Аналитика вакансий для кадрового агентства

Март 2024

Описание проекта

- Источник данных:

Исследование проводилось для кадрового агентства, предоставившего данные по закрытию вакансий за период 2023 - начало 2024 года.

- Цель:

Выявить признаки вакансий, привлекающих наибольшее внимание кандидатов и успешно закрывающихся силами агентства. В рамках выполнения данной задачи можно использовать любые методы и средства, даже если они не описаны в ТЗ.

Описание данных

В формате Excel, данные текстовые и числовые, присутствуют пропуски

#	Column	Non-Null Count	Dtype
0	id	1284 non-null	int64
1	Статус	1284 non-null	object
2	Источник лида	1274 non-null	object
3	Менеджер	1193 non-null	object
4	Дата публикации	1284 non-null	object
5	Дата закрытия	1127 non-null	object
6	Количество просмотров	1284 non-null	int64
7	Количество откликов	1284 non-null	int64
8	Позиция	1284 non-null	object
9	Зарплата от	469 non-null	float64
10	Зарплата до	372 non-null	float64
11	Город	1284 non-null	object
12	Формат оформления	1284 non-null	object
13	Формат работы	1284 non-null	object
14	Опыт	1284 non-null	object
15	Образование	1284 non-null	object
16	Занятость	1284 non-null	object
17	Ссылка на тестовое	1284 non-null	object
18	Обязательные требования	1284 non-null	object
19	Дополнительный требования	659 non-null	object
20	Этапы отбора	1284 non-null	object
21	Условия	1284 non-null	object

Концепция исследования

- Исследование было **сосредоточено только на той части данных, которые могут влиять на реакцию кандидата на вакансию**. Фактически на подробностях того, какая именно публикация вакансии эффективна.
- Остальные признаки бизнес-процесса не включены в это исследование. Часть, которая осталась за скобками несомненно очень важная. Эти параметры могут очень существенно влиять на скорость закрытия вакансии, но они заслуживают отдельного рассмотрения.
- Основной метрикой будет **конверсия просмотров в отклики**. Рассчитывается как **отклики/просмотры**, может быть от **0 до 1.**

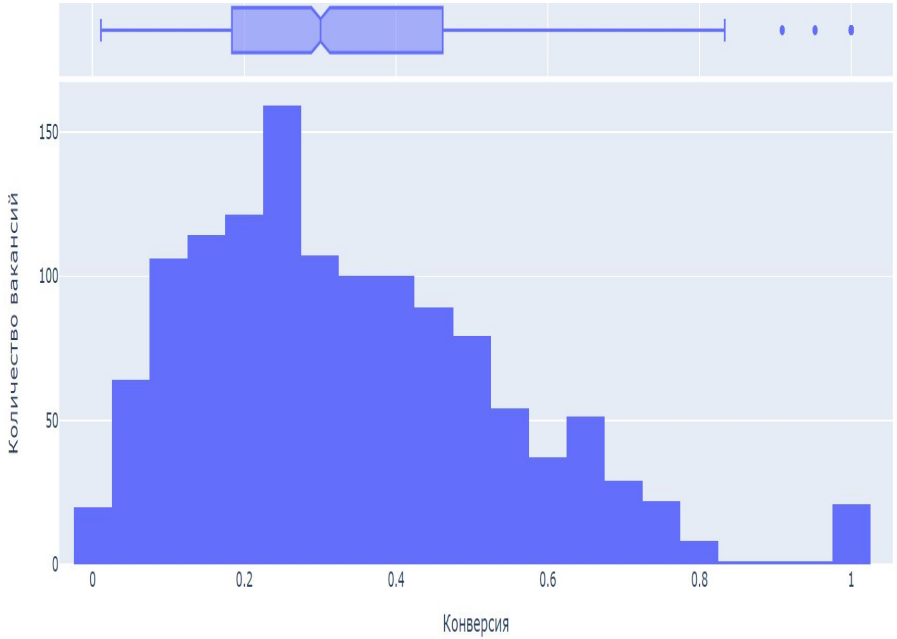
Наблюдаемые возможности на основе данных

