Работу можно увидеть по предоставленной ссылке:

<https://colab.research.google.com/drive/1OFQik3Phz6dksMhWp3YONoc2sqba5zC1?usp=sharing>

Я провела работу над проектом «Реальные/поддельные вакансии». При поиске работы на сайтах каждый из нас мог столкнутся с очень привлекательными вакансиями. Мошенничество с трудоустройством растет.  Текущая рыночная ситуация привела к высокому уровню безработицы. Экономический стресс и влияние коронавируса значительно сократили доступность работы и потерю работы для многих людей. Подобный случай представляет подходящую возможность для мошенников. Многие люди становятся жертвами этих мошенников, используя отчаяние, вызванное беспрецедентным инцидентом. Большинство мошенников делают это, чтобы получить личную информацию от человека, которого они обманывают. Личная информация может содержать адреса, реквизиты банковского счета, номера социального страхования и т. д. Мошенники предоставляют пользователям очень выгодную возможность трудоустройства, а затем просят взамен деньги. Или они требуют инвестиций от соискателя с обещанием работы.

В этом проекте я использовала данные, предоставленные [Kaggle](https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction" \t "_blank) . Эти данные содержат признаки, определяющие объявление о вакансии. Эти объявления о вакансиях классифицируются на настоящие и поддельные. Поддельные объявления о вакансиях составляют лишь малую часть этого набора данных.

Этот проект состоит из пяти этапов. Пять шагов, принятых для решения проекта:

1. Определение цели и задачи (обзор проекта, описание проекта и метрики)
2. Сбор данных
3. Очистка данных, исследование и предварительная обработка
4. Моделирование
5. Оценка

Постановка задачи

Этот проект направлен на создание классификатора, который сможет идентифицировать поддельные и настоящие вакансии работ. Окончательный результат оценивается на основе нескольких моделей. Поскольку предоставленные данные имеют числовые и текстовые характеристики, одна из моделей будет использоваться для текстовых данных. Окончательная модель будет принимать любые соответствующие данные о вакансиях и давать окончательный результат, определяющий, является ли вакансия реальной или нет.

Метрика

Модель оценивается в основном по метрике F1 : метрика F1 — это мера точности модели в наборе данных. Формула для этого метрики –



Оценка F1 используется потому, что в нашем случае решающее значение имеют как ложноотрицательные, так и ложноположительные результаты. Эта модель должна идентифицировать обе категории с максимально возможной оценкой, поскольку обе они имеют высокие затраты.

Так же мы смотрели на метрику Accuracy: процент верных предсказаний.



EDA

Данные для этой работы можно найти по ссылке: <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction>

В данном датасете - 17880 объявлений и 18 признаков. Из числовых telecommuting, has\_company\_logo, has\_questions. Остальные типа object. Целевая переменная у нас 'fraudulent'. Так же в датасете имеются пропуски. Дубликатов нет.

Краткое описание признаков:

Text

Description automatically generated

В датасете достаточно много пропусков было. Job\_id было сразу удалено после проверки на дубликаты.

Сразу был преобразован признак ‘salary\_range’ для извлечения числовых значений. Далее я получила max и min значения «предлагаемой вакансией зарплаты» и искуственный признак ‘salary\_specified’. Чтобы признак меньше коррелировал, я получила ‘delta\_salary’ – разницу максимальной и минимальной зарплаты.

Был создан бейзлайн для получения первоначальной метрики и дальнешего сравнения с изменениями в датасете. Для бейзлайна была использована наиболее новая модель – CatBoost. Высокое качество этой модели может быть получено без настройки параметров, а очень хорошие результаты могут быть получены при использовании параметров по умолчанию, что сокращает время, затрачиваемое на настройку параметров. А так же поддержка категориальных переменных без предварительной обработки. Точность измерялась на 5-fold кросс-валидации, поскольку данных у нас не очень много.

Chart, bar chart

Description automatically generatedПри применении оценки важности параметров стало ясно, что признаки ‘has\_company\_logo’ и ‘company\_profile’ достаточно сильно влияют на нашу модель. Но первый признак несет в себе важную информацию, касающуюся статистики, поэтому он оставлен. В дальнейшем я сравнила метрику после удаления ‘company\_profile’, она показала улучшение финальной метрики, поэтому было принято решение об удалении этого признака.

Из признака ‘location’ была извлечена информация о месторасположении вакансий: страна, штат(в случае с США, если имеется) и город.

