Трек#3:
Модель предсказания стоимости навыков для портала «РАБОТА В РОССИИ»

Подготовили:
Крайникова Влада
Балчиди Глеб
Ташпулатов Азиз

Лишний Носок
20 декабря 2020

Общий принцип:

Идти от простого к сложному:
базовые модели -> отбор -> усложнение -> отбор -> усложнение усложнения.

Данные: обработка и дополнение

Что было сделано с данными?

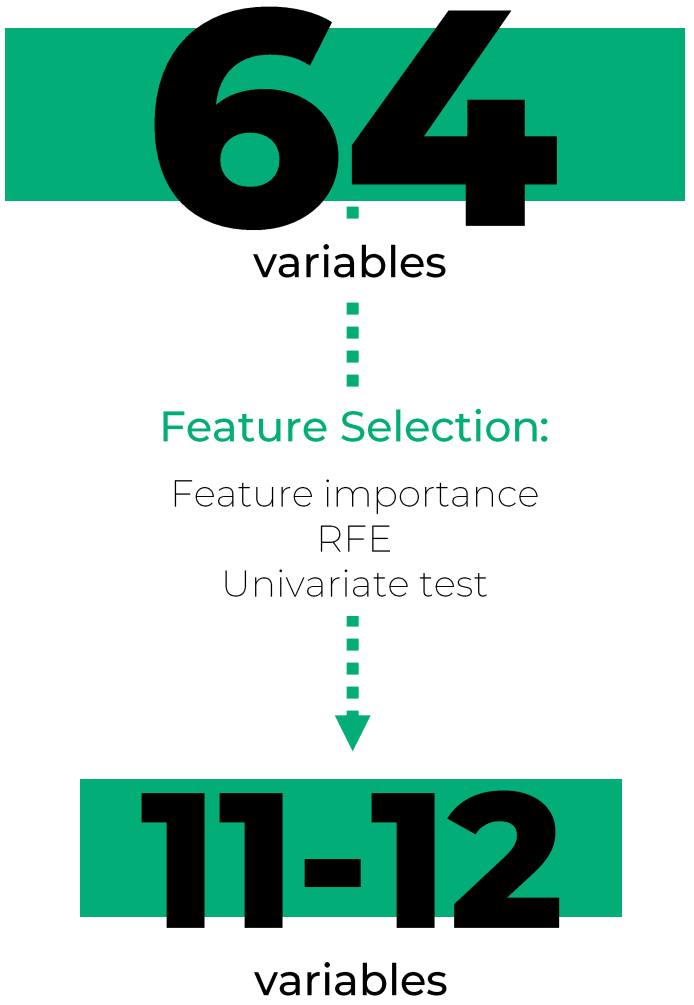
- Заменили нулевые значения отдельными категориями или медианными значениями.
- Переработка категориальных переменных (dummy или encoding)
- Тематическое моделирование для position (20 групп – специальностей)
- Векторизация текстовых переменных: responsibilities и доп. индексация категорий для TabNet

Дополнительные данные:

- Рейтинги университетов, т.к. образование значимый фактор
- Кластеризация городов на основе среднего уровня заработной платы в них

Отобранные переменные:

Тип образования, опыт, ожидаемая з/п, готовность к переезду и путешествию, заполненность анкеты, возраст, водительские права, группа специальностей, Вахтовый метод и кластер.



Модели:

Попробовали LR и DT

- Линейка не справилась с предсказаниями.
- Деревья дали и то, и другое.

Альтернативы

- Попробовали MARS (для нелинейности)
- Использовали LR для инсайтов.
- Построили NN в надежде победить.

«Деревянные» ансамбли

- Попробовали Bagging, RF, Boosting (разные вариации), Voting, остановились на GB, т.к. он позволяет учитывать ошибки предсказаний

Gradient Boost

Tuning:
Модель градиентного спуска с LAD функцией потерь была выбрана итоговой моделью, т.к. ее ошибка стабильно была ниже ошибка остальных моделей. Для ее усовершенствования был применен подбор гиперпараметров в 400 итераций.

Ошибки моделей:

Модель	OLS	Decision Tree	Extra Trees	Gradient Boost	Bagging	Voting	TabNet	MARS
RLSME:	1.13*	1.04	1.10	1.02	1.11	1.09	1.08	1.19

Немного инсайтов

«Не важно, где ты получал образование, важно, что ты его получил»
Education:
Тип образования положительно связан с зарплатой.
Рейтинг университета не стал значимым фактором.

«Готов на все!»
Relocation, travel, retraining:
Соискатели, отметившие это в своих резюме, получают более выгодные предложения.

«На минималках»
Мин. ур. з/п:
Практически нет знач. факторов, которые бы минусовали з/п, т.е. в основном указывание тех или иных пунктов положительно влияет на зарплату.

«Я не сексист, но вижу стат. разницу»
Gender:
Мужчины получают большую з/п, нежели женщины.

«Хочу много денег!»
Desired salary:
Каждый указанный рубль дает ~50 коп. к зарплате.

«Заводись, поехали!»
Driver license:
Наличие прав (в частности категории B) положительно связано с з/п.

«Больше з/п богу з/п»
Также положительно связаны с з/п:
Опыт, ненормированный график и полный график, ценится Вахтовый метод и удаленный, заполнение анкеты.

Итого:

38 часов
чистого кодинга

17 моделей
не считая их вариации

1 000 000 ₽
который мы не выиграли

1 Jupiter
пострадал во время моделирования

человека фантом
2 + 1
шутка.

Трек#3:
Модель предсказания стоимости навыков для портала «РАБОТА В РОССИИ»

Дизайн делали сами☺

Подготовили:
Крайникова Влада
Балчиди Глеб
Ташпулатов Азиз

Лишний Носок
20 декабря 2020