В качестве основной цели исследования:
изучение закономерностей реакций кандидатов на публикацию вакансии

Параметры вакансии, которые видит кандидат:	**Другие признаки бизнес-процесса, которые кандидат не видит:**
Дата публикации *	Статус
Зарплата от *	Источник лида
Зарплата до *	Менеджер
Город *	Дата закрытия
Формат оформления *	Количество просмотров *
Формат работы *	Количество откликов *
Опыт *	Этапы отбора
Образование *	
Занятость *	
Ссылка на тестовое	
Обязательные требования	
Дополнительный требования	
Условия	
***вероятные параметры поискового запроса кандидата**	***первичные метрики активности кандидата**

Основная метрика

Рассчитана метрика конверсии отклики/просмотры:
1 - максимально возможная конверсия, 0 - минимально возможная конверсия

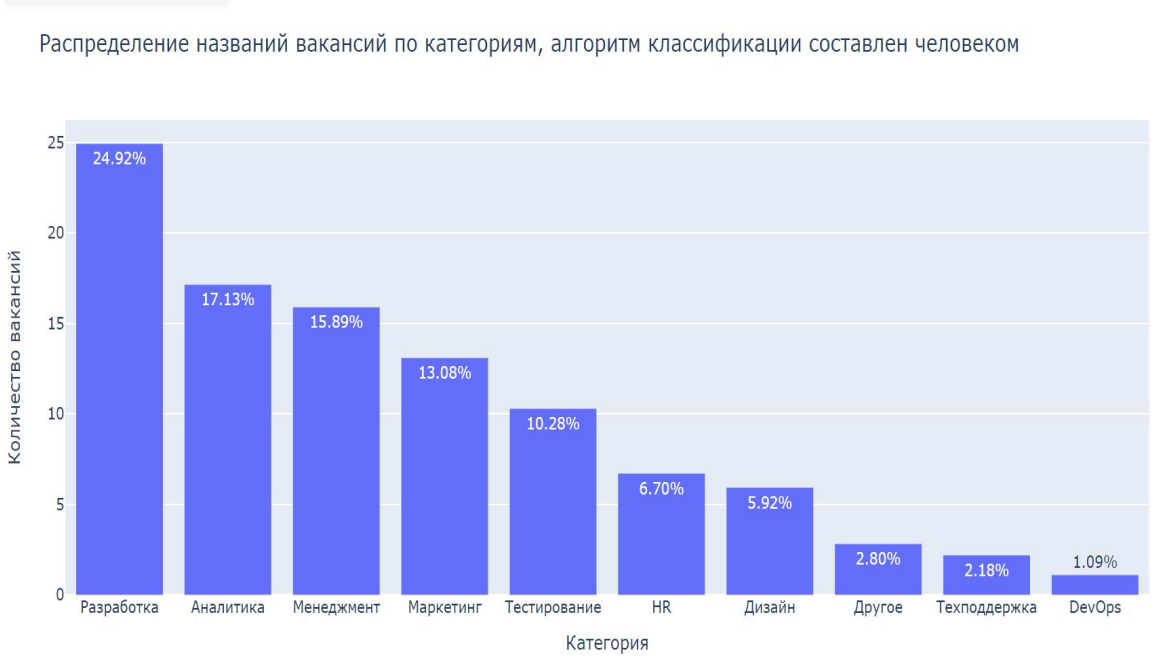


Анализ рассчитанного признака показывает нормальное распределение, это значит он будет хорошо работать как статистический критерий

