

Data Science в банке — ЖИВ ИЛИ МЕРТВ?

Афанасьев Сергей
 КБ «Ренессанс Кредит»

13 февраля 2019 г.
 Москва

Что говорят про банковский Data Science?

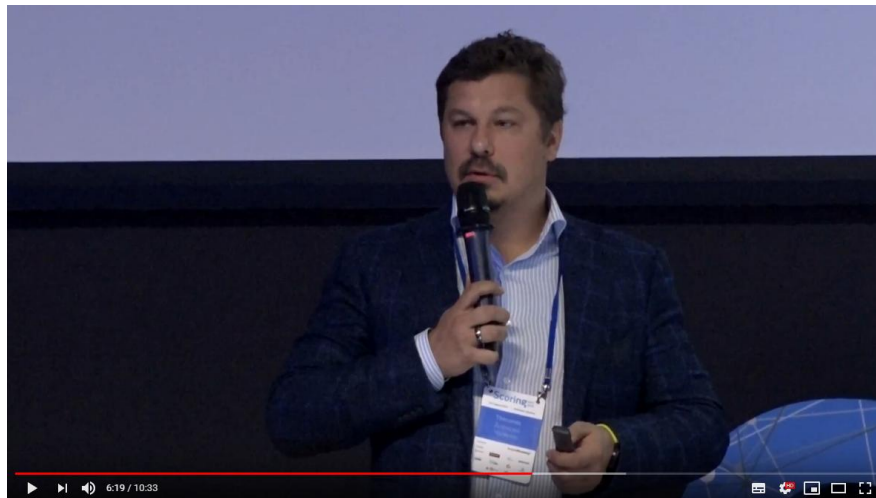
– Data Science в банках мертв!
Руководители Яндекса и
Рамблера не хотят брать на
работу людей из банков.



2016

Что говорят про банковский Data Science?

– Кто из вас работает с алгоритмами логарифмической регрессии? Можно руки поднять? Отлично! Если вас не возьмут в KPMG, вы тоже **можете менять профессию**. Это очень хорошая компетенция, для того чтобы торговать хот-догами.



2018

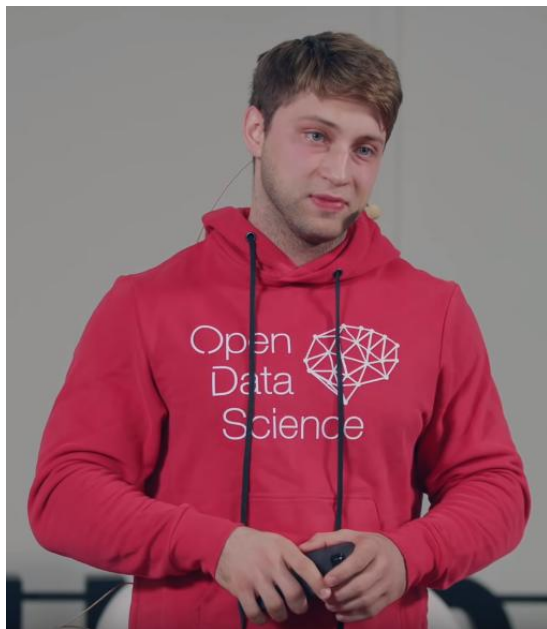
Что обсуждают звезды ODS?



Алексей Натекин

– По факту бизнесу может быть нужна пара тупых «if-then'ов». Если ваш машин лернинговый отдел и вы, как Head of Data Science, притащите какое-то решение, которое сделано на «if-then'ах», но решает бизнес-задачу, вы все равно огромный молодец – вы решили задачу.

Что обсуждают звезды ODS?



Валерий Бабушкин

– У вас нейронка может в случае отсутствия скрытых слоев свестись просто к линейной регрессии. И все, вот ваша линейная регрессия, но вы смело говорите: – **У меня нейронка!**

Что обсуждают звезды ODS?



Роман Чеботарев

– Если катить в прод какую-либо модельку, то мы отдадим, при прочих равных, **предпочтение линейной регрессии**, которая будет обвешана кучей «if'ов». Потому что мы работаем в среде с очень большой ценой ошибки. Если перекрутили где-нибудь насос на нефтедобывающей скважине и скважина остановилась – это потери приблизительно в миллион долларов.

