении ближе к 40/60).

Вот, пожалуй, и всё, что можно анонимно сообщить. Если вы забыли свой номер участника, пишите.

1. Подробнее о целях и замысле

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Допустим, у нас есть модель для предсказания будущего. Может быть, она предсказывает цену золота. Может, температуру лампочки или климата. Может, реакцию публики на рекламу. Не так уж важно, какая именно это

модель -- машинная ли, из физических формул, или же в виде глубоко интуитивного знания хирурга. Но важно, что модель эта правильна (я также иногда буду говорить "истинна"). Это означает, что модель верно предсказывает будущее. Хотя бы на узком наборе случаев. Пусть даже с ошибкой. Но если ошибка эта случайна и не слишком велика, то правильность модели, хотя бы в статистическом смысле, неоспорима. Ведь она явно что-то "знает" о мире, в чём любой скептик сможет убеждаться хоть до посинения на новых и новых прогнозах.

Гораздо интереснее другой случай. Вот есть красивая и внутренне непротиворечивая модель... которая не даёт легко проверяемых прогнозов. Потому что все её прогнозы или смотрят на 300 лет вперёд, или безумно дороги в проверке. Или "предсказывают" что-нибудь глубоко историческое. Половцы ли первыми напали на печенегов в в 1111-м году, или наоборот? А?

Как проверить такую модель на истинность?

На первый взгляд, никак. На второй, вообще проверять не надо. Ибо, если модель не делает **никаких** верифицируемых предсказаний, то и разницы нет, следовать ли ей в физическом мире. Что следуй её предсказаниям, что действуй поперёк них, вероятность стукнуться лбом о неожиданность одинакова. Ибо будь она разной, по разнице можно было бы что-то понять про истинность.

Так, да не совсем. Ибо модели не живут "в вакууме". Способы, которыми одна модель делает предсказания (пусть даже труднопроверяемые), могут частично "пересекаться" со способами, используемыми в других моделях. Которые, быть может, проще проверить. В качестве примера представим, что автор модели про половцев, оказывается, не умеет читать и пересказывать тексты. То есть, он проваливает другую, весьма простую модель "предскажи понимание этих буковок другими людьми". А умение читать и понимать текст -- ключевой элемент в построении исторической картины. Если этот элемент "поломан", то маловероятно (хотя и допустимо), что картина, построенная с его участием, окажется верной. И, соответственно, правильность модели про половцев в этом случае оказывается тогда под некоторым вопросом.

Грубо говоря, и упрощая, из вариантов написания "жи/жы" один таки является **объективно** более правильным. Не потому, что так сказано в учебнике. А потому, что механизмы, определяющие выбор, влияют, пусть и косвенно, на множество других измеримых вещей. На способность писать **все** тексты, составлять расписания, доходчиво объяснять начальству свои идеи, правильно выбирать свои войны, и т.д. И некоторые из этих способностей могут быть объективно измерены.

Это и есть наша Основная Гипотеза. Что правильность модели можно проверить не только по её предсказаниям, но по "пересечениям" с другими моделями.

Разумеется, **одна** такая проверка не сможет подтвердить (или опровергнуть) непонятную модель. Для этого надо "просто" протестировать модель на перекрытие "вычислительных путей" с как можно б**о**льшим количеством других моделей, истинность которых известна. При достаточно большом количестве таких перекрытий можно ожидать, что истинность "непонятной" модели если не установится, то хотя бы станет статистически более весомой.

А это уже немало. Ведь это открывает дверь к проверке на правильность множества вещей, традиционно считаемых "чёрными ящиками". Взглядов. Убеждений. Мировоззрений. И, конечно же, новостей, если они подаются в комбинации с определёнными выводами или политической картиной мира. Ведь выводы являются моделью, предсказывающими хоть что-то, хотя бы в прошлом. А значит, их можно подвергнуть такой проверке и выяснить, помогает ли принятие этих выводов или картины мира угадыванию будущих событий.

Разумеется, на практике подобное тестирование потребует громадных вычислительных ресурсов и вовлечения множества -- десятков, сотен, тысяч -- проверенных моделей. Мы не сможем здесь это проделать.

Но мы сможем попытаться проверить работоспособность единичной проверки. Мы можем попытаться обнаружить и измерить "перекрытие" двух достаточно разнородных моделей. В некотором смысле, мы будем тестировать... винтик. С помощью которого люди, скрепляя детали произвольной и неизвестной нам пока формы, когда-нибудь построят машины, самолёты, города... Но это потом. А пока -- пока надо протестировать просто винтик.

Итак, закатываем рукава, начинаем работу.

Но тут на сцене появляются...

2. Крокодилы!

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

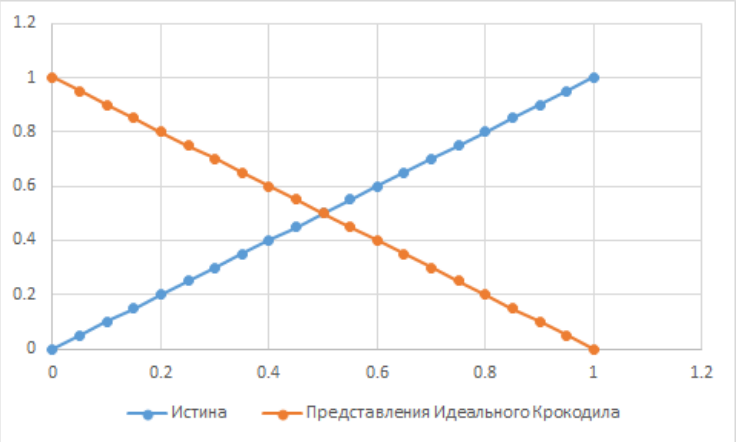
Представим себе, что мы тренируем модель, которая из индивидуальных предсказаний температуры на завтра, сделанных многими людьми, формирует объединённый прогноз.

И допустим, среди участников нашёлся человек, чьи предсказания 100% ошибочны. Когда он заявляет, что будет тепло -- жди завтра холодов. Когда предсказывает морозы -- будет жара.

Очевидно, что этот человек живёт в очень странном информационном пузыре. Он явно разбирается -- в определённом смысле -- в погоде, у него в голове есть сложная её модель. Каждый день эта модель сталкивается с реальностью, каждый день реальность говорит ему "ты неправ", но человек продолжает за свою модель по каким-то, очевидно, очень сильным причинам, держаться. Ну или просто нас троллит :)

Таких людей я назвал крокодилами. Если в простейшем случае **y** = **x** истина выглядит так:

То представления идеального "крокодила" будут ей строго обратны, по крайней мере, на некотором интервале:



С **нашей** точки зрения представления "крокодила" полностью ложны.

Но с точки зрения машинного обучения они, однако, столь же информативны, сколь и истина! Бери их, переворачивай знак, и вот тебе готовый правильный прогноз.

Возникает серьёзное проблемко. "Истинное" представление для ML неотличимо от "ложного", покуда оба достаточно информативны. А мы ведь собираемся измерять "истинность", а не информативность. Как быть?

В общем виде задача, видимо, сводится к измерению количества трансформаций, требующихся, чтобы превратить представления того или иного участника в "истинные". И выдаче "степени истинности" на основании этого количества. Крайне скользкий путь.

К счастью, для задачи бинарного предсказания на дискретных входных параметрах малой размерности есть альтернатива: монотонные регрессоры. Пытаясь описать сложную функцию $\mathbf{y} = F(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})$, они требуют, чтобы зависимость **y** от каждой переменной была неубывающей. То есть, если любой из $\mathbf{x_i}$ увеличился, то **y**, по крайней мере, не уменьшился. Соответственно, вклад "крокодилов" на участках, где их представления меняются вразрез с истиной, таким регрессором будет игнорироваться. А значит, модель будет строиться из кусочков представлений либо истинных, либо, по крайней мере, с истинными скоррелированных.