Средние показатели конверсии по всем вакансиям примерно **0.3**

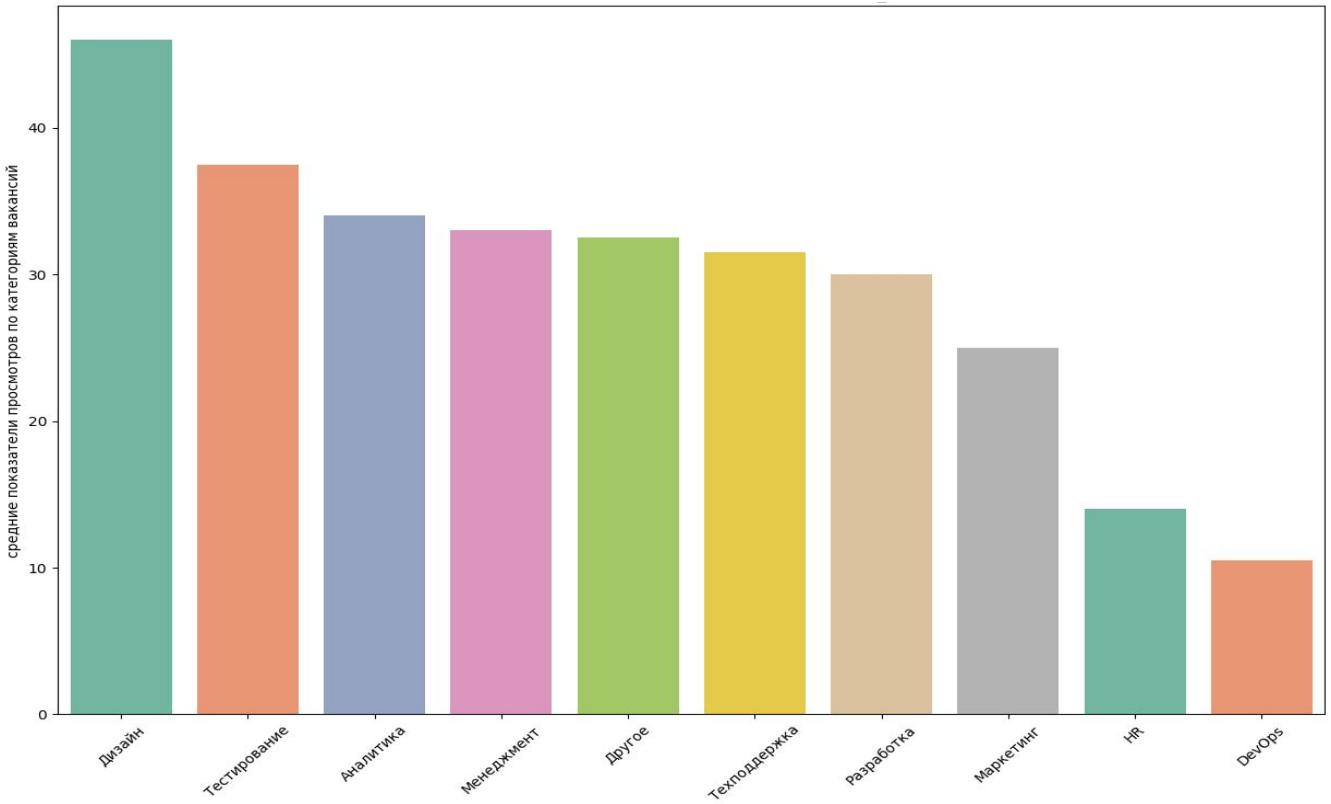
Разделение вакансий на категории

Названия позиций разделены на категории на основании экспертной оценки



Вакансии разделены по профессиональным областям (по признаку названия). Видна сравнительная статистика количества вакансий по этим категориям.

