

Хакатон Гринатом

15 октября
2019

19 ноября
2019



ГРИНАТОМ

Phystech
.Genesis

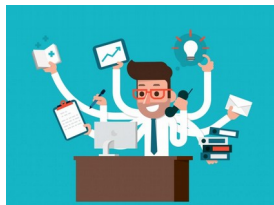
Трек: Машинное обучение и социальные графы

Команда: Polosataya

Некрасова Татьяна Валерьевна

Стоимость увольнения (проблема):

- прямые затраты (затраты на увольнение, на поиск и найм, на вхождение в должность)
- неявные затраты (снижение производительности труда, снижение лояльности клиентов, репутационные риски, снижение квалификации, недополучение прибыли в отсутствии сотрудника)



Вероятность увольнения



Задача:

Предсказание увольнения сотрудников на основе текстовой информации



Детали UI:

Адрес <http://polosataya.pythonanywhere.com/>

Роли: сотрудник, руководитель, админ.

Функционал:

Сотрудник: заполнение веб-форм, получение личного прогноза

Руководитель, hr: заполнение веб-формы, загрузка таблиц с данными о нескольких сотрудниках и получение прогнозов, доступ к социальному графу сотрудников

Админ: обновление модели, доступ к логам

Демо: вкладка для сотрудника

[Для сотрудников](#) [Для руководителей](#)

Хотите оценить, насколько сильно ваше желание уволиться?

Должность

Офисный сотрудник

Сколько лет вы проработали

<1

1-2

3-5

6-10

>10

Уровень оплаты труда

1

2

3

4

5

Начальство

1

2

3

4

5

Рабочее место

1

2

3

4

5

Коллектив

1

2

3

4

5

Карьерный рост

1

2

3

4

5

Узнать

Опишите плюсы работы:

Хороший коллектив, в нашем отделе все помогают друг другу, не шепчутся за спиной. Компания престижная.

Что нужно улучшить:

Непосредственное начальство может наорать просто так, приписывает себе все заслуги работников. Зарплата просто смешная. Остаться во время проверок до 10 вечера без доп оплаты это норма.

Вероятность вашего увольнения: 40.51%

Демо: вкладка для руководителя

Для сотрудников

Для руководителей

Загрузить CSV файл

Выберите файл test_data.csv

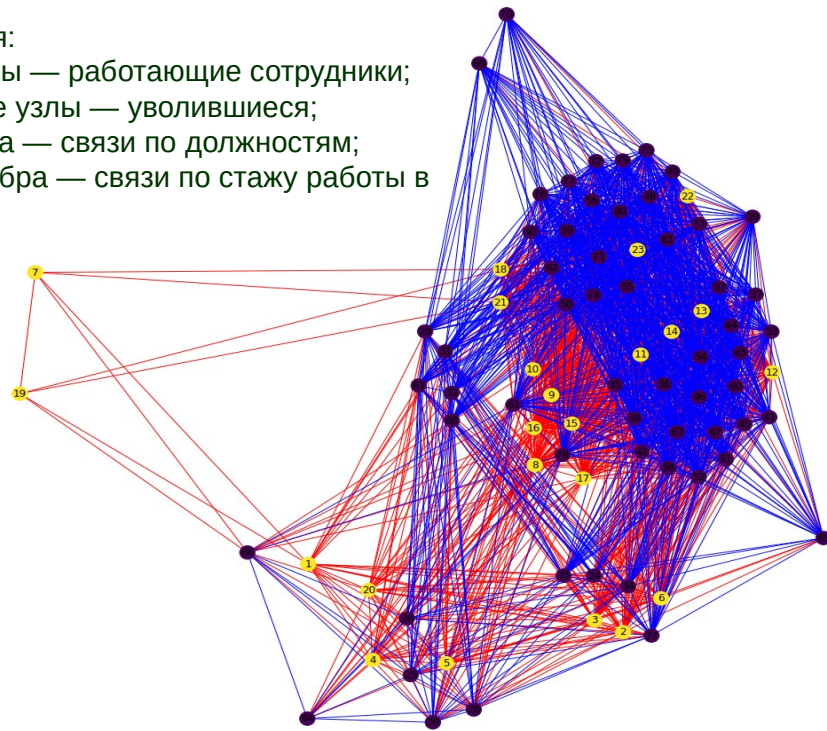
Загрузить файл

Результат:

1, "49.04%"
2, "65.15%"
3, "45.19%"
4, "40.41%"
5, "64.68%"

Обозначения:

- желтые узлы — работающие сотрудники;
- коричневые узлы — уволившиеся;
- синие ребра — связи по должностям;
- красные ребра — связи по стажу работы в компании.



Детали Tech:

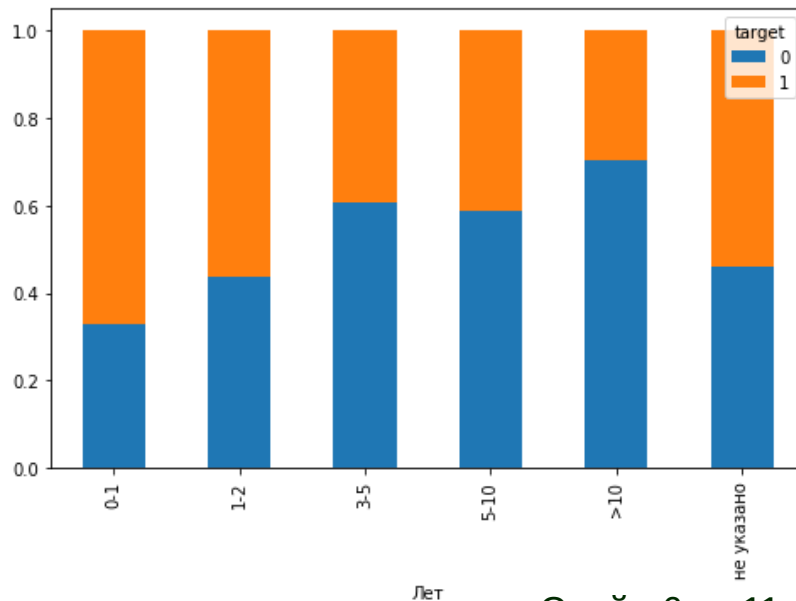
Анализ текста: оценка важности слов в контексте отзыва сотрудника о работе