Создан признак, подсчитывающий количество символов в описании вакансии – ‘character\_count’

Каждый признак был рассмотрен отдельно, произведена очистка, извлечение чисел и текста из признаков, измерена точность на экспериментах с удалением признаков. Главные выводы:

* Вакансии,которые обещают работу с дома, имеют практически в 2 раза больше вероятности мошенничества
* Вакансии, в которых компании без фирменного логотипа, имеют в 8 раз больше вероятность оказаться поддельными.
* Мошенники не добавляют проверочные вопросы. Вакансии, в которых нет вопросов к соискателю, имеют в 3 раза больше вероятность оказаться мошенническими.
* Работы,где указана в вакансии «частичная занятость» имеют самый высокий процент (почти 10%) мошенничества. На втором месте вакансии,где не указан тип работы, то есть возможно мошенники не вдаются в подробности вакансии
* Указание в вакансиях о незаконченном среднем образование являются поддельными (примерно 70%). Кстати, сертификаты тоже имеют большой процент - около 10%.
* Также был отмечен странный факт: нереальные вакансии обычно предлагают меньшую заработную плату
* В основном фейковые вакансии происходят из США. В Европейских странах это встречается реже
* Количество символов относительно одинаково как для реальных, так и для поддельных публикаций, однако реальные вакансии имеют все-таки большее количество символов в описании. Можно подтвердить ранее упомянутое,что мошенники не особо вдаются в подробности вакансии

В итоговом датасете следующие признаки: 'title', 'location', 'department', 'description’, 'requirements', 'benefits', 'telecommuting', ‘has\_company\_logo’, 'has\_questions', 'employment\_type', 'required\_experience', 'required\_education', 'industry', 'function', 'fraudulent', delta\_salary', ‘salary\_specified’, ‘character\_count’.

Перед использованием текстовых данных для любого моделирования данных требуется дополнительная предварительная обработка текста. Поэтому была проведена очистка текста, токенизация и удаление stopwords.Векторизован очищенный текст в матрицу tfidf со всеми словами с помощью TfidfVectorizer.

В целевой переменной была замечена явная несбалансированность данных, что может привести к переобучению, поэтому я применяла параметр sample\_weight при обучении и функцию SMOTE для оверсэмплинга.

Shape

Description automatically generated with low confidence

Далее исользовались базовые модели для сравнения результатов, такие как: RandomForest, GradientBoosting, KNeighbors. Была применена наиболее частая реализацию бустинга - алгоритм AdaBoost на основе RandomForest . Это мета-алгоритм, который в процессе обучения строит композицию из базовых алгоритмов обучения для улучшения их эффективности.

По итогу очистки и всех преобразований над датасетом, лучшие результаты по метрике F1 показали: CatBoostClassifier – 85%,GradientBoosting с подбором параметров – 86% .

Классификация confusion matrix CatBoost: [3373 25]

[67 111]

A screenshot of a computer

Description automatically generated with medium confidence Так же была попытка использовать  "Двунаправленные LSTM" и результат метрики F1- 83%. По confusion matrix можем заметить, что модель очень хорошо классифицировала наши данные.

Bi-Directional LSTM представляет собой две однонаправленные рекуррентные сети, одна из которых обрабатывает входную последовательность в прямом порядке, а другая — в обратном. те же алгоритмы, что и для RNN. Благодаря данной архитектуре сети доступна информация о контексте как из прошлого, так и из будущего, что решает проблему однонаправленных рекуррентных сетей. Для обучения biRNN используются.

По итогу CatBoost показал лучшую классификацию на реальные и поддельные вакансии!

Вывод:

В результате анализа, мы можем обоснованно распознать, является ли вакансия мошеннической, используя несколько особенностей, основанных на анализе информации:

1. мошеннические вакансии обычно не имеют корпоративного логотипа;
2. не добавляют проверочные вопросы;
3. обычно упоминают информацию о зарплате даже в названии, чтобы заманить кандидатов;
4. могут пропустить определенные атрибуты работы (например, отрасль, должность, уровень образования кандидата,и уровень опыта).

Модели так же предполагают, что профиль компании является не самой сильной стороной, хоть и заметно высокое влияние на модель, существуют и другие полезные функции, такие как описание должности, требования и бонусы, которые также обладают хорошей предсказательной силой.

Проведена вся необходимая работа по подготовке данных, их анализ, извлечение информации, обработка и подготовка для дальнейшей работы. Создан бейзлайн, по которому сравнивались изменения в датасете. Датасет проверен на дубликаты и выбросы, выполнена нормализация, использовались классические методы кодирования категориальных признаков. Подготовлены дополнительные признаки. По итогам анализа, из некоторых признаков были извлечены более полезная информация, и далее изначальные удалены. Использовались классические методы обработки языка,а именно очистка текста. В модели применялась токенизация и лемматизация.Я убедилась, что с помощью базовых моделей, можно получить хорошую предсказательную силу, применив NLP, поработав с обработкой текста. С базовыми моделями, основанными на NLP, мы получаем очень хорошую точность - av.F1 - 85% с минимальной настройкой гиперпараметров. С более сложными моделями, такими как "Двунаправленные LSTM" по-прежнему получаем хороший результат и неплохую классификацию по confusion matrix.