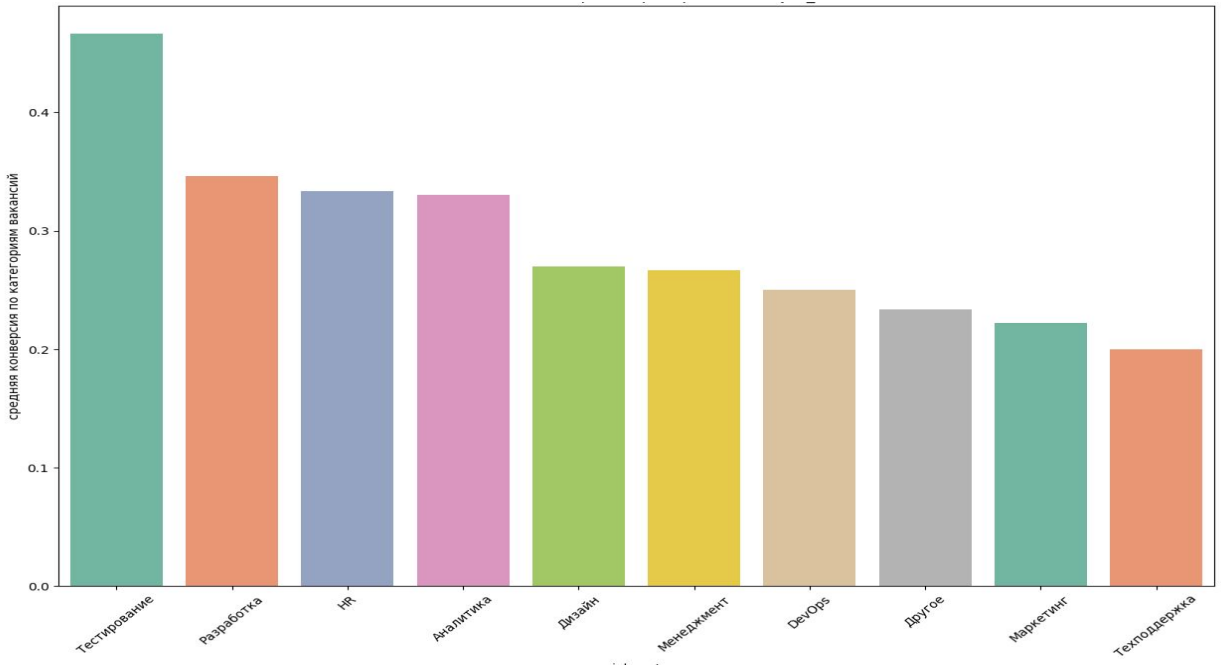
Средние показатели просмотров по категориям



Самые высокие средние показатели просмотра по категориям вакансий - в категории "Дизайн". Присутствие этой категории по количеству вакансий - менее 6%. Очевиден дефицит предложения в этой категории.

Средние показатели конверсии по категориям вакансий

Наблюдаются существенные отличия в зависимости от категории вакансии



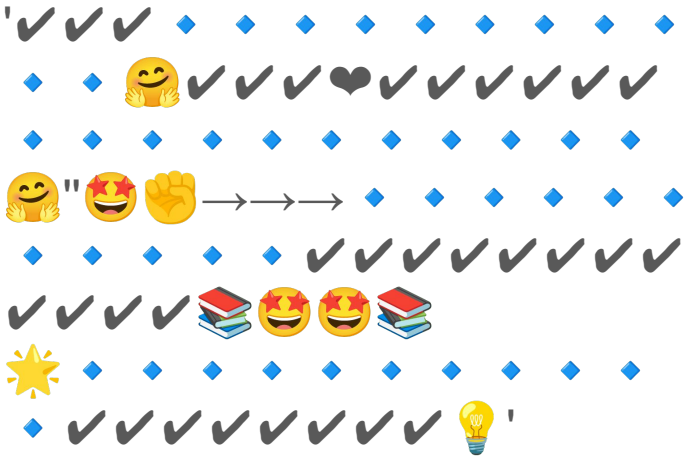
Тестировщики откликаются не прочитанное почти в половине случаев, техподдержка - почти в 2 раза реже.