Применяя монотонные регрессоры к таким задачам (разумеется, заранее оформленным, как монотонные) мы можем оценивать именно истинность вкладов и моделей. И более того, по разнице в точности между монотонным и аналогичным ему немонотонным регрессором можно количественно оценить вклад "крокодильей" части мировоззрения участников в результат.

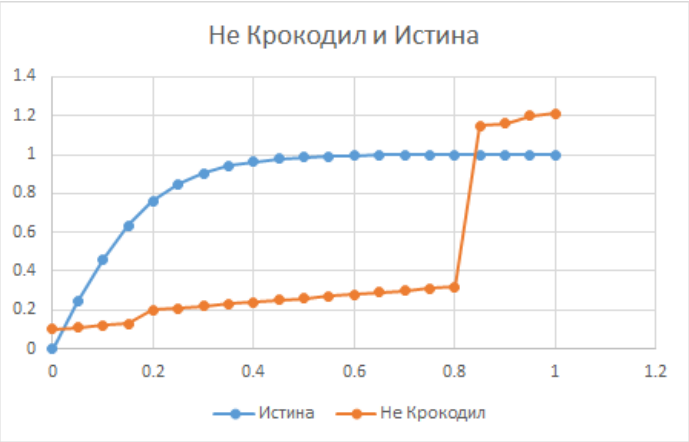
К сожалению, я не нашёл удобной имплементации монотонной версии своего любимого регрессора RandomForest (хотя технически она должна быть не так уж и сложна). Но существует родственный ему HistGradientBoostingRegressor, который можно сконструировать в монотонной версии, и которым мы и воспользуемся.

Вообще, даже сама возможность существования "крокодилов" приводит к крайне любопытным выводам. Кому любопытно, читайте следующий раздел. Ну а если нет, прыгайте дальше.

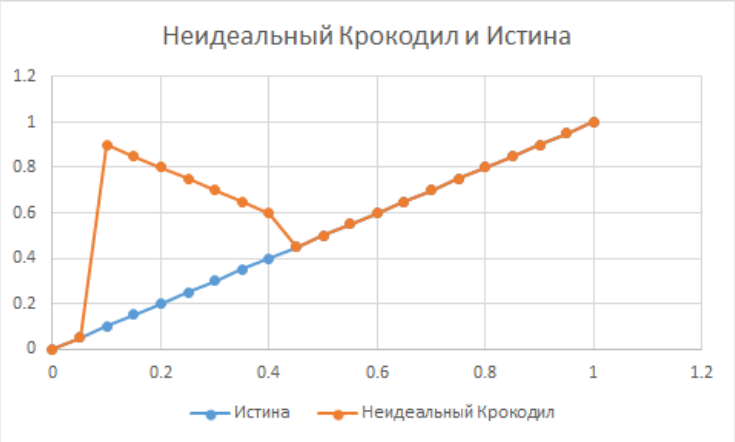
21. Баллада о крокодилах

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Важно понимать, что крокодил -- это не тот, кто просто ошибается. А тот, чьё представление о вопросе "перевёрнуто" на некотором участке данных. Так, что когда истинная ситуация меняется в одну сторону, предсказания крокодила систематически меняются в противоположную. Поэтому вот эта оранжевая кривая -- это ещё **не** крокодил:



А вот тут, между 0.1 и 0.45, имеется "крокодилий" участок:



Крокодилы могут упорно верить в невообразимую, дикую, постоянно приводящую к острым конфликтам чушь, но убеждений своих не меняют. Скорее всего потому, что локально, в некоторой **малой** окрестности их бытия, модель эта даёт верные (и даже лучшие, чем у всех вокруг) предсказания. Ну вроде как у пикаперов. Локальное убеждение, что женщина -- недосущество, частенько повышает локальный успех с женщинами. Но, очевидно, приводит к катастрофам на более серьёзных масштабах.

В момент, когда я помыслил эту картинку, я завис на пару пару дней.

Во-первых, учитывая, до какой ус%ачки люди спорят по любым поводам, легко видеть, что крокодилы не просто существуют, а многочисленны.

Далее, почти все мы наверняка ~~немного лошади~~ являемся крокодилами хоть в каком-то вопросе. Просто потому, что пространство задач громадно, и в нём наверняка найдутся такие, где мы опираемся, сами того не осознавая, лишь на пяток кривых наблюдений, через которые легко провести "перевёрнутую" модель. Независимо от уровня интеллекта или образованности. Даже хуже, умник может такую многопараметрическую загогулину в свой случайный опыт вписать, до которой "средний" человек просто не додумается и плюнет на вопрос.

С философской точки зрения здесь возникает потрясающий результат. А именно, **ложная модель может работать как истинная, будучи встроена в существенно более сильный разум**. Каковым, в данной задаче, выступает ML. Большой набор крокодилов, воображающих себе всякий бред, вконец запутавшихся в жизни, отчаявшихся и вечно ошибающихся, знающих лишь обрывочные, искажённые кусочки картины мира -- эти крокодилы **могут** владеть **большой истиной**, если мнения их правильно собрать воедино, и каждое мнение может оказаться нужным!

Конечно, лишь Господь Бог, если Он есть, может полностью решить эту задачу. Но нас вполне устроит решение маленькое и частичное, лишь бы оно было хоть чуток лучше, чем прямое голосование. "Крокодилы всех стран, объединяйтесь!" -- вот воистину правильный лозунг. Особенно в нынешние, полные вранья и разбитых идеалов, времена.

А наиболее серьёзная проблема современности -- это как включить "крокодилов" в процессы принятия решений. Все они носят в себе весьма нетривиальные модели, зачастую колоссальной предсказательной силы, по крайней мере, локально. Но глобально находятся в противоречии как с истиной, так и с доминирующими идеологиями.

Почти везде крокодилов полагают психами, придурками, приписывают им всевозможные *измы и в лучшем случае просто игнорируют.

Как быть?

Можно их уничтожить. Распихать по тюрьмам, гулагам, заставить заткнуться. Советский и, похоже, опять российский способ. Чреват тем, что у власти быстро не остаётся ни единого канала, по которому поступала бы правдивая информация. Отчего та начинает верить, что заведомо самоубийственную войну можно выиграть, да ещё за три дня.

Можно их мягко игнорировать. Что и пытается делать современный Запад. Вся эта война с "hate speech", вплоть до невозможности хранить определённые видео в облаке -- это оно. Но, помимо этической проблемы лишения всякого голоса множества живых людей, это неверно практически. Ведь эти люди несут в себе колоссальную энергию и информацию. Вышвыривание их из модели делает её хуже и приводит, в итоге, к коллапсу управления и революциям. Ещё читая Бжезинского году в 2006-м, я записал:

