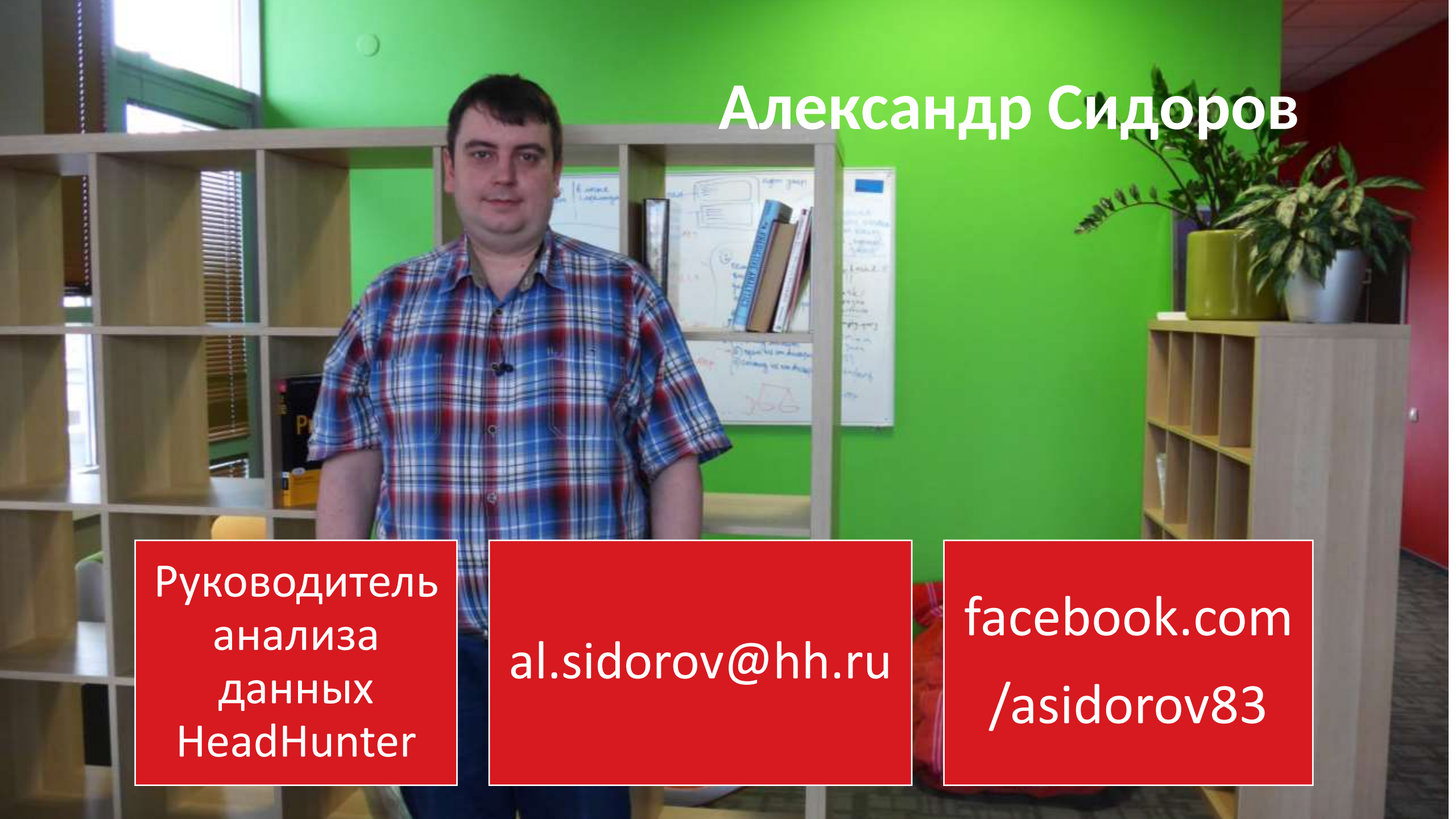
Машинное обучение в HeadHunter: умный поиск соискателей и работодателей

Александр Сидоров



Александр Сидоров

Руководитель анализа данных HeadHunter

al.sidorov@hh.ru

facebook.com/asidorov83

Кейс 1: Прескриннинг/модерация резюме

Задача	• Проверять качество заполнения резюме на самом верху воронки подбора
Проблема	• Значительная часть соискателей плохо и неполно заполняют резюме
Решение	• Использовать исторические данные по модерации резюме, чтобы автоматически подтверждать самые качественно заполненные