- Либо жив, либо мертв.



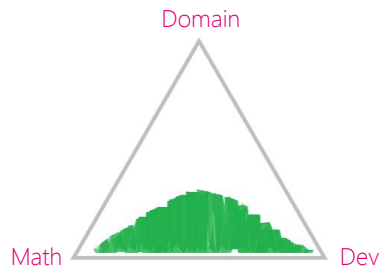
Таксономия Натекина



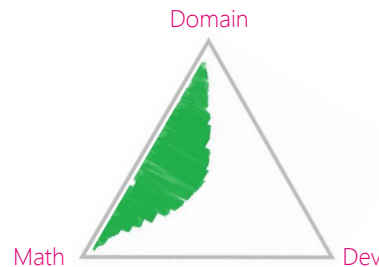
Data Scientist:

- Математик (Stats, ML, Algorithms)
- Эксперт (Business & Domain expertise)
- Разработчик (Devops, SWE, Programming)

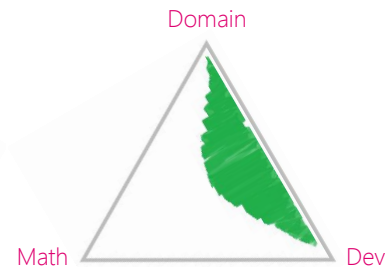
Таксономия Наткина



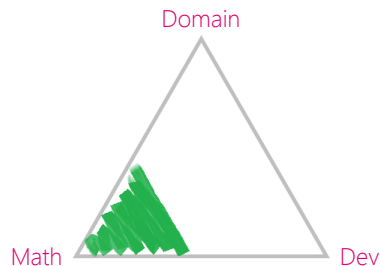
ML Engineer



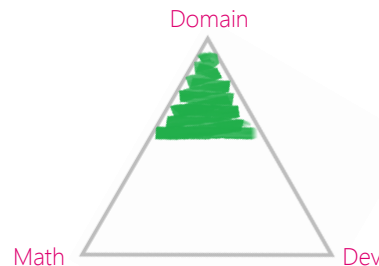
Data Analyst



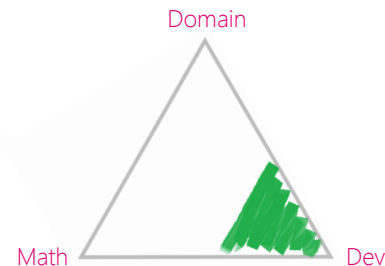
Data Engineer



ML Researcher

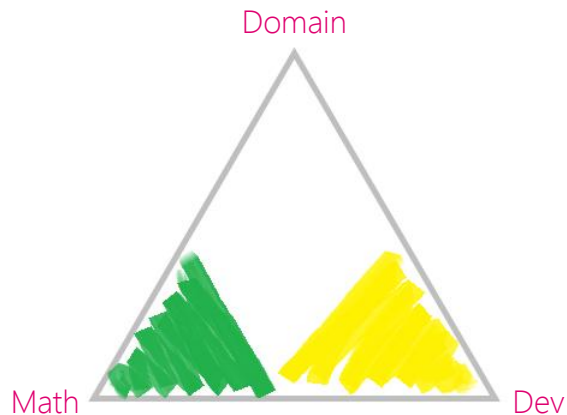


Analyst

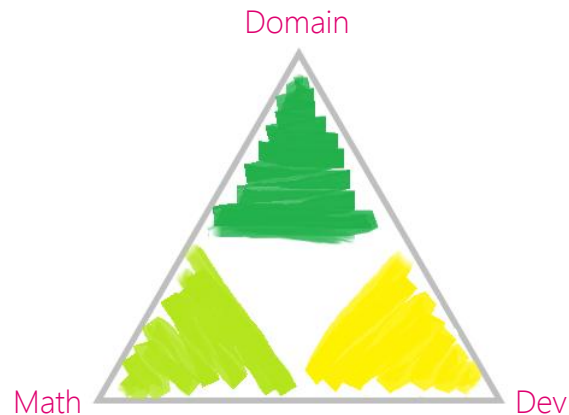


Devops

Новая школа или старая?