“Мешок слов” и Wordcloud для условий работы

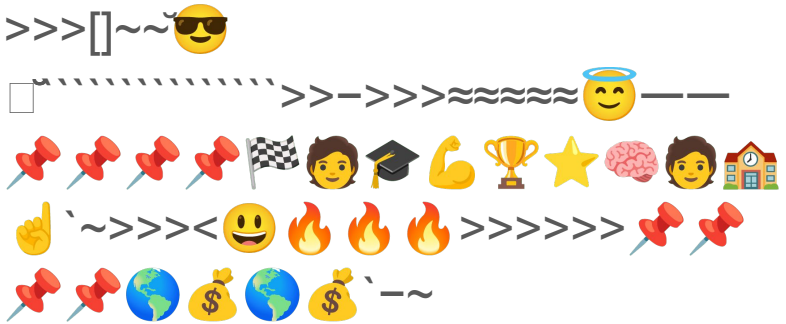
- **"Мешок слов"** - упрощенное представление текста, в котором важно только количество (частота присутствия) слов, но их порядок в предложениях не может быть учтен.
- В качестве инструмента визуализации будем использовать **Wordcloud**
- Выделяем группы объявлений с высокой и низкой конверсией. Между ними больше сходств, чем отличий. Уберем самое общее, будем искать специфичное.

Уникальные признаки текстов

Встречаются только в объявлениях с хорошей конверсией



Встречаются только в объявлениях с плохой конверсией



Именно отдельных слов в качестве исключительных признаков не обнаружено, каждое слово может быть в любой группе, но отличается частота и контекст.

Влияние текста описания условий работы на отклики

Признаки в группе с высокой конверсией

Видно существенное положительное влияние “ТК РФ” в условиях работы

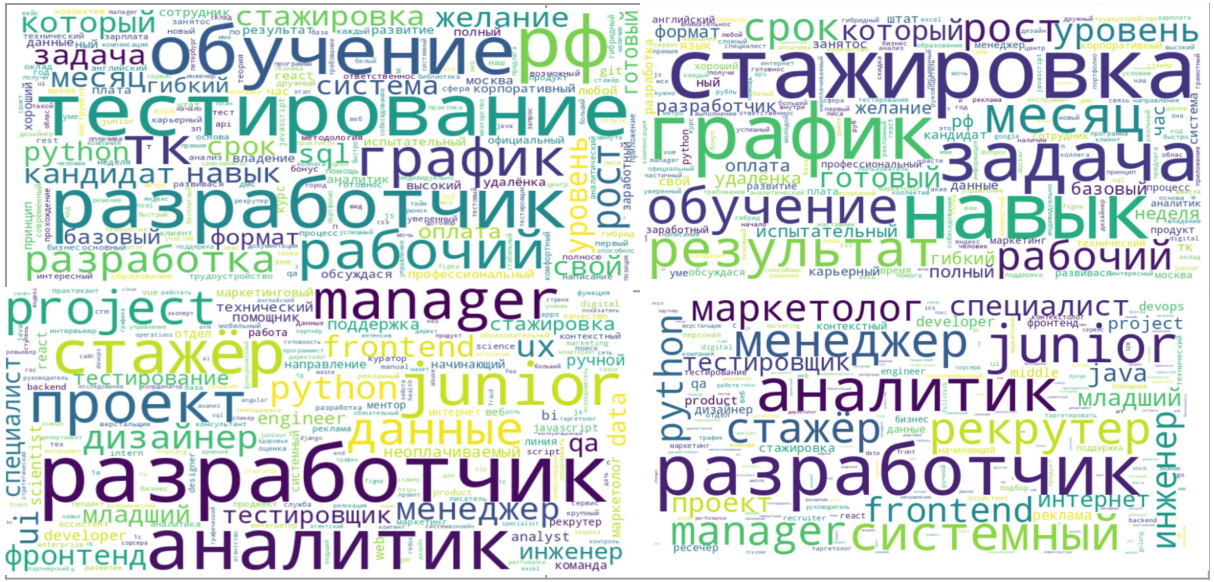
Влияние текста описания условий работы на отклики

Признаки в группе с низкой конверсией

График, проект и стажировка - признаки, снижающие конверсию откликов.