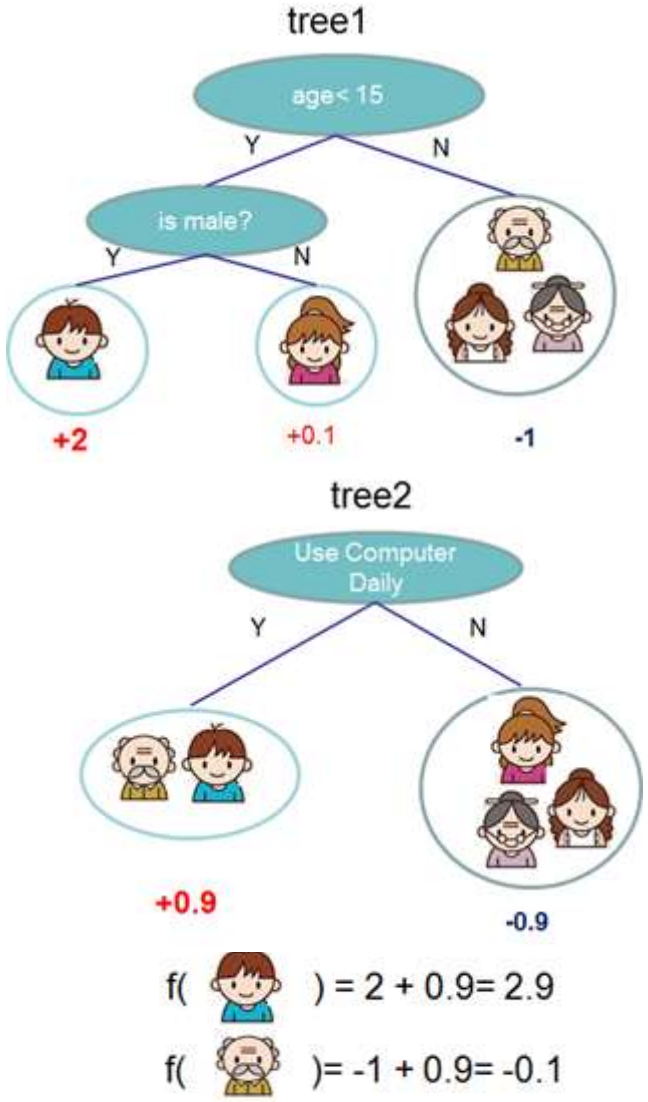
Схема работы автомодерации

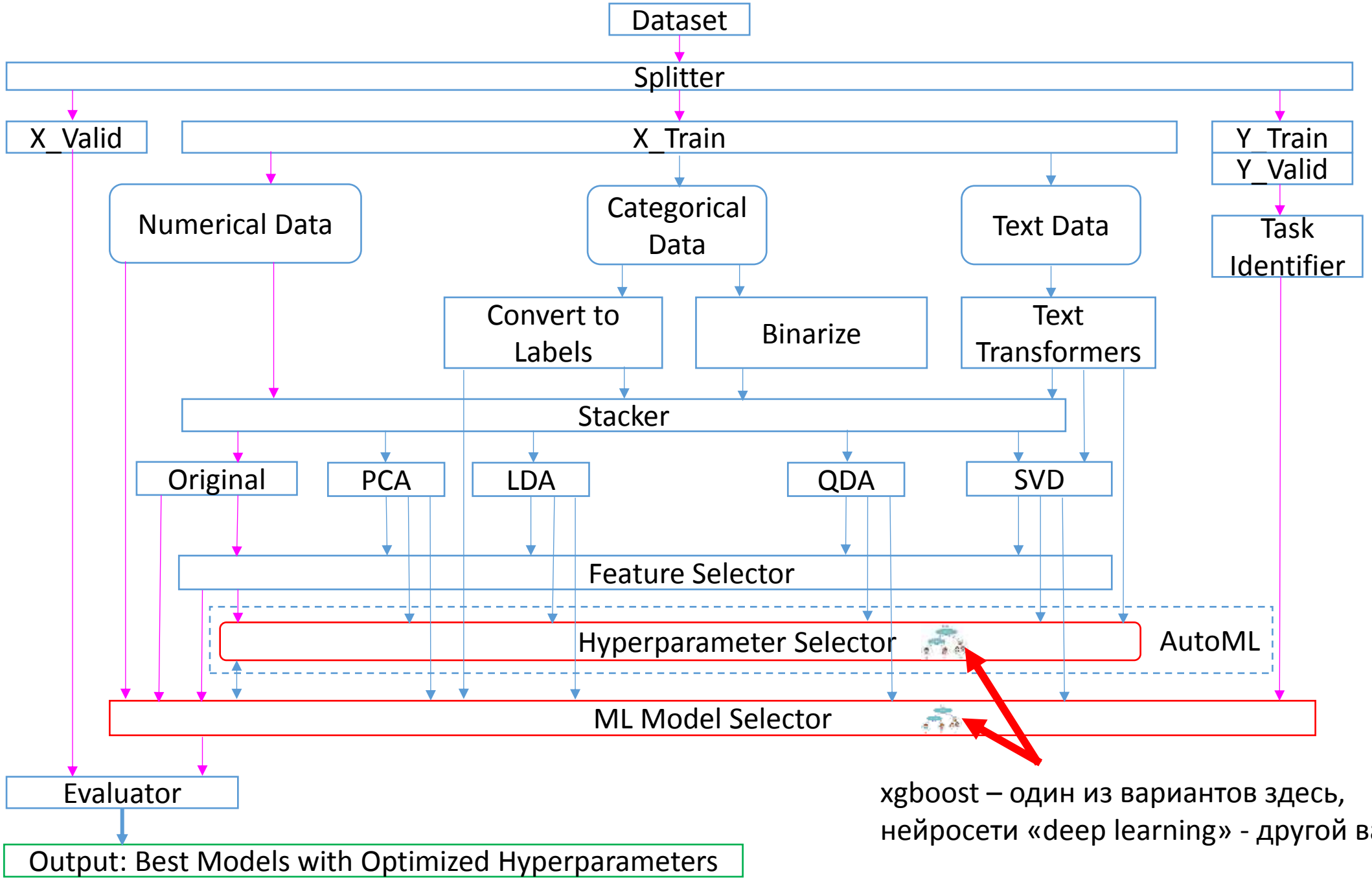
1. Результаты ручной модерации - исторические данные
2. Научили модель автомодерации
3. Новые резюме проверяются автомодератором
4. Качественные резюме автоматически подтверждаются
5. Спорные случаи отправляются людям

>30k новых резюме в день, ↑, кол-во модераторов и качество модерации const, auc_roc 0.94
200k резюме в обучающем множестве, 1k признаков

Градиентный бустинг с xgboost

- хорошо работает с разнородной информацией
- не очень чувствителен к шуму
- довольно чувствителен к переобучению
- многопоточный, хорошая параллельность
- и бинарная классификация, и регрессия
- подбираем eta, n_estimators, max_depth
- можно полуавтоматически, через hyperopt
- «when in doubt – use xgboost» Owen Zhang, Kaggle