Неофит



Скорингист

Даже если вам немного за 30...



johnpateha 24 октября 2018 в 14:01

Как стать датасайентистом, если тебе за 40 и ты не программист

Блог компании QIWI, Data Mining, Карьера в IT-индустрии, Машинное обучение, Учебный процесс в IT

Бытует мнение, что стать датасайентистом можно только имея соответствующее высшее образование, а лучше ученую степень.

Однако мир меняется, технологии становятся доступны и для простых смертных. Возможно, я кого-то удивлю, но сегодня любой бизнес-аналитик в состоянии освоить технологии машинного обучения и добиться результатов, конкурирующих с профессиональными математиками, и, возможно, даже лучших.

Дабы не быть голословным, я расскажу вам свою историю — как из экономиста я стал дата-аналитиком, получив необходимые знания через онлайн-курсы и участвуя в соревнованиях по машинному обучению.

<https://habr.com/company/qiwi/blog/427311/>



Цели Data Science в Банке

1

**Повышение эффективности
банковских процессов**

- Scoring
- Collection scoring
- CRM-models
- Fraud-models

2

**Автоматизация ручного труда
и сокращение издержек**

- Голосовые помощники
- Чат-боты
- Auto Help-Desk

3

**Безопасность и качество
клиентского сервиса**

- Биометрия
- Мобильные технологии
- Голосовые помощники
- Чат-боты

Задачи Data Science в банке

	Risk, Antifraud	Collection	CRM	CC, TM, CS	IT, IT Security	Other
Classical Machine Learning	Credit Scoring <ul style="list-style-type: none"> - Application scoring - Behavioral scoring Anti-fraud Models <ul style="list-style-type: none"> - Внутренний фрод - Внешний фрод - Транзакционный фрод 	Collection Scoring <ul style="list-style-type: none"> - Модели Recovery Rate 	Recommender System <ul style="list-style-type: none"> - Модели отклика на коммуникацию Targeted Advertising <ul style="list-style-type: none"> - Таргетирование рекламы на сайте банка Churn Prediction <ul style="list-style-type: none"> - Модели оттока 		Biometrics <ul style="list-style-type: none"> - Keystroke Dynamics & Mouse Movements - Device Print для ИБ/МБ Error Detection <ul style="list-style-type: none"> - Выявление сбоев в системах Банка 	AML <ul style="list-style-type: none"> - AML правила и модели Time Normalization <ul style="list-style-type: none"> - Оптимальная загрузка массовых подразделений Staff Recruitment <ul style="list-style-type: none"> - Отбор резюме по ключевым словам
Natural Language Processing	Voice Recognition <ul style="list-style-type: none"> - Аутентификация клиентов по голосу Fraud Text Mining <ul style="list-style-type: none"> - Анализ корпоративной переписки 	Text-To-Speech <ul style="list-style-type: none"> - Голосовой коллектор 	Recommender System <ul style="list-style-type: none"> - Анализ транзакции с помощью RNN для построения моделей отклика 	Speech recognition <ul style="list-style-type: none"> - Извлечение инфо - Анализ тональности - Чат-боты (сайт, ИБ/МБ) CS Text Mining <ul style="list-style-type: none"> - Извлечение инфо - Анализ тональности - Чат-боты (сайт, ИБ/МБ) Text-To-Speech <ul style="list-style-type: none"> - Синтез речи 	Security Text Mining <ul style="list-style-type: none"> - Проверка email-писем на утечку информации Topic Model <ul style="list-style-type: none"> - Распределение заявок в HD по тематикам 	Post Mail Classifier <ul style="list-style-type: none"> - Распределение почты по тематикам HR Text Mining <ul style="list-style-type: none"> - Анализ переписки и выявление проблем в работе сотрудников
Computer Vision	Photo Biometrics <ul style="list-style-type: none"> - Вход в ИБ/МБ - Вход в системы Банка Object Detection <ul style="list-style-type: none"> - Скоринг по фото - Проверка фото ТТ Spoofing <ul style="list-style-type: none"> - Выявление подделок 			OCR (Text recognition) <ul style="list-style-type: none"> - Перевод скан.копий документов в текст. 		Photo Biometrics <ul style="list-style-type: none"> - Аутентификация клиентов в ДО/ККО - Пропускные системы для сотрудников

