analytics-notebook

8 июня 2019 г.

1 Анализ вакансий в канале #_jobs в слаке Open Data Science

1.1 Вступление

Рынок труда в области Data Science стремительно рос последние 5 лет. Вместе с тем, как все больше компаний используют у себя анализ больших данных, все больше появляется курсов для подготовки специалистов по анализу данных и, соответственно, их слушателей.

Поэтому интересно было, как со временем менялась зарплата и какие есть тренды. Вакансии можно посмотреть, например, в канале #_jobs в слаке сообщества Open Data Science. Это и будет набором данных для анализа.

1.2 Инструменты анализа

Данные изначально были представлены в виде постов в канале Слак. Я использовал утилиту ParseHub и библиотеки Питона `BeautifulSoup` и `Selenium` для того, чтобы спарсить все посты. На выходе получался Pandas датафрейм с колонками ссылок на сообщение, автора, текста сообщения и даты и времени сообщения. Также был проведен предварительный анализ сообщений: все сообщения, не содержащие цифры, были убраны из рассмотрения из соображения, что пост с вакансией имеет цифровую запись предлагаемой зарплаты.

Получилось порядка 2400 сообщений, которые нужно было проверить на то, что они действительно являются вакансиями и содержат вилку. Для этого был предпринят следующий подход: 1. разметка сообщений по ключевым словам; 1. обучение алгоритма машинного обучения по размеченным данным для определения того, является ли сообщение вакансией и является ли строка сообщения строкой с вилкой; 1. применение алгоритма ко всем данным.

По ключевым сообщениям были определены 797 сообщений, затем после ручной проверки была составлена обучающая выборка из 100 постов с вакансиями и 100 постов без вакансий. В качестве алгоритма машинного обучения был использован `CatBoostClassifier` из библиотеки `catboost`. На вход алгоритм получал датафрейм из 7 признаков сообщений (длина сообщения, количество слов, количество строк, количество цифр, количество слов длины 1, количество слов длины 2, количество специальных символов), и он должен определять вероятность события, является ли данное сообщение вакансией или нет. Целевой функцией была кросс-энтропия, в обучении использовалась кросс-валидация. Затем аналогичный алгоритм был составлен для получения вероятности события нахождения вилки в строке с небольшим изменением входных параметров (количество строк заменялось на номер строки и добавлялась вероятность сообщения быть вакансией). При обучении обоих алгоритмов при стандартных параметрах получились значения метрики `roc_auc` 0.92.

При применении алгоритма количество сообщений увеличилось до 929, и из них были выделены значения вилок. К сожалению, всех их пока не представляется возможным проанализи-

ровать, так как это вакансии в разных валютах и с удельной зарплатой за разные промежутки времени. Поэтому эти 929 сообщений были отфильтрованы по тому диапазону, где находится вилка, и по запрещенным ключевым словам, чтобы получить сообщения с вакансиями с зарплатой в рублях за месяц (534 сообщения).

1.3 Результаты анализа

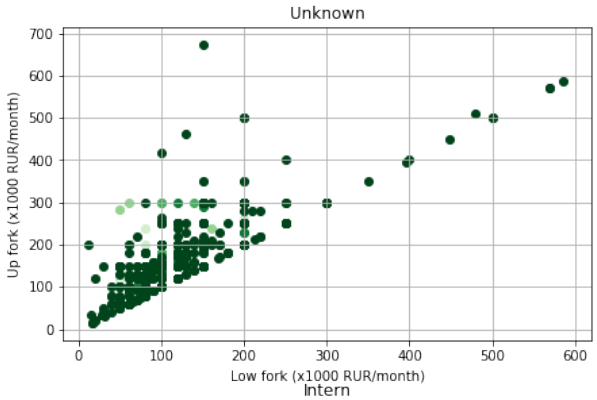
Итак, были рассмотрены 534 сообщения с вакансиями с апреля 2015 года по май 2019 года, а именно были посчитаны графики минимальных, максимальных и средних значений за месяц, и количество вакансий в месяц.

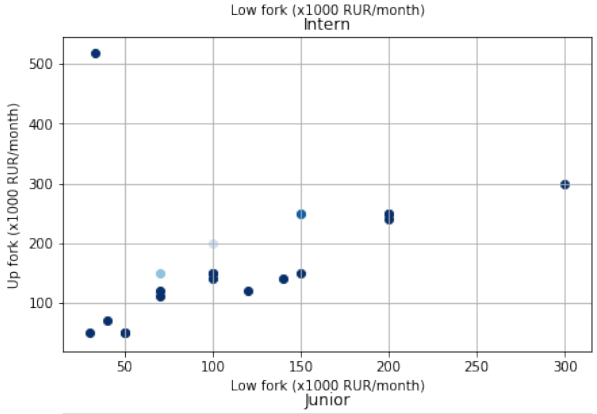
Количество вакансий постепенно растет последние 2 года и составляет порядка 20 вакансий в месяц. Пока непонятно, для каких позиций преоладают вакансии.