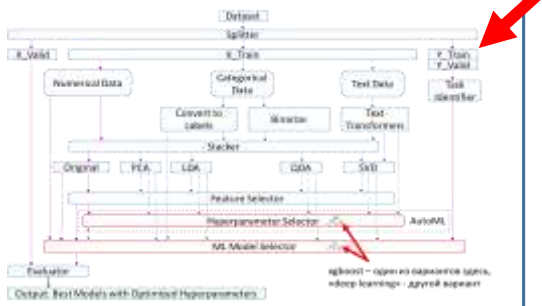


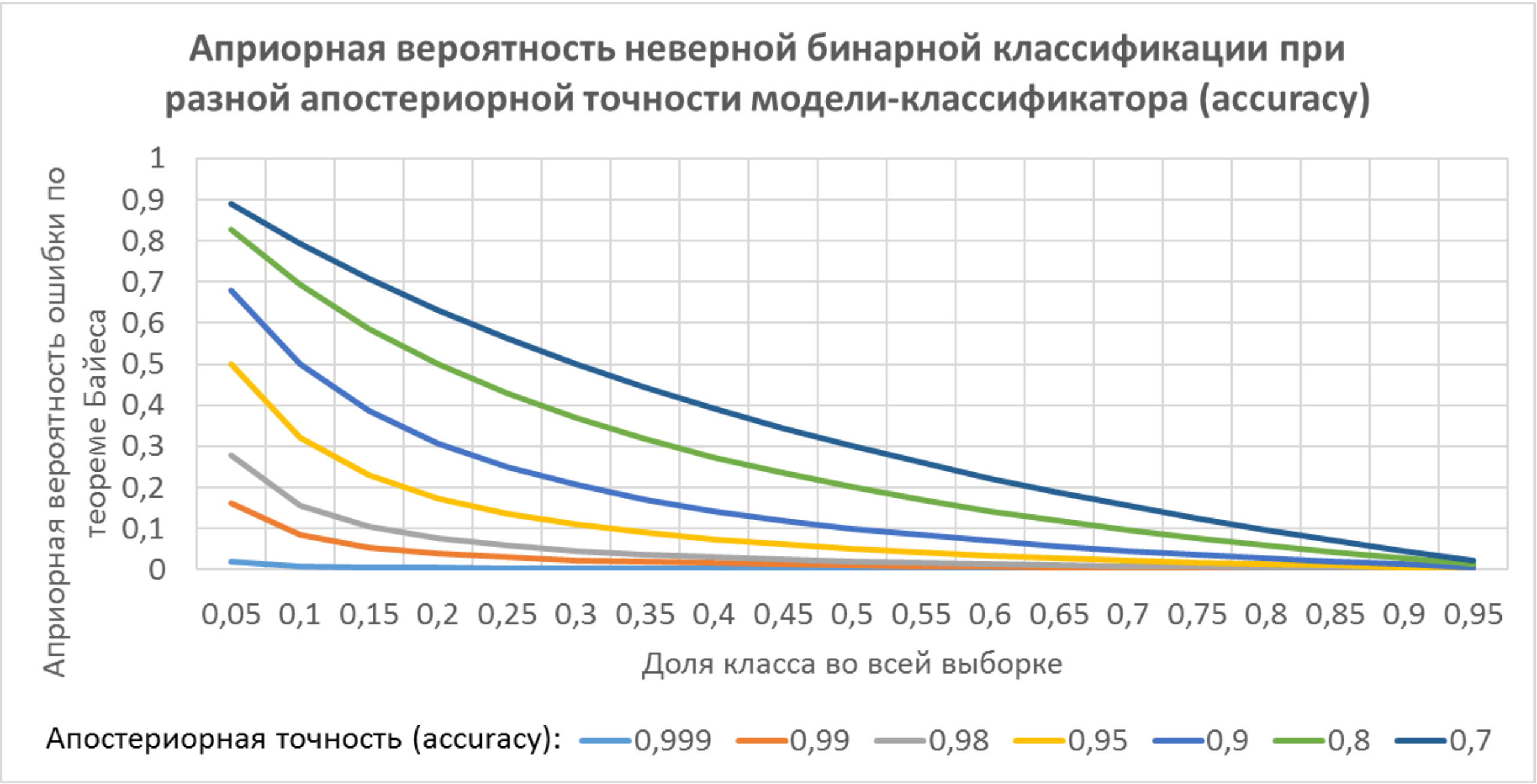


xgboost – один из вариантов здесь,
нейросети «deep learning» - другой вариант

http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/

вся Kaggle-подобная работа здесь

Сбор исторических данных	Выбор признаков и обучение моделей 	Автоматическое тестирование и синхронный выкат	Оптимизация ресурсов	Регулярный пересчёт признаков и переподбор моделей	Мониторинг качества и ресурсоёмкости
	Расчёт байесовской априорной вероятности				
	Расчёт признаков и работа моделей в runtime	Работа с экспериментальным кодом			
	A/B-тесты, баланс «польза/ресурсы»				


Априорная вероятность неверной бинарной классификации при разной апостериорной точности модели-классификатора (accuracy)
Априорная вероятность ошибки по теореме Байеса
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,05 0,1 0,15 0,2 0,25 0,3 0,35 0,4 0,45 0,5 0,55 0,6 0,65 0,7 0,75 0,8 0,85 0,9 0,95
Доля класса во всей выборке
Апостериорная точность (accuracy):
0,999
0,99
0,98
0,95
0,9
0,8
0,7