Коллекшн-модели банка из топ-1

Предконтрактная стадия

1. Модели для оценки цессионных портфелей
2. Модели для workforce management
3. Модели для resource management

Верификация и обогащение данных

4. Модели для распознавания captcha при автоматизированном сборе данных
5. Модели для оценки вероятности найти информацию при помощи skip tracing
6. Модели для оценки вероятности принадлежности найденного контакта должнику
7. Модели для мэтчинга профилей физлица в различных онлайн-источниках

Робот-коллектор

8. Модели для определения наличия голоса в звуковом канале
9. Модели для распознавания речи
10. Модели для выбора направления движения по скрипту общения робота в зависимости от слов должника

Soft Collection

11. Модели для оценки мат.ожидания Recovery Rate
12. Модели для оценки вероятности дозвона по телефонному номеру
13. Модели для оценки ожидаемого финансового эффекта от звонка
14. Модели для прогноза длительности звонка
15. Модели для оценки вероятности получить обещание об оплате
16. Модели определения наличия автоответчика
17. Модели для определения результата переговоров с должником
18. Модели для рекомендации оптимальных мотиваторов
19. Разговорные модели для текстовых коммуникаций
20. Модели для динамического управления процессами взыскания на базе RNN

Hard / Legal / Executory collection

21. Модели для оценки мат.ожидания Recovery Rate
22. Модели для оценки финансового эффекта от выезда сотрудника выездного взыскания



Радость

Это может быть бесплатно

Многие AI-технологии являются Open Source:
Одна из лучших систем распознавания лиц FaceNet от компании Google выложена в открытый доступ. Бесплатно можно пользоваться биометрией от Microsoft и других ИТ-гигантов.

AI эффективнее людей

В розничном кредитовании скоринговые карты предсказывают просрочку лучше, чем андеррайтеры. Нейронные сети распознают лица на фотографиях лучше, чем человек.

И – инновации

Сегодня Data Science – это быстро растущая инновационная сфера. Развитие Data Science в банке позволит быть на передовой, быстро реагировать на изменения рынка и не отставать от конкурентов.

Любовь сотрудников

Молодые специалисты любят компании, в которых занимаются AI-технологиями. Для них это своего рода «тусовка».

AI заменяет людей



Боль

Это может быть дорого

Даже при использовании Open Source затраты на доработки ИТ-инфраструктуры могут быть огромными. Например, по некоторым исследованиям на цифровую трансформации среднего европейского банка требуется порядка €1 млрд.

Люди эффективнее AI

Чат-боты, голосовые помощники, роботизированные колл-центры и т.д. пока не приближаются к человеку по качеству понимания контекста. Это сильно раздражает клиентов.

Возможно это пузырь

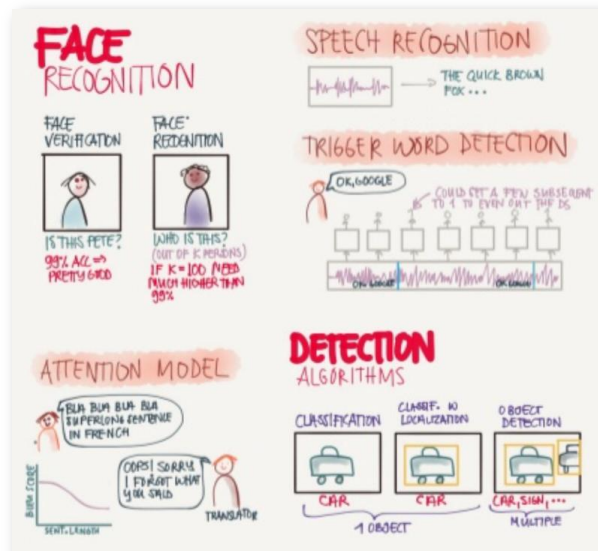
Многие AI-технологии были придуманы десятки лет назад. Программа для распознавания текста Tesseract «пролежала на полке» 10 лет. Возможно многие AI-технологии сейчас переоценены и не будут применяться в ближайшие десятилетия.