> *Допустим даже, мировая демократизация более-менее удалась и в мире действительно наступила "стабильность". Бжезинский неоднократно употребляет это слово, надо бы остановиться на этом подробнее.*
>
> *Под стабильностью он подразумевает сохранение статус-кво нынешнего принципа мирового устройства: основной способ производства – капиталистический; основная ценность – права человека; в мире по-прежнему существует масса бедных стран.*
>
> *По мне, такая "стабильность" скорее сродни успешному подавлению победителем побеждённого. В данном случае победитель, может, и впрямь не так плох, как многочисленные исторические примеры. Yet I believe, что более правильным термином была бы всё же "метастабильность", что-то вроде равновесия седока на скачущей лошади. Нет никаких гарантий, что подобная стабильность действительно является минимумом потенциальной энергии как функции социального и экономического устройства.*
>
> *<...> Не все люди разделяют культурные ценности США. Допустим даже на минутку, что они все ошибаются. Но ведь им-то от этого не легче! Человек может в душе полагать какую-нибудь нацию грязными свиньями; он может верить, что человеческая жизнь не имеет ценности; он может быть в душе коммунистом или анархистом; он, наконец, может просто считать своим неотъемлемым правом пьяную езду, а ему этого не позволяют! Так или иначе, но у очень многих людей в мире есть те или иные подобные заморочки. Если их не осуществлять, во что выльется подавленное недовольство? Не окончится ли оно массовыми, бессмысленными, полупсихопатическими бунтами, саботажем, бардаком?*
>
> *Уже сегодня консульства США выполняют функцию подавления. Так, если девушка занимается проституцией, в США её не пустят. Если человек не имеет стабильной работы, нормального дома/квартиры и приличного дохода, то есть "не встраивается" в капиталистическоей общество, то, скорее всего, в визе ему опять же откажут. Если у него были проблемы с законом в своей стране, въезд в США тоже может оказаться затруднительным.*
>
> *Подобное разделение на "правильных" и "неправильных" сильно сегодня недооценивается. Между тем оно порождает широко распространяющуюся и глубокую социальную обиду, и может быть крайне опасным для демократических государств через создание негативного и презрительного общественного образа этаких ханж, "слишком правильных чистоплюев". Собственно, в России данное отношение к США уже массово сформировано. Не без активной помощи воплей официальной власти, ибо ей выгоден антиамериканизм :)*
>
> *Нет, всякая глобальная мировая система ценностей должна оставлять какую-то приемлемую роль и для подобных взглядов. Нельзя выкидывать человека, как отходы, на свалку общественной и экономической жизни за присущие ему атавистические вгляды. Точнее, можно, но это кончится "тихим бунтом" отверженных, тем более страшным, что у них не будет ни какой-либо созидательной цели, ни альтеративной программы улучшения жизни, а будет лишь одно желание: разрушение всего, ассоциируемого с подавлением человека в их глазах. То есть, в нынешней обстановке, тех самых демократических и гуманных ценностей.*

Можно попытаться переучить и перевоспитать их. Это "мягкий" советский и китайский способ. К сожалению, даже если перевоспитание срабатывает, на выходе получается убогий "неотрицательный" классификатор-болванчик, синхронный с линией партии.

Раскрыть им глаза? Редко работает. Иначе они не были бы "крокодилами", ибо реальность и так предоставляет им немало намёков "перевёрнутости".

Позволить каждому верить в свою чушь, но собирать предсказания ну вот хоть машинными обучением? Возможно, будет работать... до тех пор, пока крокодил не входит в противоречие с реальностью. Пока он не увидит, что чем сильнее он отстаивает свою позицию, тем более обратно оной принимаются коллективные решения. Которые да, работают ко всеобщей, в том числе и его, выгоде, но его лично это всё равно не убеждает, а лишь только злит и бесит.

Засунуть каждого в свой виртуальный "пузырь"? Это пытается делать Фейсбук. Доводя людей до катастрофических столкновений в реальности физической.

Учить всех терпимости, понижать примативность и самоуверенность? Приличный вариант. Вся христианская цивилизация построена на этом. Но очень, очень легко перегнуть палку. Ведь НЕ ВСЕ мнения одинаково ценны. НЕ ВСЕ одинаково безопасны при столкновении с реальностью. Что возвращает проблему на круги своя.

К сожалению, увы, даже Запад качественно решить эту проблему сегодня не готов. Про Китай и Россию я вообще молчу.

30. Данные и предсказатели

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Итак, нам потребуется несколько моделей. К счастью, их технологию создания мы уже отработали в предыдущем эксперименте.

Предсказатели

1. RandomForestRegressor -- как самый мощный и стабильный на зашумленных небольших табулированных данных, далее фигурирует под именем RF.

2. HistGradientBoostingRegressor -- наш основной предсказатель, под именем XG. Близкий родственник RandomForest, хотя заметно отличается в деталях поведения. Более склонен к оверфиттингу. Если находит в фичах одну сильную, может почти всё предсказание "повесить" на неё, игнорируя более слабые компоненты. Принятый риск.

3. Он же, но в монотонной версии (параметр monotonic_cst выставлен в единички для всех входных фич) -- для использования только не-крокодильего вклада. Зовут его XG+.

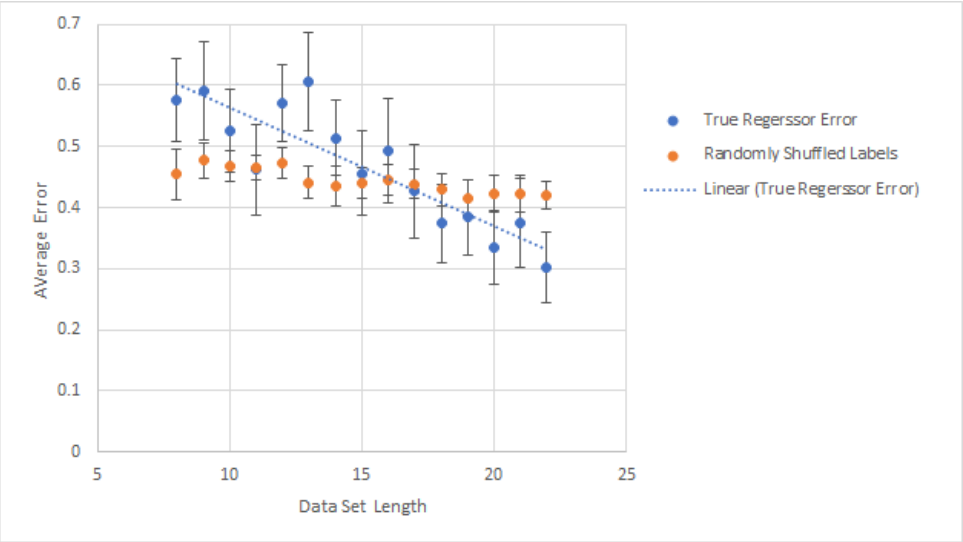
4. ElasticNet (EN) и она же в монотонной версии (EN+). Исключительно для сравнения с нелинейными предсказателями.

Модель №1. Цена активов.

Собираем кучу индивидуальных предсказаний по цене на завтра, заполняем пропущенные данные нулями, низкую уверенность обозначаем ±1, высокую ±2, сравниваем с реально случившейся ценой, тренируем регрессоры её предсказывать:

Date	Q	Label	p0	p1	p2	...
7/5/2022	MSFT	1	0	1	1	...
7/6/2022	MSFT	1	1	1	-1	...
7/7/2022	MSFT	0	1	0	2	...
7/8/2022	MSFT	0	1	-1	1	...
7/9/2022	MSFT	0	1	-1	2	...
7/10/2022	MSFT	1	-1	0	2	...
...	...	...	...	...	...	...

По нефти и по курсу рубля ничего не получилось. А вот предсказатель цены Майкрософтовских акций таки заработал:



Голубое -- средняя ошибка кросс-валидации, полученная на случайной выборке данных, как функция длины этой выборки (test size = 4). Оранжевое -- то же самое, но со случайно перемешанными метками. Видно, что регрессор "учится" по мере увеличения объёма данных. Наклон линии статистически достоверен на уровне 98%. На полном наборе конечная ошибка составляет 0.35 - 0.40 в разных "забегах" во всех трёх "сильных" регрессорах (RF, XG, XG+). Не слишком хорошо, но тащить с людей ответы ещё целый месяц мне совесть не позволяла. Может быть, зря -- об этот недостаток точности я ещё побился головой.

Отмечу, что то, что мы здесь вытворяем, нагло противоречит "классическим" рекомендациям по машинному обучению. Они говорят: не пытайтесь сделать систему лучше, чем человек (т.н. HLP – Human Level Performance). Они говорят: если ваши эксперты расходятся в расстановке меток, в первую, вторую, в третью очередь займитесь устранением этого расхождения **до** обучения. Пусть даже ценой потери части информативности. Потому что "переламывать" такое расхождение методами ML -- безумно дорого. Это правда. Дорого. Но не невозможно. Это можно понять как из статистических соображений, так и экспериментально (что мы и показали в предыдущем эксперименте).

Модель использовалась в двух вариациях, "малой" и "полной".