Кейс 2: Подбор вакансий по резюме

Задача

- Подобрать по резюме вакансии, интересные соискателю

Решение

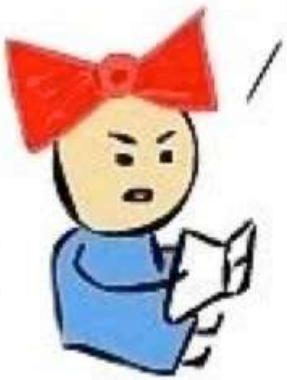
- Фильтр по параметрам вакансии
- Машинное обучение

Функционал

- Главная страница, рекомендуемые вакансии в списке резюме, рассылки с подходящими вакансиями

М-а-а-ш-и-и-н-н-о-е
о-б-у-ч-е-е-н-и-и-е

$$p(C \mid F_1, \ldots, F_n) = \frac{p(C)\, p(F_1, \ldots, F_n \mid C)}{p(F_1, \ldots, F_n)}.$$



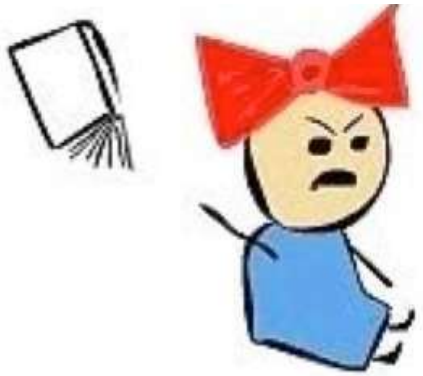
$$C = \sum C_{ij}, \quad C_{ij} = \log(1 + \exp((-1)^{[l_i > l_j]}(s_i - s_j)))$$

$$\frac{\partial C_{ij}}{\partial s_i} = -\frac{\partial C_{ij}}{\partial s_j} = -\frac{1}{1 + e^{s_i - s_j}} = -\rho_{ij}, \quad \frac{\partial C}{\partial s_i} = \sum_{j:l_i > l_j} (-\rho_{ij}) - \sum_{j:l_i < l_j} (-\rho_{ji})$$

$$\frac{\partial^2 C_{ij}}{\partial s_i^2} = \frac{\partial^2 C}{\partial s_j^2} = \rho_{ij}(1 - \rho_{ij}), \quad \frac{\partial^2 C}{\partial s_i^2} = \sum_j \rho_{ij}(1 - \rho_{ij})$$