Дорогие специалисты

Сейчас наблюдается высокий спрос на Data Scientist'ов, поэтому они стоят дорого. Но их становится много и з/п должны падать.

AI заменяет людей

Кейсы



Примеры внедрений в банках

Голосовая биометрия

Топ-26

В банке из топ-26 разработали свою собственную голосовую биометрию, которую внедрили на процессах Колл-центра и Коллекшн (идентификация клиентов).

Рабочая группа: ~10 человек

Срок разработки: ~1,5 года

Запуск: 2018 г.

Боль:

Потребовались дорогие доработки ИТ-инфраструктуры.



Примеры внедрений в банках

Post Mail Classifier

Топ-34

В банке из топ-34 построили AI, который позволяет разделять почтовую корреспонденцию по тематикам.

Технология позволяет распознавать печатные сканы писем, разделять их на 250+ тематик и формировать автоматический ответ на распознанные письма.

Для распознавания текста использовали **Open Source** технологию OCR: программу Tesseract от компаний Google.

End-to-end время: **10-40 сек**

Качество распознавания печатных текстов: **78%**

Про проект

Задача:
Ускорить процесс анализа бумажной корреспонденции
Понять набор тематик, определить обработчика
Сформировать ответ автоматически

Поставим OCR перед NLU?

Текущий статус:
WIP

18

2018. Frankenstein Forum & Awards. The Retail Finance.

Примеры внедрений в банках

Open Source фотобиометрия

Топ-6

Специалисты из банка топ-6 (украинский филиал) использовали **бесплатные** Open Source технологии от Microsoft и Google для обработки фотографий клиентов.

Из фото клиента извлекалась информация:

- 1) Эмоции на лице;
- 2) Различные объекты (машины, пляж, дети и т.п.);
- 3) Задний фон

Извлеченную информацию использовали:

1. Как дополнительные предикторы в скоринге: например, выяснилось, что **лучшие клиенты** – это **блондинки**, а клиенты, сфотографированные на фоне дорогих авто – чаще допускают просрочки;
2. Для выявления «**черных брокеров**»: выявлялись одинаковые предметы на фоне разных фотографий клиентов, сделанных якобы в разных ТТ.


ИДЕНТИФИКАЦИЯ КЛИЕНТОВ ОТ БРОКЕРА

amazon web services
amazon RECOGNITION

Google Cloud Platform
CLOUD VISION API

Microsoft Azure
FACE API

IBM Watson
IMAGE CLASSIFICATION




Описание: ["tags": ["person", "indoor", "woman", "sitting", "table", "holding", "front", "black", "smiling", "young", "man", "glass", "background", "blouse", "winter", "looking", "tags"], "captions": [["text": "a woman sitting on a table", "confidence": 0.7737947]]]

Текст: [["name": "person", "confidence": 0.996421], ["name": "wall", "confidence": 0.976095], ["name": "indoor", "confidence": 0.9215844], ["name": "woman", "confidence": 0.9023877]]

Формат изображения: "jpg"

Размеры изображения: 275 x 297

БЛОНДИНКИ – ИДЕАЛЬНЫЙ КЛИЕНТ



ALL	30-3MOB
3 993	2.5%
173	0%

2017. Антифрод в банке. MSB Evens.

Примеры внедрений в банках

Чат-боты

Топ-1

В рамках конференции SDSJ-2017 Сбербанк провел соревнования по разработке чат-бота. Призеры получили предложения работать в лаборатории AI Сбербанка.

Топ-34

В банке Хоум Кредит разрабатывают чат-бота на основе AI, который расширяет возможности «кнопочного» чат-бота. Разработка находится на стадии пилотирования.

Топ-26

В банке Тинькофф разработали чат-бота, который выдает подсказки на вопросы клиентов. В диалоге пока используют живых операторов, которым чат-бот также помогает с подсказками для ответов.