Дело в том, что далеко не все участники ответили на все вопросы. Первые же запуски показали неустойчивость результата относительно ответов с большим количеством пропусков. Поэтому в "малой" версии модели их, к сожалению, пришлось исключить. При этом возник вопрос: а кого же оставить? Я перепробовал несколько критериев и в итоге сошёлся на том, что надо просто оставлять людей, ответивших на 80-90% вопросов. Это оставляет в "малой" модели (чей график точности приведён выше) пять человек.

И здесь мы, на самом деле, достигли цели №1 эксперимента: показать, что относительно небольшая группа не-экспертов, используя относительно скромное оборудование, может формировать совместные предсказания такого будущего, о котором на момент предсказания не знает **никто** из них.

Чем это важно, расписано в следующем разделе. Ну а если кому не интересно, продолжаем про эксперимент.

31. Лирическое отступление о бюрократах

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Роберт Хайнлайн частенько задавался вопросом: какое действие индивидуума **не** является этичным, будучи выполненным в одиночку, но становится таковым, будучи выполненным группой? На первый взгляд, разницы нет, и Хайнлайн частенько использует это для продвижения либертарианских идей.

Вопрос красивый, но чёткий ответ на него есть.

Простейший пример -- доктор, на руках у которого пациент с нетривиальной и плохо изученной болезнью. Доктор, без сомнения, может сам определить курс лечения. Но это действие будет неэтичным, ибо несёт в себе высокий шанс ошибки, легко понижаемый, если собрать консилиум специалистов по этой болезни. Вероятность того, что пятеро экспертов, осмысленно обсудив проблему, ошибутся, всё же значительно ниже. Коллективное решение в данном случае более этично, чем индивидуальное, благодаря принятию разумных мер по предотвращению бесчеловечных последствий.

Просматривается утверждение: более правильно (и потому обычно более этично) принимать решение коллективно, если это ведёт к меньшей вероятности ошибки. По крайней мере, пока участники коллектива не идиоты и благожелательны, а задержка на обсуждение некритична.

(Хайнлайн, экзистенциалист, сказал бы, что пофигу, ибо следовать ли коллективному решению, решает всё равно доктор. Это так, но это не снимает ответственности с консилиума за явные глупости -- то есть, за решение хотя бы в некоторой степени отвечают все.)

Итак, вовлекая в процесс других, мы иногда можем сделать его менее ошибочным и, вероятно, из-за этого более этичным. Поэтому имеет иногда смысл требовать, чтобы проект нового моста, софта, производства был внимательно просмотрен экспертами по сопромату, безопасности, химии, экологии, и всем вот этим на первый взгляд нудным и гнусным "кодам", диктующим, куда обязана в доме открываться дверь и сколько лампочек должно быть на комнату. Потому что пусть не все, но многие из этих правил писаны кровью. Отсюда вырастает (в общем случае) необходимость госрегулирований, общественного контроля, предписаний и бюрократии.

Но, как часто бывает, дотрагиваясь до этой идеи, бюрократия превращает её в говно. Если по задумке эксперты и проверки должны делать процесс лучше и правильнее, на практике они сплошь и рядом его тормозят насмерть. Судебное решение, требующее пары **часов** человеческого внимания, занимает пару месяцев. Покупка рентгеноструктурного анализатора требует **бизнес-**лицензии. Не потому, что, заведя ломбард, его владелец внезапно начинает лучше обращаться с радиацией, а потому, что только в такой постановке бюрократия имеет силы безопасность проконтролировать. Несчастный security/privacy review в софтверной компании может занять недели и съесть метры нервов. А пропихивание нового закона (не говоря уж об отмене старого) в правительстве -- вообще многолетняя клоунада. И это, предполагая, что бюрократы всё-таки разбираются в вопросе и хотя бы стараются решить его правильно. А не превратились в бессмысленных простановщиков птичек напротив требований, или в безумный принтер постановлений.

Почему? Потому что **бюрократия неэффективна**. Несмотря на колоссальный прогресс в технологиях, принятие **коллективного** решения людьми -- по-прежнему пещерный процесс. Он безумно медленен, он нестабилен, он легко подавляется единичным крикуном и легко подпадает под власть рвача, и его почти невозможно заранее проверить на наличие ошибок. Оттого даже самые простые инструменты типа регламента или ведения протокола уже дают своим пользователям колоссальные преимущества.

Тем не менее, именно бюрократии правят современным миром. Демократия, коммунизм -- все эти идеи давно ушли в песок. Бюрократы, начертав на знамёнах "всеобщее благо", пытаются к нему рулить, медленно но верно уменьшая количества степеней свободы цивилизации. Зачастую, быть может, и не желая этого.

Нам нужен, и нужен срочно, **более эффективный** способ принятия коллективных решений. В сто раз быстрее, чем современная бюрократия. В десять раз реже ошибающийся. Способный работать одинаково как для двух человек, так и для двадцати миллиардов. Со встроенной на уровне процесса проверкой на противоречие "твёрдо" проверенным фактам. И, разумеется, распределённый, чтобы ни захватить, ни запретить, ни отменить его было невозможно.

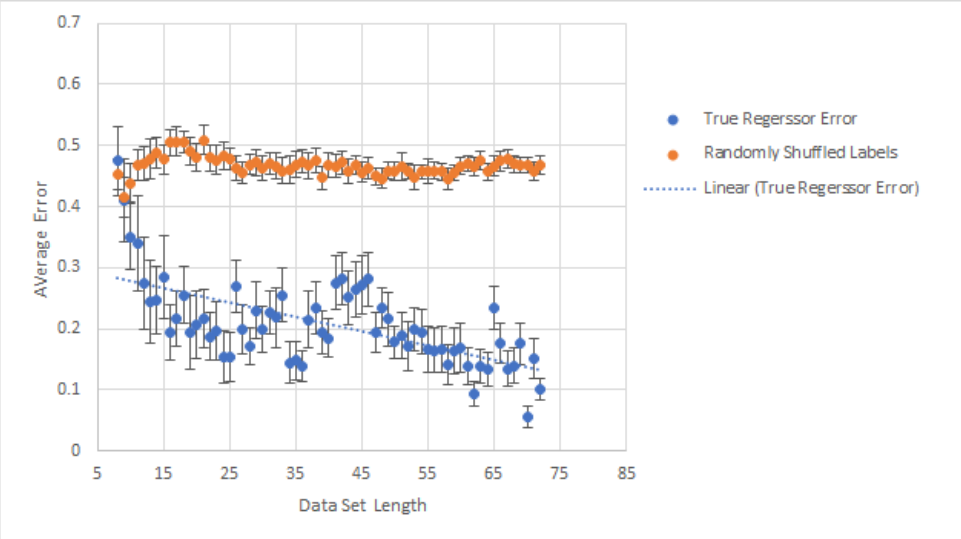
Быть может, распределённое машинное обучение способно поддержать эту задачу. Или хотя бы эффективно приступить к ней.

32. Данные и предсказатели, продолжение

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Модель №2. Задачки по физике и математике.

Ну, тут работоспособность модели куда очевиднее:



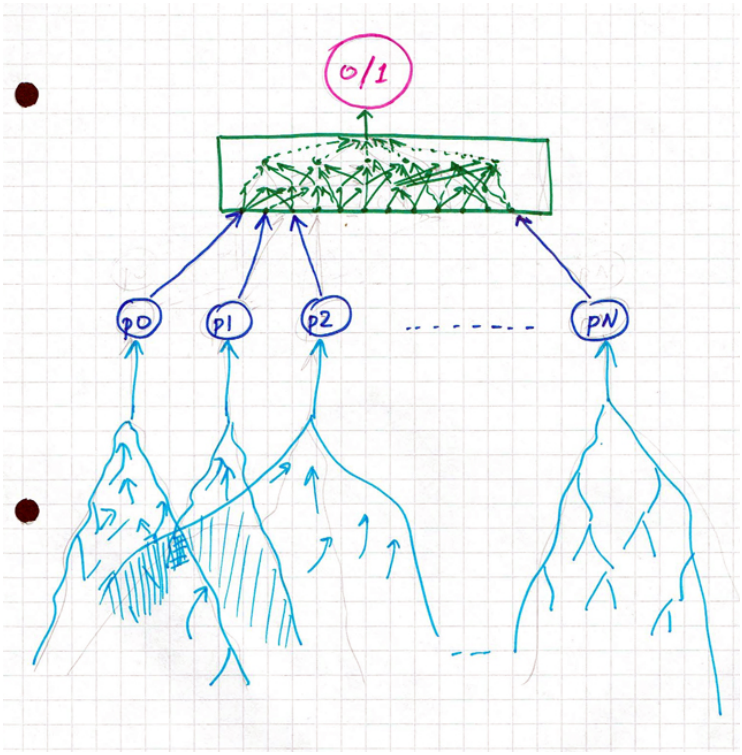
Вы спросите, откуда на графике 72 точки, если задачек было всего три десятка? Дело в том, что каждая задача, имея несколько вариантов ответов, генерила столько же и строк для бинарной классификации данных. Минус дубликаты, разумеется.