Влияние других текстовых признаков на отклик

Также было исследовано влияние других текстовых признаков: всех полей описания вакансии на конверсию, также влияние названия позиции на количество просмотров

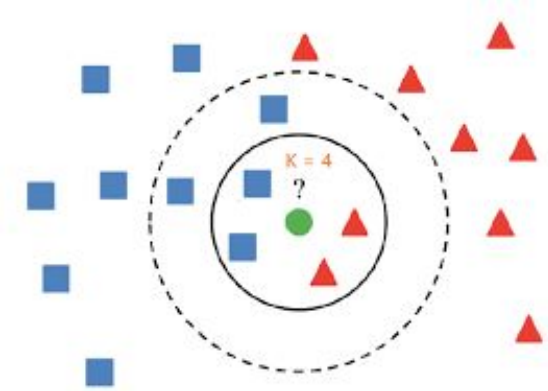


Слева: текстовые признаки для высокой конверсии и большого количества просмотров, справа: для низких значений.

Верхний ряд: все тексты вакансии, влияние на конверсию
Нижний ряд: влияние названия вакансии на просмотры

Метод KNN (k Nearest Neighbor)

- Попробуем сравнить классификаторы вакансий предложенных человеком и ИИ.
- Будем искать группы похожих названий с помощью машинного обучения: поиск ближайших соседей в многомерном пространстве признаков.

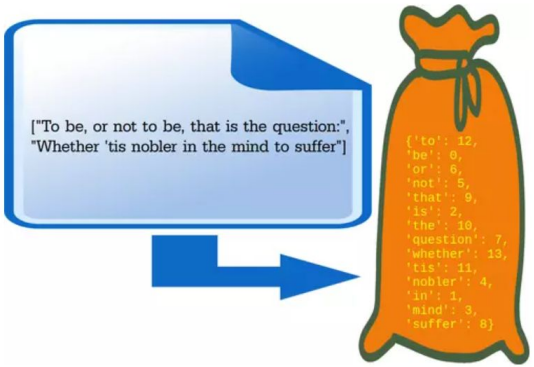


Метод KNN, 22 кластера

Метод KNN, 10 кластеров

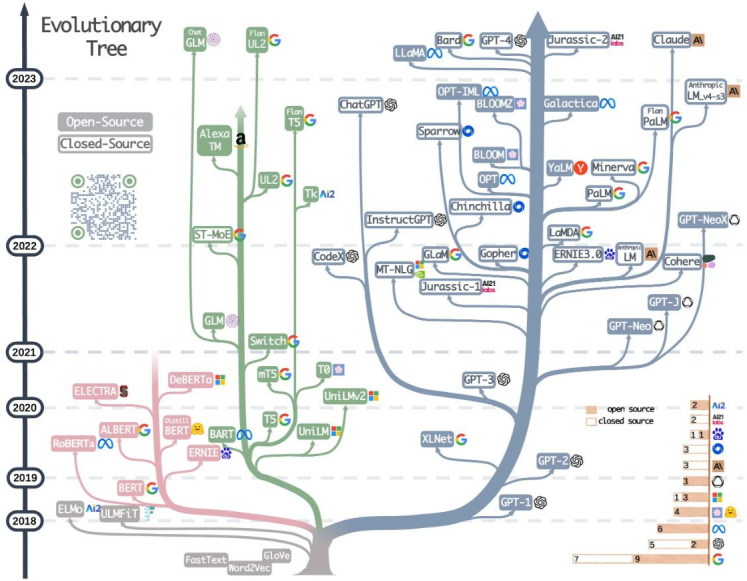
Выводы по применению ‘Мешка слов’

- Дает результаты для высвечивания статистически значимых слов, хорошо помогает визуализировать данные с помощью Wordcloud
- Для практических целей годится, но по сравнению с теми алгоритмами классификации, которые пишет человек, может быть недостаточно точен, потому что не учитывает связи между словами.



Трансформеры, LLM, большие языковые модели

- **Могут учитывать не только набор слов, но и их порядок и связи**.
- Работают с огромным количеством данных и требовательны с вычислительным ресурсам, поэтому их появление стало возможным только недавно (примерно с 2017 года).
- Активно разворачиваются в системах автоматического перевода, но человечество их заметило недавно благодаря чату GPT.
- Эволюция их развития на рисунке.
- Существует много open-sourse решений.