2017. Sberbank Data Science Journey.

Бывают неудачи



Kaggle – что мы оттуда узнали?

Соревнования по скорингу

Топ-34

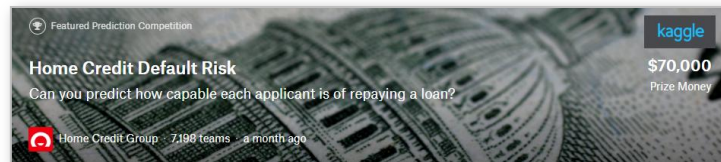
Летом 2018 г. банк Хоум Кредит провел соревнования на Kaggle по разработке скоринговой модели:

- Соревнования собрали более 7.000 команд
- Призовой фонд составил \$70.000

В рамках соревнований был выложен обезличенный нормированный датасет банка. В датасете были данные с характеристиками домов заемщиков:

- | | |
|-----------------------------|---------------------|
| 1) Год ввода в эксплуатацию | 7) Спорт. площадка |
| 2) Количество квартир | 8) Площадь (жилая) |
| 3) Количество этажей | 9) Площадь парковки |
| 4) Количество лифтов | 10) Аварийность |
| 5) Количество подъездов | 11) Кап. ремонт |
| 6) Детская площадка | И т.д. |

Эти данные могут косвенно подтверждать финансовый статус клиента. Данные о домах выложены на открытых интернет-ресурсах: <http://dom.mingkh.ru/>



<https://www.kaggle.com/c/home-credit-default-risk>



<https://youtu.be/H4sl3iMTCns>



KonstantinKG 19 июня в 23:12

Соревнование Kaggle Home Credit Default Risk — анализ данных и простые предсказательные модели

<https://habr.com/post/414613/>

Что мы узнаем из научных статей?

Скоринг по данным cookies

Немецкие исследователи обогатили скоринг европейских БКИ данными cookies и получили рост AUC ROC модели на 5,3%.

Для обогащения были использованы открытые cookies:

1) Computer & Operating system

- Desktop/Windows
- Desktop/Macintosh
- Tablet/Android
- Tablet/iOS
- Mobile/Android
- Mobile/iOS

2) Email Host

- Gmx (partly paid)
- Web (partly paid)
- T-Online (affluent customers)
- Gmail (free)
- Yahoo (free, older service)
- Hotmail (free, older service)

3) Channel

- Paid
- Affiliate
- Direct
- Organic

4) Check-Out Time

- Evening (6pm-midnight)
- Morning (6am-noon)
- Afternoon (noon-6pm)
- Night (midnight-6am)

5) Name In Email

6) Number In Email

7) Is Lower Case

8) Email Error

On the Rise of FinTechs – Credit Scoring using Digital Footprints

Tobias Berg[†], Valentin Burg[‡], Ana Gombovic^{*}, Manju Puri^{*}

July 2018

Abstract

We analyze the information content of the digital footprint – information that people leave online simply by accessing or registering on a website – for predicting consumer default. Using more than 250,000 observations, we show that even simple, easily accessible variables from the digital footprint equal or exceed the information content of credit bureau scores. Furthermore, the discriminatory power for unscorable customers is very similar to that of scorable customers. Our results have potentially wide implications for financial intermediaries' business models, for access to credit for the unbanked, and for the behavior of consumers, firms, and regulators in the digital sphere.

We wish to thank Frank Ecker, Falko Fecht, Christine Laudénbach, Laurence van Lent, Kelly Shue (discussant), Sascha Steffen, as well as participants of the 2018 RFS FinTech Conference, the 2018 Swiss Winter Conference on Financial Intermediation, and research seminars at Duke University, FDIC, and Frankfurt School of Finance & Management for valuable comments and suggestions. This work was supported by a grant from FIRM (Frankfurt Institute for Risk Management and Regulation).

[†] Frankfurt School of Finance & Management, Email: t.berg@fs.de, Phone: +49 69 154008 515.