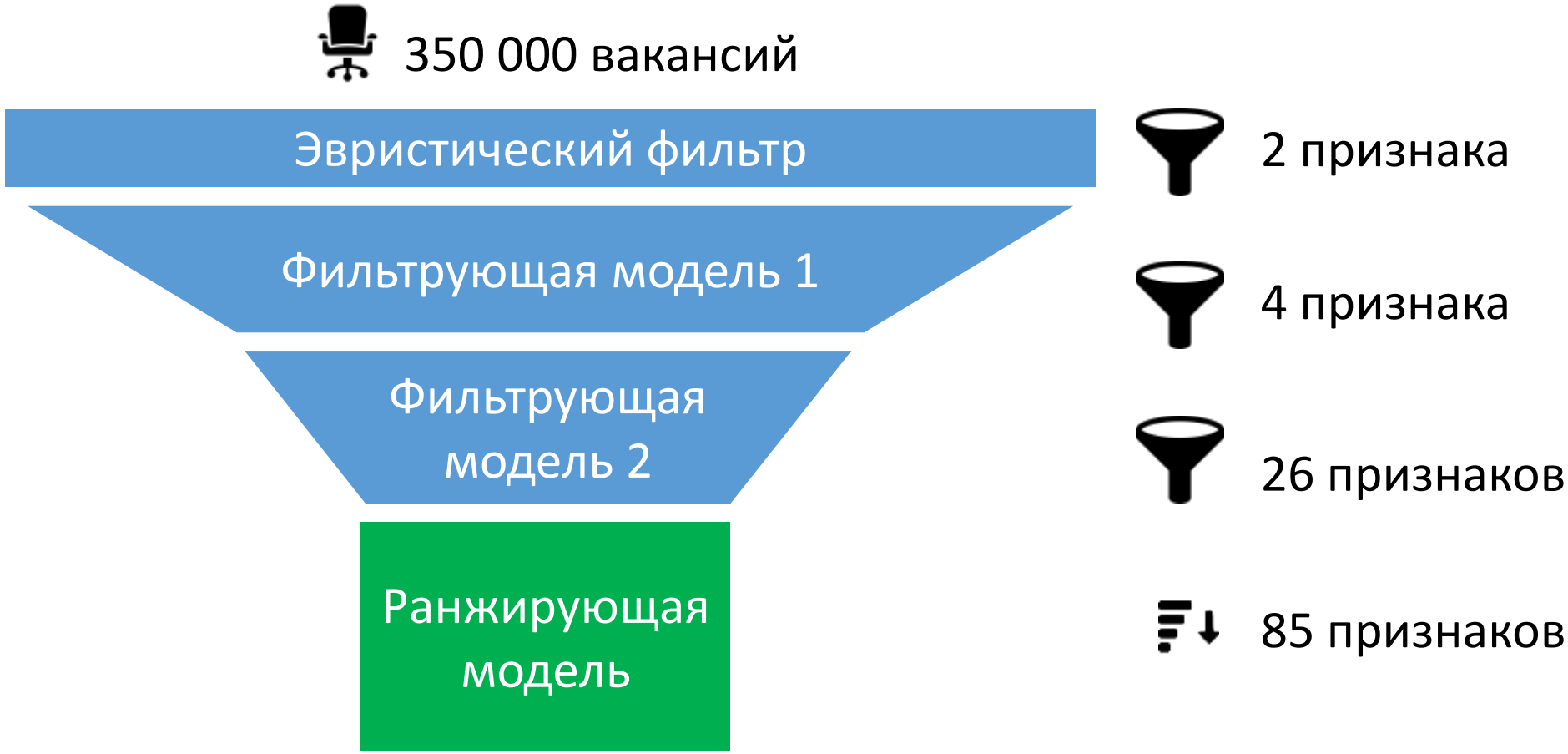
$$g_i = \sum_{j:l_i > l_j} |\Delta Z_{ij}|(-\rho_{ij}) - \sum_{j:l_i < l_j} |\Delta Z_{ij}|(-\rho_{ji}), \quad h_i = \sum_j |\Delta Z_{ij}| \rho_{ij}(1 - \rho_{ij})$$