Дерево развития LLM

Алгоритм работы

- LLM преобразует все текстовые признаки вакансий в матрицы чисел (embedding)
- Поиск закономерностей в полученных данных проводится с помощью различных моделей машинного обучения.
- Используем метрику RMSE.
- Выбираем самую успешную по точности предсказаний модель.

$$RMSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(y_i - f_i)^2}$$

BERT (Bidirectional Encoder Representations from Transformers)

Hugging Face

BERT “понимает” связи между словами и преобразует текстовые поля

Все **произвольные текстовые поля вакансии**:

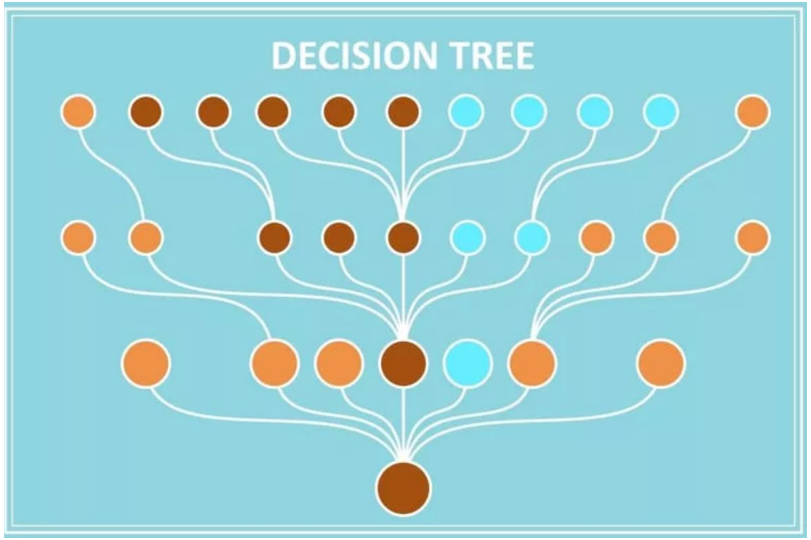
- Название
- Обязательные требования
- Дополнительные требования
- Условия работы

Embeddings, числовые представления слов, с учетом их порядка

5	6	7	8	9	...	1014	1015	
38	-0.563705	0.307822	0.596808	-1.534599	...	-0.101405	1.682038	(
37	-0.633104	0.118598	0.616011	-1.284846	...	0.317867	1.660743	(
55	-0.623898	-0.077078	0.544618	-0.444710	...	0.224897	1.368433	-(
70	-0.554586	0.138702	0.456975	-1.261138	...	-0.209950	1.832704	-(
39	-0.339269	0.025940	0.537210	-0.376054	...	-0.510443	1.706598	-(

“Деревья решений, лес решений”

В библиотеках машинного обучения множество моделей, которые могут работать с таким количеством данных и находить закономерности



- Сложный разветвленный алгоритм работы
- Находят оптимальный вариант среди множества
- Лучшая метрика **RMSE 0.20** (чем меньше, тем лучше)

Catboost - может работать с категориями

Yandex
CatBoost

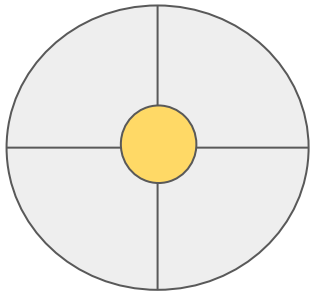
Embeddings

+

Категории из датасета:

- Город
- Формат оформление
- Формат работы
- Опыт
- Образование
- Занятость
- Наличие тестового

- Появляется возможность расширить данные для обучения модели
- Модель стала сильнее
- Лучшая метрика **RMSE 0.17** (чем меньше, тем лучше)



Результаты работы

Модель машинного обучения, предсказывающая конверсию просмотров в отклики на основании текстовых и категориальных признаков в объявлении о вакансии.