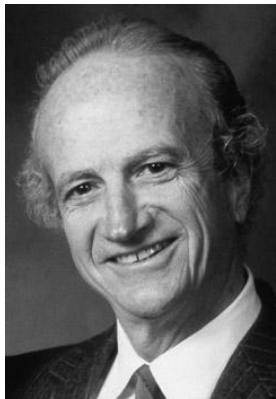
[‡] Humboldt University Berlin, valentin.burg@gmail.com.

^{*} Frankfurt School of Finance & Management, Email: a.gombovic@fs.de, Phone: +49 69 154008 830.

^{*} Duke University, FDIC, and NBER, Email: mpuri@duke.edu, Tel: (919) 660-7657.

Экономика преступления и наказания

Концепция Гэри Беккера



$$(1 - \pi)U(W_C) - \pi S > U(W_L)$$

π – вероятность быть пойманным;

$U()$ – функция полезности индивида;

S – санкции, понесённые в случае поимки (например, штраф);

W_C и W_L – соответственно доходы от преступной и легальной деятельности

Левая часть неравенства характеризует то, что связано с преступностью, правая – полезность от законного заработка. Логично, что если неравенство будет выполнено, индивид при прочих равных предпочтёт нарушить закон.

Основные факторы образования равновесного уровня преступности:

- Готовность людей совершать преступления ради **выгоды**: вероятность ареста, осуждения и наказания; меры наказания; доходы от альтернативных видов легальной и нелегальной деятельности; **риск безработицы; изначальный уровень благосостояния.**
- Поведение потенциальных жертв и потребителей **нелегальных товаров**: величина спроса на нелегальные товары; спрос на средства защиты (сейфы, сигнализацию, охрану).
- Меры, принимаемые государством: осуществляя поимку преступников, государство вводит своеобразный «налог» на преступные виды деятельности, выражаемые в риске быть пойманным.

“Take Home Test” для кандидатов

Задача 1: Baseline model

1. Из открытых источников собрать sample в разбивке «регион РФ + год»:

Target: Уровень преступности;

Feature_1: Уровень безработицы;

Feature_2: Уровень средних заработков.

2. Обучить модель регрессии и проверить теорию Беккера на переменных Feature_1 и Feature_2

Задача 2: User model

1. Самостоятельно сформулировать критерии задачи и придумать признаки.

2. Обучить модель регрессии, проверить теорию Беккера и сравнить с Baseline model.

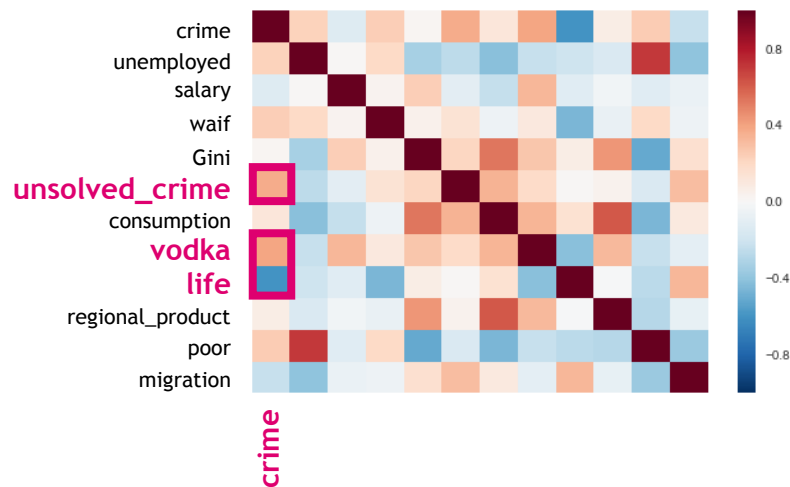
Baseline model:

R2 (Reg) = 0.1343

User model:

R2 (Reg) = 0.5633

R2 (GBR) = 0.7146



«Делать внутри компании
датасайнс отделы - это очень
классно и хорошо вас продвинет»

16:17



16:18 ✓✓

А кого продвинет? Вас или
компанию?

16:35 ✓✓

Спасибо за
внимание!

Афанасьев Сергей

Head of Data Science

Head of Antifraud

КБ «Ренессанс Кредит»

safanasev@rencredit.ru