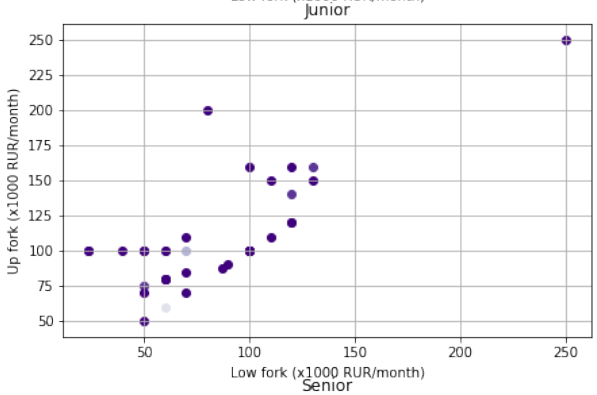
Средняя вилка для джуниоров составляет 50-120 тысяч рублей в месяц, для мидла — 100-170 т.р., для сеньора — 150-330 т.р. Средняя вилка вообще равна 145-170 т.р.

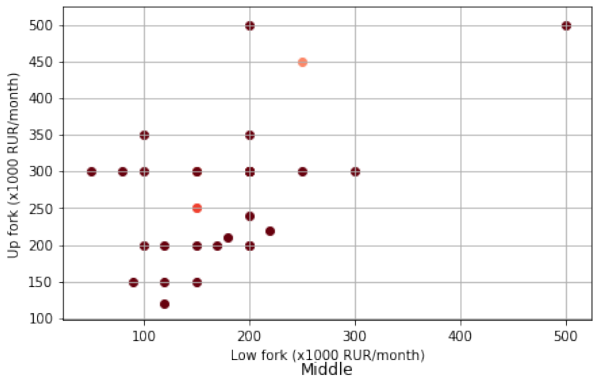
Также были построены по данным линейные тренд средней зарплаты. Коэффициенты в трендах составляют -1.6 т.р./год для джуниоров, 8.4 т.р./год для мидлов, 6.6 т.р./год для сеньоров и в среднем 6.7 т.р./год.

Рис. 1 Карта вилок (чем ярче, тем позже опубликована вакансия)


Unknown
Up fork (x1000 RUR/month)
Low fork (x1000 RUR/month)
0
100
200
300
400
500
600
700


Intern
Up fork (x1000 RUR/month)
Low fork (x1000 RUR/month)
50
100
150
200
250
300
400
500


Junior
Up fork (x1000 RUR/month)
Low fork (x1000 RUR/month)
50
75
100
125
150
175
200
225
250


Senior
Up fork (x1000 RUR/month)
Low fork (x1000 RUR/month)
100
150
200
250
300
350
400
450
500

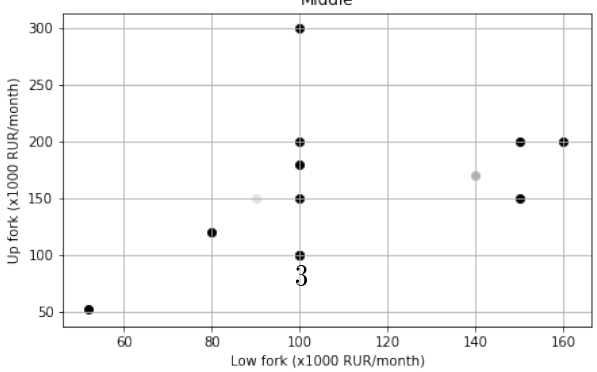

Middle
Up fork (x1000 RUR/month)
Low fork (x1000 RUR/month)
50
100
150
200
250
300
60
80
100
120
140
160