Ну, нафиг…

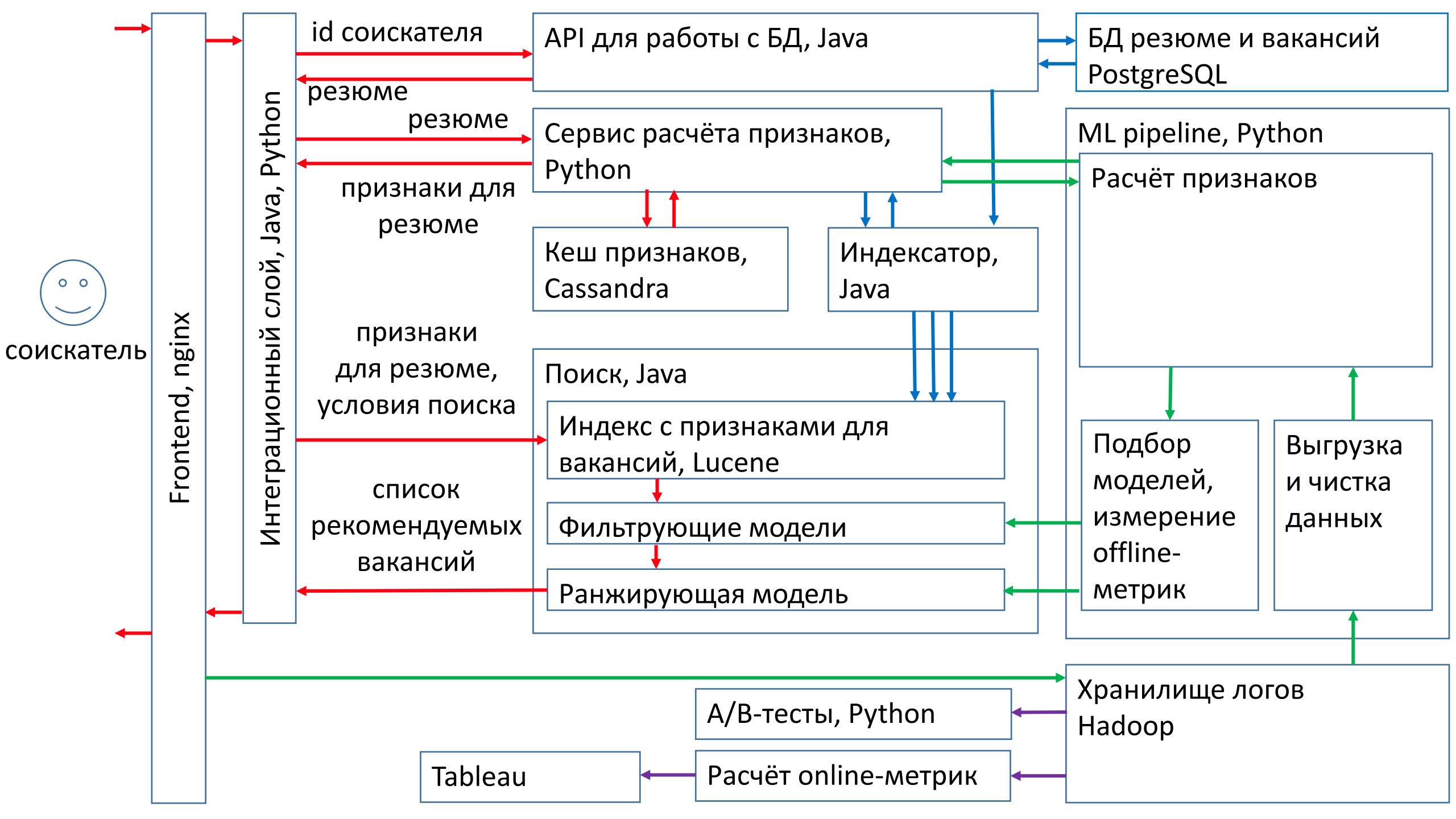


…буду экспертом
напишу формулу
вручную!

Схема работы рекомендательной системы



500-44m пар в обучающих множествах, переподбор за ночь


соискатель
Frontend, nginx
Интеграционный слой, Java, Python
id соискателя
резюме
резюме
признаки для резюме
признаки для резюме, условия поиска
список рекомендуемых вакансий
API для работы с БД, Java
БД резюме и вакансий PostgreSQL
Сервис расчёта признаков, Python
Кеш признаков, Cassandra
Индексатор, Java
Поиск, Java
Индекс с признаками для вакансий, Lucene
Фильтрующие модели
Ранжирующая модель
ML pipeline, Python
Расчёт признаков
Подбор моделей, измерение offline-метрик
Выгрузка и чистка данных
Хранилище логов Hadoop
A/B-тесты, Python
Tableau
Расчёт online-метрик