Отметим, что все эти задачки про ожидаемое число ударов по морде, плотность олова и свинца, количество оборотов рулона скотча -- это тоже про предсказание будущего. Конечно, в данном случае будущего известного многим людям. И, возможно, с небезупречными оценками истинности, ибо расставлял их я, а не суровая реальность. Но даже с такой поправкой можно ожидать, что **в среднем** эта модель всё-таки верна и что-то предсказывает. Чем нам и пригодится.

40. А где у ней неонка? (Замечание).

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

Небесполезно отметить, что "разум", делающий предсказание, "живёт" не только в моём компьютере, а распределён между участниками довольно нетривиальным образом:



Верхний, зелёный уровень модели, действительно образует машинный регрессор, обученный на предыдущих ответах, и крутящийся в моём компьютере. Но, чтобы выдать завтрашнее предсказание, ему нужны прогнозы от индивидуальных участников p_0, p_1, ... p_N (обозначены синим). А они свои предсказания берут из всей совокупности увиденного, услышанного, передуманного и ранее пережитого (голубое). Возможно, где-то эти совокупности друг с другом пересекаются. Для получения простого ответа 0/1 надо, чтобы вычисления прошли по всей системе, с самого дна и до самой до макушки. И изменения в любом месте могут вызвать изменение результата.

50. Главная часть.

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

А вот здесь начинается шаманство, ради которого всё и затевалось.

Итак, как проверить, что две модели (по ценам и по физике) "перекрываются" в вычислительном смысле, и измерить хотя бы знак этого перекрытия?

Тут возможно множество подходов. Я приведу два, на мой взгляд, наиболее осмысленных, а остальные свалю в отдельный документ.

*** Постановка первая**, или **Метод Разделения.** Он отвечает на вопрос: перекрываются ли модели на большинстве участников **в среднем**?

Идея такая. Давайте возьмём и разделим всех участников на две равные группы A и B *некоторым образом*.

Затем посчитаем качество предсказания цен, полученное обучением на каждой группе. Пусть R(A) и R(B) обозначает остаточную ошибку такого предсказания. Затем точно так же посчитаем качество предсказания ответов по физике каждой группой. Пусть T(A) и T(B) будут остаточными ошибками в этом вопросе.

Если существует некоторое общее знание, одновременно помогающее решать задачи каждой группы, то статистически можно ожидать, что когда R(A) > R(B), то и T(A) > T(B). Потому что, когда больше этого общего

знания попадает в группу B, то она должна **в среднем** лучше решать задачи **каждой** категории.

С другой стороны, если это знание помогает решать задачи одной из групп, но мешает другой, то, наоборот, при R(A) > R(B) будет в среднем ожидаться T(A) < T(B).

Если же никакого общего знания нет вообще, то мы увидим, с точностью до погрешности, нулевую корреляцию между R(A) - R(B) и T(A) - T(B).

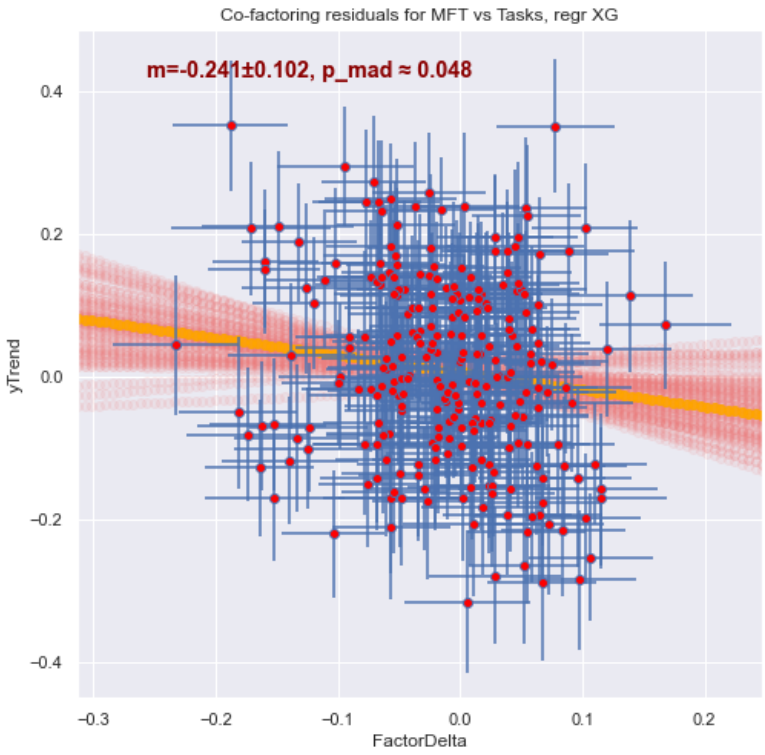
Самый интересный случай такой: а что, если это знание существует, но оно **мешает** обоим группам? В этом случае зависимость между (R(A) - R(B)) и (T(A) - T(B)) окажется положительной... но только на немонотонных регрессорах. Монотонный регрессор эту общую ошибку исключит, и мы увидим, в среднем, ноль.

Вроде бы всё понятно. Остался вопрос: как будем разбивать?

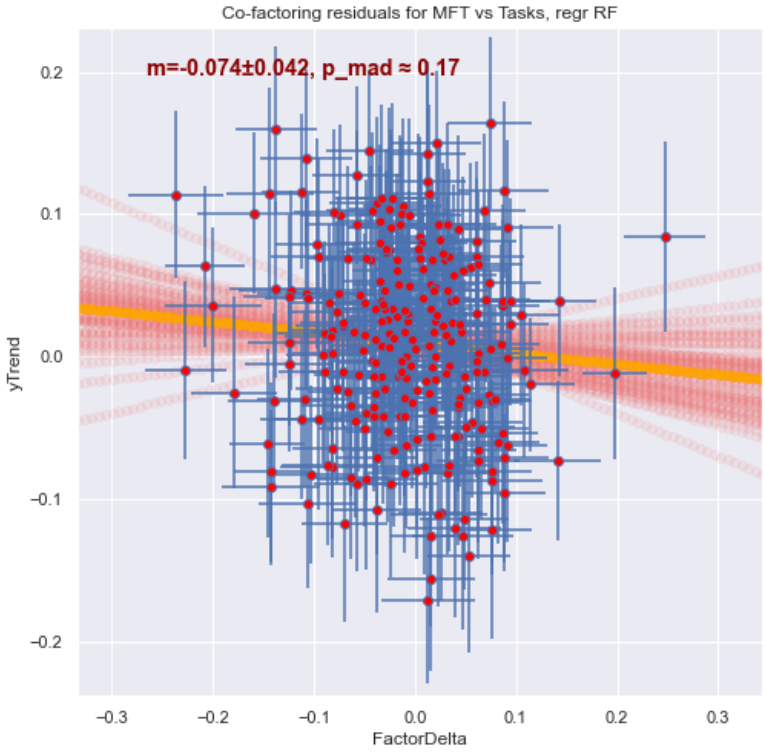
Я потратил кучу своего и процессорного времени, чтобы дойти до простого ответа: только случайно. Разумеется, на группы одинакового размера. Всё. Все прочие варианты так или иначе привносят в задачу всякие предположения, эхо которых мы и можем услышать вместо ответа. Поэтому -- только много-много случайных разбиений, с последующим выделением сигнала статистикой.

