Разработка модели ранжирования соответствия пула кандидатов предлагаемой вакансии от FriendWork

SmartAnalytics

Черных Иван @iceman_o_O г. Москва

Мурзина Ольга @olga_murzina г. Н.Новгород

Стек технологий

- Python 3.7
- Colab

- Pandas
- TensorFlow
- CatBoost
- SkLearn
- NLTK
- Re
- Matplotlib

Совместная работа

- Colab
- Google Drive
- Telegram

Работа с моделями

- Загрузка, обработка и подготовка данных
- Выбор моделей
- Обучение моделей
- Контроль качества на тестовой выборке
- Итоговый прогноз на базе двух моделей
- Форматирование полученного результата под требования Заказчика

Подготовка данных 1

- Загрузка из .csv-файлов
- Проверка данных на пропуски и дубли
- Объединение данных
- Создание синтетических признаков
- Оцифровка категориальных данных через OHE-кодирование (для категорий) и токенизацию (для текстов)

Подготовка данных 2

Генерация дополнительных данных:

Изначально входными данными являются относительно "позитивные" данные, когда кандидат был рассмотрен на должность, т.е. заведомо связанные данные вакансии и кандидата.

Однако для поиска признака, на сколько подходят вакансия и кандидат, необходимо иметь негативный набор данных. За основу негативных данных можно взять такие связь вакансия-кандидат, у которых схожесть в названии вакансий максимально низкая

В рамках проекта были сгенерированы дополнительные данные, что позволило значительно повысить точность моделей

Модель – TensorFlow

- **Вход** - только числовые данные
- **Подбор параметров** – по-слойно
- **Обучение модели** – отслеживание прогресса и визуализация значений loss-функции и выбранной метрики
- **Контроль качества** – Accuracy
- **Результат** – вероятность получения / неполучения работы кандидатом

Модель – CatBoostRegressor

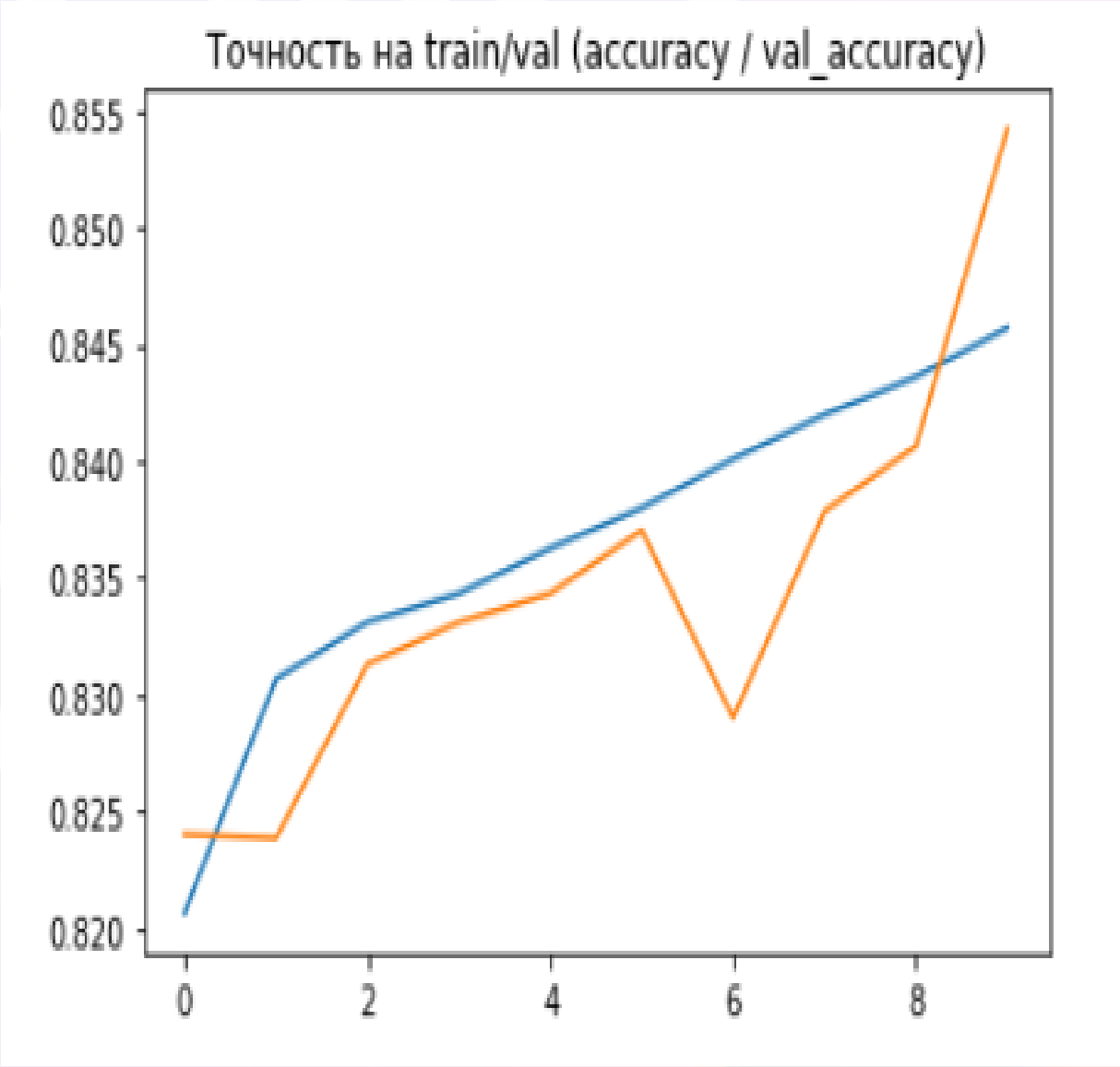
- **Вход** – категориальные и числовые данные
- **Подбор параметров** – можно в авто-режиме
- **Обучение модели** – отслеживание прогресса и визуализация значений loss-функции и выбранной метрики только для Jupiter
- **Контроль качества** – AUC-ROC и R2
- **Результат** – вероятность получения работы кандидатом

Оценка качества (TensorFlow)

Нейронная сеть:

loss-функция – **binary_crossentropy**

метрика - **accuracy**



Оценка качества (CatBoost)

CatBoostRegressor:

loss-функция – **RMSE**

метрика – **ROC-AUC**

Результат 1

Механизм для сортировки - **синтетический "рейтинг"** зависит от:

- **региона проживания**,
- **схожести вакансии** и желаемой кандидатом **позиции**,
- **оценки модели + схожести** характеристик вакансии и набора характеристик кандидата.

Результат 2

- "Рейтинг" представлен в числовом виде и для каждого кандидата вычисляется по сложной формуле.

- К выдаче идет ранжированный по рейтингу в порядке убывания список кандидатов в формате candidateId - rating.

SKOLKOVO HACK

ВАШ КОД К УСПЕХУ!

2022

23-25 сентября
ONLINE

Спасибо за задание!