Результаты в откликах по результатам A/B-тестов

+ x1000 в сутки при внедрениях там, где не было

+ x100 в сутки при улучшениях моделей, факторов, таргетов

Иногда «просто» уменьшаем ресурсоёмкость

Кейс 3: Ранжирование откликов

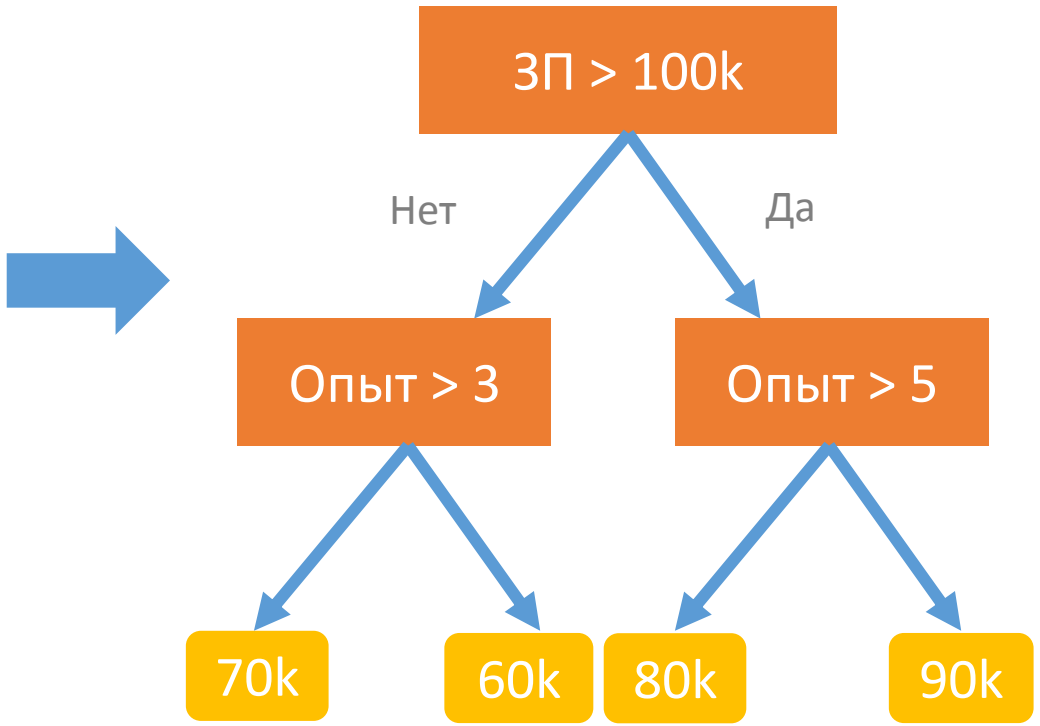
Задача	• На вакансию поступили отклики, необходимо понять кого приглашать на собеседование или телефонное интервью
Проблема	• HR-ры тратят значительное количество времени на разбор резюме и определение списка для собеседования
Функционал	• Кнопка «Лучшие» на странице откликов на вакансию

Человек: делает по инструкции, учится на небольшом опыте ошибается, субъективен, дорогое время

Неформализованный список:

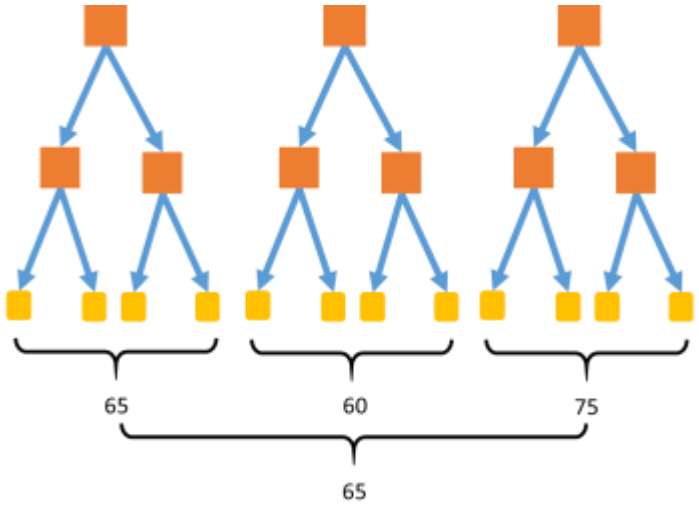
1. Опыт работы
2. Образование
3. Навыки
4. Зарплатные ожидания
5. …
6. …
7. …

Дерево решений



Машина: учится на большом опыте
тоже ошибается, объективна, дешёвое время

1. Исторические данные: кого пригласили на собеседование
2. Учим модель
3. Применяем модель к кандидатам
4. Улучшаем модель



Результат:

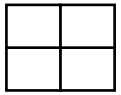
- 3,5m откликов в обучающем множестве
- 500 признаков
- 800k пар в день, 95%: 600ms

Кейс 4: поиск по соцсетям

Задача	• Найти соискателей, у которых нет резюме или у которых они закрыты
Решение	• Собрать все профили людей, сопоставить их между собой, выделить полезное для работы
Функционал	• Профили в поиске по резюме

Схема работы сопоставления профилей

365 млн. записей, 53.856 трлн. потенциальных пар



15 признаков

15 признаков

14k пар в обучающем множестве
точность: 99,9% - 150k пар, 80% - 4m
пересчёт за месяц

Рост использования машинного обучения в рекрутинге

Выводы и прогнозы

Использование машинного обучения дает хороший эффект в рекрутинге: в среднем целевые метрики вырастают на десятки процентов.

В краткосрочной перспективе

- вырастет эффективность подбора за счет экономии времени рекрутеров
- увеличится скорость поиска работы для соискателей

В среднесрочной перспективе

- подбирать сотрудников и искать работу будут ансамбли систем машинного обучения с уменьшающимся участием человека

17 лет меняем мир рекрутинга и HR к лучшему!