Вот что получается, если это проделать.

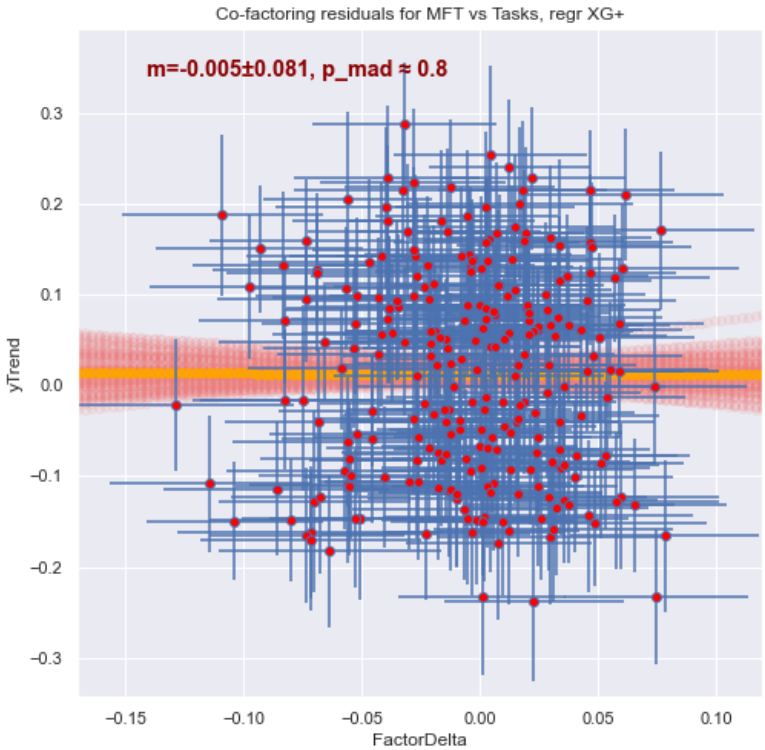
Регрессор XG довольно уверенно видит отрицательную зависимость (p_mad -- примерная оценка вероятности получить такой же, или более отличный от нуля наклон в предположении, что точки разбросаны случайно и никакой зависимости между ними на самом деле нет):



RandomForest менее чётко видит то же самое:



А вот "безкрокодильный" регрессор не видит ничего:



Что это означает? Что между задачами про цены и про физмат **есть** общее знание. Оно помогает решать один из классов задач и мешает решать другой. Но оно, к сожалению, имеет "крокодилью" природу. То есть, это не "знание", а некоторая ошибка. Которую ML может использовать, но которая человеку, опирающемуся на это "знание", лишь мешает.

* **Вторая постановка, или Объединение Моделей.** Она спрашивает: а существует ли общее между задачами знание **хотя бы для некоторых** участников, или хотя бы для некоторой (пусть малой, пусть сложной) их комбинации?

Тут следите внимательно за руками, потому что дальнейшее выглядит почти обманом.

Мы берём данные по акциям, пишем их в табличку. А затем берём данные по задачкам и... тоже пишем их в хвост той же самой таблички. **Как будто** это та же задача. Но, чтобы быть честными, добавляем колонку IsStock, сообщающую ML, к какому же типу задач относится каждая строка:

IsStock?	Label	p0	p1	p2	...
1	1	0	1	1	...
1	1	1	1	-1	...
1	0	1	0	2	...
1	0	1	-1	1	...
1	0	1	-1	2	...
1	1	-1	0	2	...
...	...	...	...	...	...
0	0	0	1	0	...
0	1	1	-2	1	...
0	1	-1	-1	2	...
...	...	...	...	...	...

Что сделает ML, если задачи эти **абсолютно** разные? Правильно. Он тупо разделит данные на две группы и натренирует для каждой свою отдельную модель:

$$ML(x) = \begin{cases} if\ (1 == IsStock)\ \ ML1(stock) \\ if\ (0 == IsStock)\ \ ML2(math) \end{cases}$$

В этом случае точность модели ML для данных по акциям будет такой же, как и точность отдельно натренированной на них модели ML1.

Но что, если между задачами имеется некоторая общность? Если выясняется, к примеру, что когда участник p1 говорит "1", и его мнение **не** равно мнению участника p7, то его ответ является прямым предсказанием метки, причём в обоих случаях? Тогда совместная модель, обученная на **смеси** данных, окажется несколько точнее на акциях, хотя дополнительный материал для её обучения и поступил из совершенно чуждой области знания. Просто потому, что если между паттернами предсказаний есть общность, её можно использовать. И измерить количественно, посмотрев, насколько улучшается предсказание цен акций **после** добавления к данным ответов на задачки. Разумеется, аналогичное сравнение надо провести, добавив в модель вместо истинных решений задачек такие же, но с "перепутанными" метками, чтобы вычесть случайный overfitting.

Этот метод отличается от предыдущего нечувствительностью к знанию, которое помогает решать **только одну** задачу. Он видит лишь "общее", то, что помогает обоим. И видит он вот что:

Regressor	residual MSFT	residual Random	residual Combined	Effect	Overfitting Effect	PureEffect	σ	pValue
RF	0.370 ± 0.012	0.378 ± 0.011	0.337 ± 0.011	0.033 ± 0.016	-0.008 ± 0.016	0.041 ± 0.020	2.05	2.0%
XG	0.391 ± 0.014	0.353 ± 0.016	0.311 ± 0.015	0.080 ± 0.021	0.038 ± 0.021	0.042 ± 0.026	1.62	5.3%
XG+	0.413 ± 0.012	0.417 ± 0.012	0.418 ± 0.011	-0.005 ± 0.016	-0.004 ± 0.017	-0.001 ± 0.020	-0.05	52.0%

Что это значит?

1. Самый мощный регрессор RF однозначно улавливает структурную общность моделей. Выигрыш после их объединения, даже после поправки на overfitting, составляет 0.041 ± 0.020, или 2.05σ. Вероятность случайно получить такой эффект на несвязанных наборах данных не превышает 2% (последняя колонка).

2. Регрессор XG чуть послабее, но тоже видит улучшение в 0.042 ± 0.026 (достоверность 94.7% = 100% - 5.3%)

3. А вот "антикрокодильный" регрессор... не видит ничего! Измеренная им разница составляет -0.001 ± 0.020, то есть ноль с шумом.

Получается забавная штука: да, структурная общность между моделями есть. Но возникает она не за счёт истинного знания, а за счёт "крокодильего". Грубо говоря, **одни и те же ошибки** как в первой, так и во второй модели позволяют ML (как объединяющему людей "разуму") "перевернуть" их и предсказать будущее. Но на индивидуальном уровне это знание является именно ошибкой, и лишь мешает участникам предугадывать картину мира.

60. А про новости?