Рис. 2 Количество вакансий в месяц

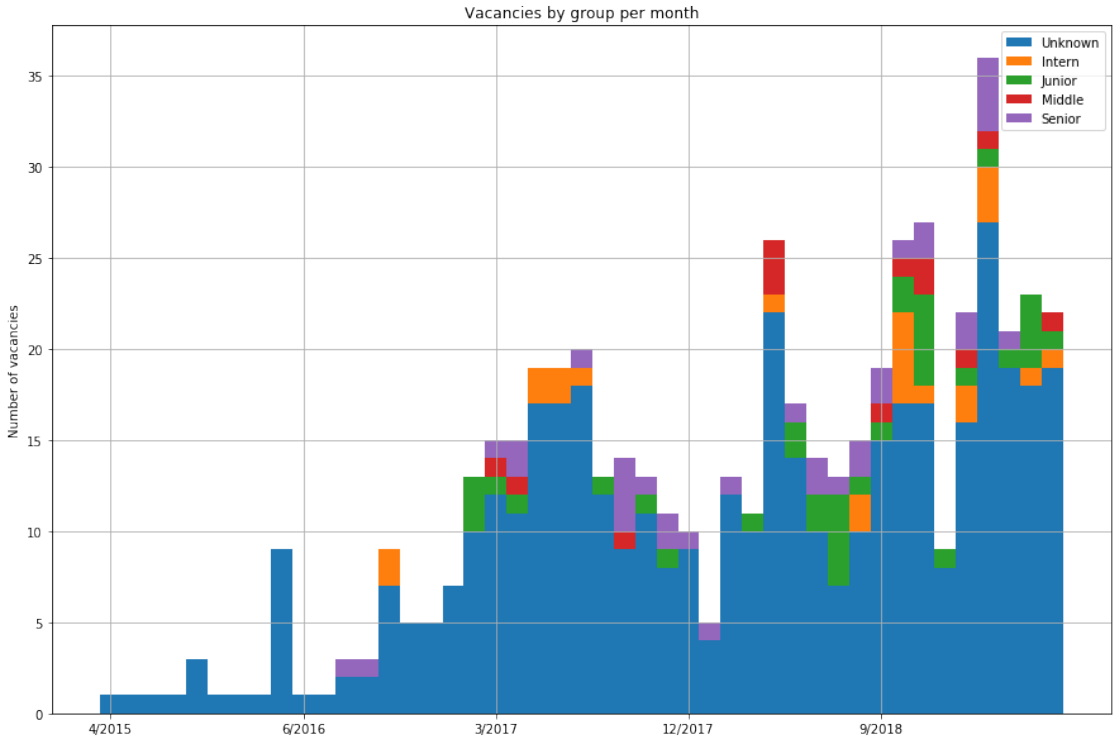
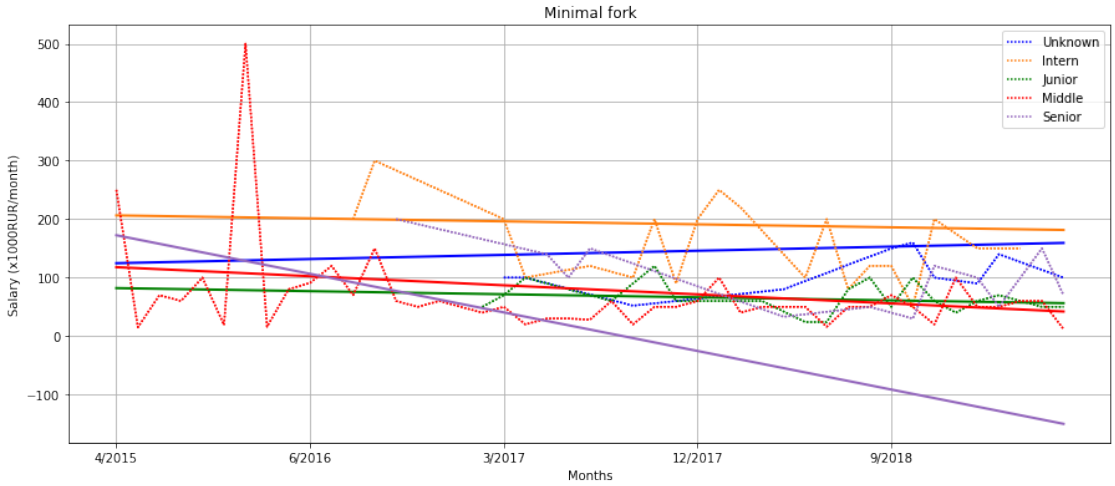
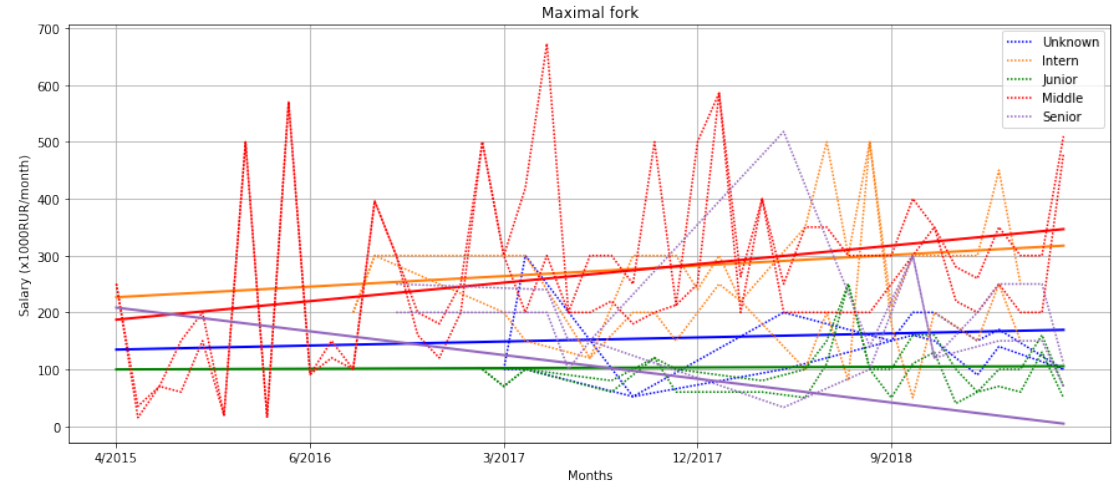
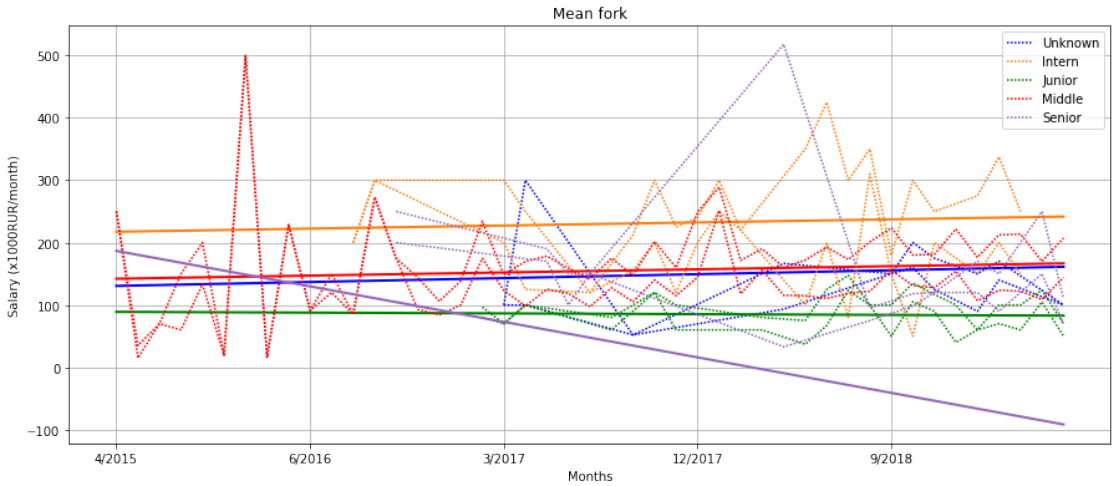


