Применение алгоритмов машинного обучения для автоматического анализа вакансий из портала HeadHunter.ru

Автор работы: Валерия Морозова
Научный руководитель: Александр Югай

4. Baseline модели

Для постороения бейзлайна рассмотрим два подхода к формированию обущающей выборки: из категориальных и из текстовых признаков.

4.1. Категориальные признаки

За основу выборки были выбраны следующие поля:

– тип занятости

– тип графика

– требуемый опыт

– локация

– доступность для несовершеннолетних соискателей

– доступность для соискателей с ограниченными возможностями

– тип графика работы

– ключевые навыки

Поля “ключевые навыки” и “локация” потребовали дополнительной обработки. Оба поля являются категориальными, но допустимо большое разнообразие меток (4200 для навыков и 1000 для локаций), что затрудняет их кодирование. Препроцессинг данных полей заключался в группировке редких на данной выборке меток в отдельный класс. Для навыков было выбрано 150 наиболее частотных меток, для локаций – 100.

Прочие признаки были закодированы при помощи ordinal encoding.

В качестве модели бейзлайна был применен регрессионный случайный лес модуля sklearn со стандартными параметрами.

4.2. Текстовые признаки

Для обучения текстовой модели было использовано поле описание вакансии. Брендированное описание вакансии было исключено из эксперименто, поскольку оно заполнено достаточно редко в текущей выборке (только для 18% вакансий).

Тексты описаний были предварительно обработаны путем токенизации, лемматизации, удаления стоп-слов и специальных символов. Для препроцессинга были применены модули nltk и pymorphy.

В качестве бейзлайна мною была выбрана архитектура, содержащая векторизацию документов при помощи TF-IDF векторизатора и линейной регресии со стандартными параметрами из модуля sklearn.

4.3 Результаты

Для обоих экспериментов из выборки были отобраны семлы с заполненной целевой переменной (около 15 000 вхождений). Полученные данные были разбиты на обущающую и тестовую выборку в соотношении 70% и 30%. Для оценки качества использовалась метрика MAE (mean absolute error).

Таблица 1. Метрики Baseline моделей

Модель	**MAE**
RandomForestRegressor	28526
TfidfVectorizer + LinearRegression	59233

Качество случайного леса оказалось вдвое выше, что было ожидаемо: некоторые категориальные переменные значительно скоррелированы с целевой переменной. Результат оценки feature importance совпал с выводами разведочного анализа: наиболее скоррелированы с целевой переменной признаки “опыт”, “тип занятости”, “локация” и “ключевые навыки”.

Векторизация при помощи TF-IDF, возможно, является слишком примитивной для данного датасета.

5. Архитектура проекта

Проект состоит из следующих этапов:

1. Сбор и подготовка данных.

Сбор данных произведен посредством строенных библиотек python. Хранение данных осуществляется в нереляционной базе данных ElasticSearch.

2. Обучение и эксперименты.

Препроцессинг данных реализован при помощи классических библиотек NLP для в приложении к русскому языку: nltk, sklearn.

Обучение моделей происходит на базе библитек sklearn (для неглубоких моделей) и pytorch (для глубокого обучения).

Трекинг экспериментов производится при помощи сервиса MLFlow.

3. Деплой.

Интерфейс итоговой модели планируется осуществить в виде REST API сервиса на основе фреймфорка fastapi, контейнеризованного при помощи docker. Для мониторинга сервиса планируется использовать платформу graphana.

6. План экспериментов

Предусмотрено два направления экспериментов: работа над векторным представлением описания вакансии и проектирование регрессионной архитектуры.

Эксперименты над векторным представлением текстовых переменных включают в себя испытание:

- классические эмбеддинги – как предобученные, так и обученные с нуля на текущей выборке: Word2Vec, Fasttext, GloVe;
- более глубокие модели, такие как ELMO – предобученные и fine-tuning;
- трансформеры – предобученные и fine-tuning;

Также рассматривается возможность включение в архитектуру модели предсказания целевой переменной обработки категориальных признаков, так как, согласно разведочному анализу и бейзлайну, они достаточно сильно скоррелированы с целевой переменной и могут значительно отразиться на качестве.

Основная метрика для оценки качества предсказания уровня заработной платы – MAE (mean absolute error).

Singh, Lovedeep. (2022). Clustering Text: A Comparison Between Available Text Vectorization Techniques. 10.1007/978-981-16-1249-7_3.

Kozhevnikov, Vadim & Pankratova, Evgeniya. (2020). Research of the Text Data Vectorization and Classification Algorithms of Machine Learning. Theoretical & Applied Science. 85. 10.15863/TAS.2020.05.85.106.