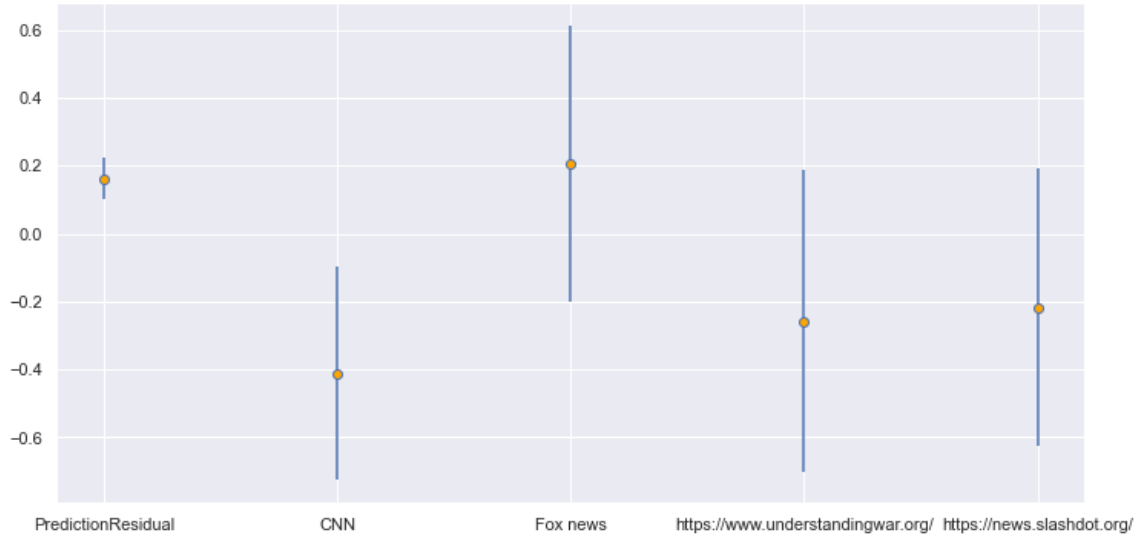
перепрыгнуть к следующему разделу - к предыдущему - к содержанию

В отличие от решений предыдущих задач, мы **не знаем** истинности новостных источников. Поэтому "проверить" их на истинность мы не можем. Но мы можем погонять метод теста ради.

В основе лежит та же идея, что и выше. Разделяем участников на две группы. Смотрим, какая из них лучше предсказывает задачу. Смотрим, как различается уровень доверия к тому или иному источнику между группами. Повторяем сотни раз, смотрим, есть ли статистически значимая корреляция. В принципе, подход можно улучшить, попросив ML предсказать эту разницу как функцию доверия... но, во-первых, для этого мало данных. А во-вторых, как быть, если эта функция вдруг окажется нетривиальной? "Помогает, пока вот настолько не доверяю, а потом мешает, а потом, если сильно доверяю, снова помогает"? Всякое, конечно, в природе бывает, но начнём лучше с того, что можем проглотить.

Итак, как способность предсказывать цену Майкрософта коррелирует с уровнем доверия к разным новостям?

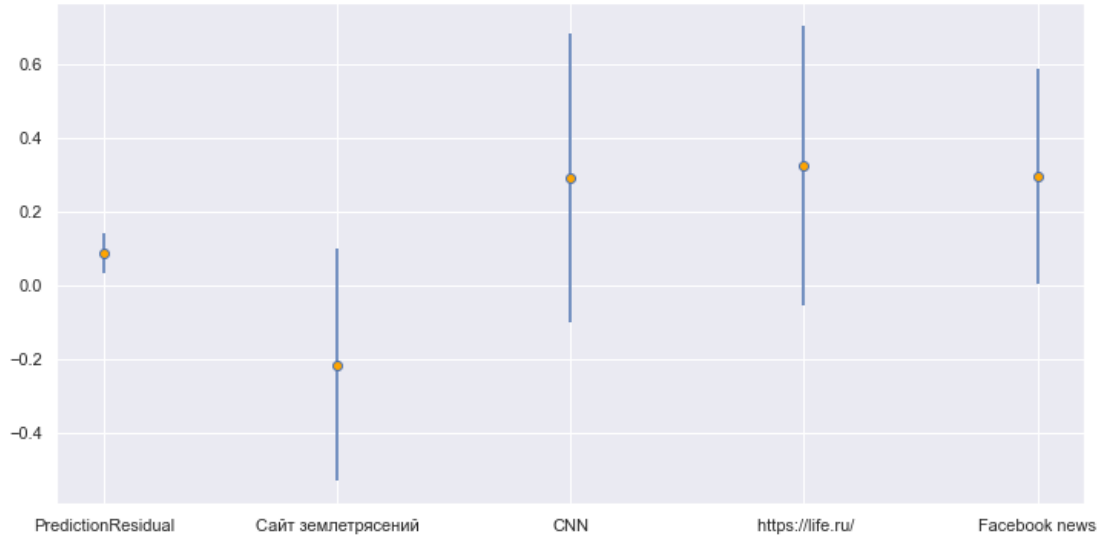
Ответ: слабо. Вот регрессор XG, оставлены только те новости, где значимость сигнала составляет хотя бы ±0.5σ:



Первая колонка -- разница в степени угадывания между "слабыми" и "сильными" группами. Как видно, она существенна и достоверна (0.16 ± 0.06).

Из второй колонки с достоверностью около 90% (разница -0.41 ± 0.32) вытекает, что если вы верите CNN, то вы **хуже** предсказываете цену акций Майкрософта. И это единственное, что более-менее точно. Влияние всех прочих источников имеет достоверность не выше 72%. Использование регрессора XG+, как ни странно, эту картину почти не меняет, что означает, что "крокодилья" составляющая здесь мала.

С техническими задачками картинка получается такая (регрессор XG, новости со значимостью сигнала хотя бы ±0.5σ):



Хотя доверие к CNN, life.ru, и новостям Фейсбука позитивно коррелирует со способностью решать технические задачи, достоверность всех трёх сигналов весьма сомнительна. Лишь Facebook News с большим скрипом дотягивает

до 1σ надёжности (0.30 ± 0.29), но и это означает лишь 1-pValue=85%. Может быть шумом, может быть слабым сигналом.

70. Итоги

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

1. Да, группа не-экспертов может, правильно натренировав ML на объединение своих мнений, получать предсказания по нетривиальным вопросам, в том числе таким, ответы на которые пока неизвестны никому. Например, ценам акций. Точность полученного предсказателя не очень велика (~60%), но ведь и данные у нас крошечные. Однако даже с такой точностью уже можно играть на рынке и в среднем выигрывать.

2. Подтверждена правильность представлений о существовании "крокодилов" и их вкладе в моделирование. Разработаны способы, измеряющие "крокодилью" компоненту в модели.

3. С помощью ML удалось детектировать перекрытие "вычислительных путей" принципиально разных моделей. Что в принципе открывает путь для проверки мировоззрений/представлений на истинность, даже когда они не дают прямых верифицируемых прогнозов, но...

4. ...но в нашем конкретном случае измеренный эффект оказался мал. И, что хуже, происходит он от не от общего правильного знания, а от общих ошибок участников.

5. Поэтому нам не удалось проверить Основную Гипотезу. А именно, что между достаточно разнородными моделями бывает общее **позитивное** знание. Мы его здесь не увидели.

Но нет ли его всегда, или нет только между данной парой моделей, или оно есть, но слишком мало, чтобы быть здесь замеченным? Мы не знаем. Вполне возможно, что из-за слабости предсказателя про Майкрософт (точность 60-65%) нужный сигнал просто не пробился через шум.

Весьма вероятно также другое объяснение. Какие участники внесли более 10% вклада в предсказание цен акций? Это номера 4, 6, 10, 11, 12, и 16. А кто внёс аналогичный вклад в задачи по физике и математике? Номера 5, 7, 12, 15, 17. За исключением номера 12, полностью другой список. Могло ли быть, что одним участникам были интереснее задачи по физике, и они честно про них думали, а про цену отписывались наугад, а другим было интереснее угадывать цену, а с математикой они не заморачивались? Вполне возможно. Вне зависимости от того, случилось ли это на самом деле, необходимо выработать механизм для "выравнивания" подобного перекоса на случай, если он появится.

6. Задним умом я сейчас понимаю, что опыт надо бы повторить на нескольких парах моделей, про которые мы **достоверно** знаем, что общее знание между ними есть. И таких, чтобы это знание при переходе от пары к паре убывало. Если удастся увидеть этот убывающий тренд, гипотезу можно будет считать доказанной, а метод -- работающим.

7. Мы побаловались измерением корреляции между доверием к новостным источникам и способностью модели предсказывать будущее. Увидели кое-что любопытное. Но, поскольку Основная Гипотеза не доказана, мы не можем достоверно сказать ничего объективного об истинности того или иного источника.