Рис. 3 Средние зарплаты по месяцам


Minimal fork
Unknown
Intern
Junior
Middle
Senior
Salary (x1000RUR/month)
500
400
300
200
100
0
−100
4/2015
6/2016
3/2017
12/2017
9/2018
Months


Maximal fork
Unknown
Intern
Junior
Middle
Senior
Salary (x1000RUR/month)
700
600
500
400
300
200
100
0
4/2015
6/2016
3/2017
12/2017
9/2018
Months


Mean fork
Unknown
Intern
Junior
Middle
Senior
Salary (x1000RUR/month)
500
400
300
200
100
0
−100
4/2015
6/2016
3/2017
12/2017
9/2018
Months

1.4 Выводы

Анализ был проведен по вакансиям российских компаний, так что он отражает ситуацию преимущественно на российском рынке специалистов в Data Science.

Явно видно, что количество требуемых специалистов увеличивается, и в среднем зарплаты увеличиваются, однако этот рост, можно сказать, только покрывает инфляцию (инфляция в 2018 году составила 4.2%, и отношение роста средней зарплате к средней зарплате сейчас составляет те же самые 4.2%). Только для позиции джуниоров наблюдается спад средней зарплаты.

Набор данных оказался довольно информативным относительно зарплат, однако пока не хватает разделения по позициям. Возможно, эту информацию можно почерпнуть из данных с помощь дополнительных алгоритмов. В наборе данных содержится информация про вакансии за рубежом, которую также можно добавить в анализ, так как рынок труда в Data Science доволльно сильно глобализован.