Категориальные фичи: должность, стаж, оплата труда, начальство, рабочее место, коллектив, карьера

Числовые фичи: год, количество слов

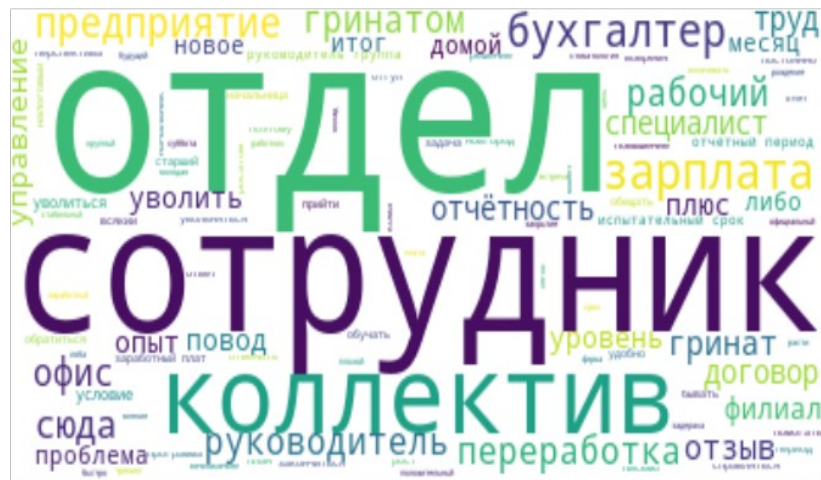
Потенциальные фичи: отрасль, город,
выделение более 10 должностей

Модель: CatBoostClassifier (точнее)
или LGBMClassifier (быстрее)

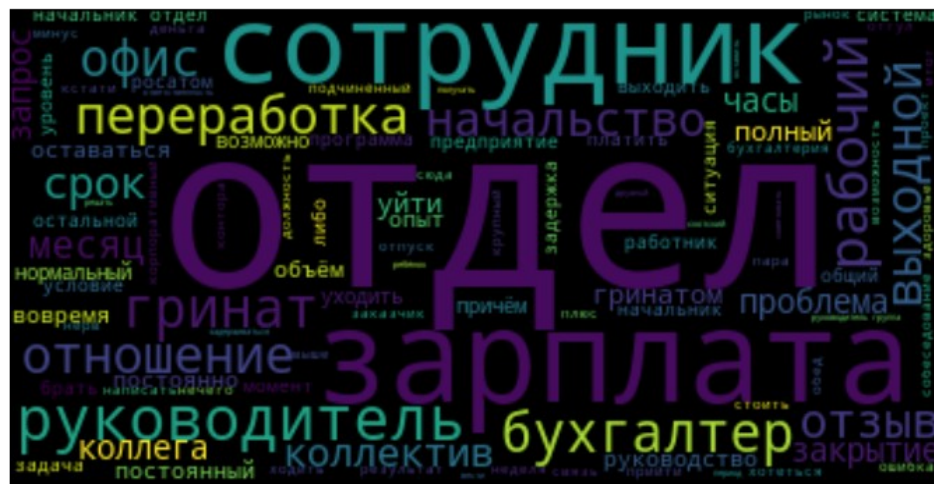


Киллер фича — text mining

Отзывы работающих сотрудников



Отзывы уволившихся сотрудников



To sum up + Business

Что сделано:

- сбор датасета, обучение модели,
- разработка демоверсии приложения.



Почему это нужно компании/пользователю:

- получение информации о вероятности увольнения сотрудников заранее;
- выявление причин увольнения;
- сокращение финансовых и репутационных потерь компании.

Оценка модели:

2000 наблюдений - ROC AUC 0,73-0,74

(без текстовых фич 0,67; только с текстовыми фичами 0,69)



Стек и команда

Моделирование: TfidfVectorizer, CatBoostClassifier, LGBMClassifier

Социальные графы: graphviz, networkx

Демоверсия приложения: python, flask

Пути улучшения: nltk, eli5

Команда: все роли — polosataya

Направления улучшения:

- увеличение датасета до 5 000 наблюдений (увеличение словарного запаса)
- добавление в выборку работающих сотрудников
- добавление функциональности и улучшение дизайна приложения (ведение логов для пополнения датасета, добавление новых параметров)
- для графа добавление интерактивности
- усложнение языковой модели (вместо мешка слов полноценные эмбединги и т.д.).

Спасибо за внимание

Материалы:

[https://drive.google.com/drive/folders/
1gavBHfWBdWFyXTM2HD1APVxhrlsok-U-?usp=sharing](https://drive.google.com/drive/folders/1gavBHfWBdWFyXTM2HD1APVxhrlsok-U-?usp=sharing)

Демоверсия:

<http://polosataya.pythonanywhere.com/>

E-mail:

gntv1977@gmail.com