8. Очень большой проблемой оказались шумы и пропуски данных. Не меньше половины ответов выглядят простыми случайностями, и лишь четверть участников заполнила хотя бы 80% вопросов. Это не обвинение, это констатация реальности, с которой я не совсем сумел справиться. Я ожидал и шума, и пропусков, но заготовленные для борьбы с ними методы оказались не на высоте. Тут потребуется ещё немало работы.

80. Разное. Веса участников.

перепрыгнуть к следующему разделу - к предыдущему - к содержанию

В задачах по физике и математике:

Feature	Weight RF	%RF	Sigmas RF	crocRange
p0	0.0004 ± 0.0014	(0.2 ± 0.8)%	0.27	NA
p1	0.003 ± 0.003	(1.7 ± 1.8)%	0.98	0%...100%
p2	0.0043 ± 0.002	(2.5 ± 1.1)%	2.2	NA
p3	-0.0016 ± 0.0022	(-0.9 ± 1.3)%	-0.73	NA
p4	0.0015 ± 0.0027	(0.9 ± 1.6)%	0.54	NA
p5	0.0424 ± 0.0136	(24.9 ± 8)%	3.11	45%...100%
p6	-0.0003 ± 0.001	(-0.2 ± 0.6)%	-0.31	NA
p7	0.0247 ± 0.0111	(14.5 ± 6.5)%	2.23	60%...100%
p8	0.0019 ± 0.0018	(1.1 ± 1)%	1.09	NA
p9	-0.0024 ± 0.0027	(-1.4 ± 1.6)%	-0.91	NA
p10	0.0007 ± 0.0005	(0.4 ± 0.3)%	1.33	NA
p11	0.0013 ± 0.0015	(0.8 ± 0.9)%	0.86	NA
p12	0.0303 ± 0.0087	(17.8 ± 5.1)%	3.51	1%...100%
p15	0.0352 ± 0.0154	(20.6 ± 9)%	2.29	35%...100%

p16	-0.0008 ± 0.0011	(-0.5 ± 0.6)%	-0.74	NA
p17	0.0301 ± 0.0193	(17.7 ± 11.3)%	1.56	13%...100%

Первая колонка -- номер участника. Вторая колонка -- абсолютный вклад в точность предсказания методом RandomForest. Третья (%RF) -- процентная доля вклада участника в общем предсказании. Зелёным обозначены участники, внёсшие хотя бы 10%. Четвёртая (Sigmas RF) -- статистическая достоверность вклада. Всё, что больше единицы, более-менее достоверно. Отрицательные значения означают, что сигнал в ответах был подавлен шумом. Наконец, последняя колонка (crocRange) -- попытка оценить долю "крокодильей" компоненты во вкладе участника -- там, где это было статистически возможно.

Аналогичная табличка к задаче про Майкрософт выглядит несколько печальнее:

Feature	Weight RF	%RF	Sigmas RF	crocRange
p0	0.0028 ± 0.0051	(7.3 ± 13.3)%	0.55	NA
p1	0.0021 ± 0.0078	(5.4 ± 20.5)%	0.26	NA
p2	0.003 ± 0.0055	(7.8 ± 14.5)%	0.54	NA
p3	-0.0017 ± 0.0027	(-4.5 ± 7)%	-0.65	NA
p4	0.0124 ± 0.0173	(32.4 ± 45.2)%	0.72	0%...100%
p5	-0.0054 ± 0.0073	(-14 ± 19.2)%	-0.73	NA
p6	0.0095 ± 0.0074	(24.9 ± 19.4)%	1.28	NA
p7	-0.0009 ± 0.0033	(-2.3 ± 8.5)%	-0.27	NA
p8	-0.0019 ± 0.0053	(-5.1 ± 13.8)%	-0.37	#DIV/0!
p9	-0.004 ± 0.0067	(-10.4 ± 17.6)%	-0.59	NA
p10	0.0049 ± 0.0064	(12.9 ± 16.8)%	0.77	NA
p11	0.0063 ± 0.0158	(16.6 ± 41.2)%	0.4	#DIV/0!
p12	0.0075 ± 0.0121	(19.6 ± 31.8)%	0.62	NA
p15	-0.0038 ± 0.0049	(-10.1 ± 12.9)%	-0.78	NA
p16	0.0065 ± 0.0098	(16.9 ± 25.6)%	0.66	NA
p17	0.001 ± 0.0038	(2.6 ± 9.9)%	0.26	NA

Шума больше, и лишь для одного участника (№6) его вклад вычислен более-менее достоверно. Всё прочее сильно зашумлено.

82. Разное. А где код?

к предыдущему - к содержанию

Сначала код писался по правилам. Был замысел. Были функции. Были юнит-тесты для всей функциональности.

Потом время стало поджимать, и я забросил юнит-тесты. Но структурирование в виде функций сохранилось.

Потом пошёл просто копипаст. Когда уже нет времени обобщать, а надо решить "вот эту вот конкретно узенькую задачу", и пофиг, что вокруг плавают три копии отличающегося мелочами кода.

Здесь проявилась главная ошибка: недооценка разнообразия потребовавшихся пайплайнов данных. Как я рассуждал? Ну, вот есть сырые данные. Читаем, преобразуем, разделяем. Учимся, тестируем, собираем ответ. Ну, может быть в варианте A/B. Ну, может быть, в цикле. Делов-то, что я, транспонирование или join нужный не напишу? Напишу. Но оказалось, что нужны были десятки разных путей от исходных данных до ответа. Которые, в итоге, были написаны "грубой силой". А по-хорошему надо было создать класс Transform, от которого наследовать разные трансформеры, каждый со своим конструктором, определяющим, как и что будет преобразовываться. И лепить из них pipelines, как из лего. Если буду повторять подобный эксперимент, сделаю так.

Потом... потом на это были наложены хаки, патчи, и "волшебные" if-ы. В надежде уж досчитать, как есть, лишь бы не переписывать архитектуру под конец проекта.

Ну и самый последний расчёт вёлся уже хаками, навешенными на хаки. Код превратился в исландскую сагу, которую я сам с трудом понимаю.

Но, кажется, он сделал всё, что я пытался из него выжать.

Только вот непонятно, стоит ли им делиться. Если кому-то Очень Хочется, пишите.

Евгений Бобух,

25.10.2022

===

Text Author(s): Eugene Bobukh === Web is volatile. Files are permanent. **Get a copy:** [PDF] [Zipped HTML] === **Full list of texts:** http://tung-sten.no-ip.com/Shelf/All.htm] === **All texts as a Zip archive:** http://tung-sten.no-ip.com/Shelf/All.zip] [mirror: https://1drv.ms/u/s!AhyC4Qz62r5BhO9Xopn1yxWMsxtaOQ?e=b1KSiI] === **Contact the author:** h o t m a i l (switch name and domain) e u g e n e b o (dot) c o m === **Support the author:** 1. **PayPal** to the address above; 2. **BTC**: 1DAptzi8J5qCaM45DueYXmAuiyGPG3pLbT; 3. **ETH**: 0xbDf6F8969674D05cb46ec75397a4F3B8581d8491; 4. **LTC**: LKtdnrau7Eb8wbRERasvJst6qGvTDPbHcN; 5. **XRP**: ranvPv13zqmUsQPgazwKkWCEaYecjYxN7z === **Visit other outlets:** Telegram channel http://t.me/eugeneboList, my site www.bobukh.com, Habr https://habr.com/ru/users/eugeneb0/posts/, Medium https://eugenebo.medium.com/, Wordpress http://eugenebo.wordpress.com/, LinkedIn https://www.linkedin.com/in/eugenebo, ЖЖ https://eugenebo.livejournal.com, Facebook https://www.facebook.com/EugeneBo, SteemIt https://steemit.com/@eugenebo, MSDN Blog https://docs.microsoft.com/en-us/archive/blogs/eugene_bobukh/ === **License**: Creative Commons BY-NC (no commercial use, retain this footer and attribute the author; otherwise, use as you want); === **RSA Public Key Token:** 33eda1770f509534. === **Contact info** relevant as of 7/15/2022.

===