



ELTE

Лучшее решение на коде


Лучшая концепция решения

Кейс 2. Система определения предлагаемой заработной платы

Виталий Наумов
+79256014973



Этапы

- Сбор данных
 - Обработка данных
 - Построение модели
 - Обучение модели
 - Оценка данных
 - Интеграция с веб версией
- 

Сбор данных

- Транслитерация на английский
- Формирование ссылок на вакансии
- BeautifulSoup + анализ html кода
- Запись в CSV

In [94]: frame

Out[94]:

	experience	schedular	description	key_skills	url	prof	salary_n
0	1–3 года	Полная занятость, полный день	Любишь моду, стильных людей и авторский почерк...	[помощник менеджера по продажам, Организаторск...	https://hh.ru/vacancy/37128903	administrator	30000
2	не требуется	Полная занятость, сменный график	Хочешь интересную и стабильно оплачиваемую раб...	[Работа в команде, Креативность, Грамотная реч...	https://hh.ru/vacancy/37879197	barista	25000
3	не требуется	Полная занятость, полный день	Приглашаются в команду-бармены Требуемый о...	[Пользователь ПК, Барное ремесло, Работа в ком...	https://hh.ru/vacancy/37774700	barmen	50000

Обработка данных

Определение и преобразование категориальных данных

- Тип занятости
- Специализация
- Опыт работы

Выявление ключевых навыков в некатегориальных данных

- Лемматизация
- Группировка

```
def change_employment(cdf):
    cdf['employment_num'] = (cdf['employment'] == 'full').apply(to_int)
    return cdf.drop('employment', axis = 1)

def change_schedule(cdf):
    type_schedule = list(cdf.groupby('schedule').count().index)
    for typ in type_schedule:
        cdf[typ + '_schedule'] = (cdf['schedule'] == typ).apply(to_int)
    cdf = cdf.drop('schedule', axis = 1)
    return cdf

def delete_duplicates(x):
    return list(dict.fromkeys(x))

def spec_to_int(string):
    spec = string.split()
    res = []
    for num in spec:
        res.append(int(float(num)))
    return delete_duplicates(res)

def add_spec(cdf):
    cdf['spec'] = cdf['specializations'].apply(spec_to_int)
    cdf['spec'] = cdf['spec'].apply(lambda x: x + [0]*6)
    for i in range(6):
        cdf['spec_' + str(i)] = cdf['spec'].apply(lambda x: x[i])
    return cdf

def nor_experience(s):
    if s == 'between1And3':
        return 2
    if s == 'between3And6':
        return 4
    if s == 'moreThan6':
        return 8
    return 0
```

Построение модели

- CatBoost

```
[ ] model= CatBoostRegressor(iterations=1000, learning_rate=0.05, bootstrap_type='MVS', depth=16, leaf_estimation_method='Newton', score_function='NewtonL2',  
                             od_type='IncToDec', one_hot_max_size=4, l2_leaf_reg=3, random_seed=1)
```

- word2vec

```
In [84]: df[finale_skills]
```

```
Out[84]:
```


	creativity_num	hard_work_num	accuracy_num	leader_num	computer_num	mind_tech_num	communication_num	language_num
0	1	1	0	1	1	0	0	0
1	1	1	1	1	1	1	1	1
2	0	0	0	0	1	0	0	0
3	1	1	0	1	0	0	1	1

Обучение модели

```
[ ] model.fit(full_train_pool, eval_set=(X_test,y_test), verbose=100)  
model.score(X_test,y_test)
```

```
0:      learn: 21401.5773737      test: 21464.5682504      best: 21464.5682504 (0) total: 1.47s      remaining: 24m 27s  
100:    learn: 16198.3283763      test: 16224.4991762      best: 16224.3724190 (99)      total: 3m 26s      remaining: 30m 35s  
200:    learn: 15891.5661998      test: 16133.2731655      best: 16133.2731655 (200)      total: 7m 43s      remaining: 30m 42s
```

Оценка данных



	Original	Predicted
index		
83091	45000.0	45000.0
30447	35000.0	36000.0
57575	27000.0	25000.0
63928	50000.0	51000.0
109981	60000.0	55000.0

Интеграция с веб-версией

Activities Google Chrome Sat Jul 11, 01:20:14

Document UN2020 - Jupyter Notebook

Not secure 192.168.0.115

Хакатон труда

Форма заявки

Название вакансии	Название компании	
Регион	Город	
Тип занятости	Тип графика	Опыт
Ключевые навыки	Специализация	
Описание		
Расчитать зарплату		

Пути развития



Внедрение в существующую
систему по подбору
персонала



Использовать как
самостоятельный продукт

Необходимые ресурсы

- Больше количество данных для обучения
- Сервера для быстрого обучения и сбора данных

Наша команда



Золотарева Екатерина
Data Science – 1 год
Web-development – 3 года



Наумов Виталий
Студент ELTE, Будапешт
Опыт в Data Science 1 год
Сфера интересов NLP,
Deep Learning.



Владимир Соловьев
Студент 3 курса ВШЭ
Факультет экономических
наук

The background features two large, decorative, curved lines. One line starts from the top right and curves downwards towards the center. The other line starts from the bottom left and curves upwards towards the center. Both lines are composed of multiple overlapping layers in shades of light blue and light green, creating a sense of depth and movement.

Спасибо за внимание!