НЕТОЛОГИЯ

ПРОГРАММА “АНАЛИТИК ДАННЫХ”

Предсказание оттока сотрудников

Дипломная работа

Обучающийся _____ Анна Владимировна Кошелева

Руководитель _____ Анастасия Сергеевна Успенская

Москва 2021

Содержание

1. Постановка задачи

Для многих крупных компаний текучесть кадров является серьезной проблемой, так как стоимость замены сотрудника может быть достаточно высокой, а отток ценных кадров снижает производительность компании. Таким образом, задачей HR-аналитиков является не только поиск сотрудников, но и предотвращение ухода уже существующего персонала. Для этого HR-ам необходимо знать как факторы, увеличивающие вовлеченность сотрудников, так и наоборот, побуждающие сотрудников уйти. Часто HR-менеджеры лично отслеживают комфортность рабочей среды, наблюдая за сотрудниками и общаясь с ними. Но для компаний с большим количеством сотрудников, и следовательно, с большим количеством данных, имеет смысл цифровизировать HR-процессы. С развитием машинного обучения и ростом объема данных одним из методов анализа и предсказания поведения сотрудников стала предиктивная HR-аналитика ([10], [1], [3]-[7], [15]).

К компаниям, активно использующим предиктивную аналитику для предотвращения оттока и контроля эффективности сотрудников, относятся такие гиганты, как Google, Hewlett-Packard, IBM, JPMorgan, Credit Suisse, Experian, Nilsen, EY — крупнейшая аудиторско-консалтинговая компания в мире, Best Buy — американская компания, владеющая крупной сетью магазинов бытовой электроники и сопутствующих товаров, а также US Special Forces — силы специального назначения Армии США ([1]-[7], [15]). В России это такие компании, как Альфа-банк, Сбербанк, ВымпелКом, МегаФон, МТС, Ростелеком, X5 Retail Group ([1], [3], [6], [7]).

1.1. Постановка бизнес-задачи, бизнес требования и стейкхолдеры

Бизнес-задачей, решаемой в нашем проекте, является снижение оттока сотрудников и выявление факторов, которые влияют на решение сотрудников уволиться.

Для решения этой задачи мы построим модель машинного обучения с учетом следующих бизнес-требований:

1) модель должна предсказывать решение сотрудника уволиться,

2) выявлять факторы, влияющие на это решение,

3) легко модифицироваться под дата-сеты с другим количеством признаков,

4) язык программирования — Python,

5) модель должна иметь подробную документацию.

Стейкхолдерами проекта являются крупные предприятия.

1.2. Метрики для оценки качества модели, гипотезы для проверки

Построенная нами модель является моделью бинарной классификации, которая определяет, с какой вероятностью уволится сотрудник и попадает ли он в класс сотрудников, планирующих уйти или в противоположный класс. Но также мы вычислим и степень влияния каждого из признаков модели на целевую переменную, следовательно, можно, например, проверить гипотезу: "признак A влияет на решение сотрудника уволиться".

Качество модели мы будем оценивать с помощью стандартных для моделей классификации метрик: accuracy — доли правильных ответов, precision — точности и recall — полноты, а также с помощью анализа площади под precison-recall и ROC-AUC кривыми, [8].

2. Анализ данных

2.1. Исследования аналогичных решений

Известны примеры успешного применения HR-аналитики крупными компаниями по уменьшению оттока сотрудников. Например, Ростелеком смог удержать до 70% ключевых сотрудников, планировавших уволиться и сэкономил миллиарды рублей [7],

Hewlett-Packard, JPMorgan, Credit Suisse, Experian, IBM, Nilsen, EY также, предотвратив уход ключевых сотрудников, смогли сэкономить до нескольких десятков миллионов долларов ([2], [3], [5], [15]).

Пример прогнозирования оттока сотрудников на данных, предоставленных IBM и на данных с платформы Kaggle, можно посмотреть в статьях [9] и [12].

2.2. Предварительный анализ данных

Мы будем обучать модель на данных с платформы Kaggle: «Employee Attrition. Fictional dataset on HR Employee attrition and performance», https://www.kaggle.com/patelprashant/employee-attrition.

В выборке имеется 4410 исторических наблюдений и 20 переменных, одна из которых — переменная Attrition, является целевой. Таким образом, про каждого из 4410 сотрудников мы знаем значения 20 их характеристик (возраст, пол, образование, место и область работы, расстояние от дома и т.д.), в том числе значение целевой переменной: 1 — если сотрудник уволился и 0 — если не уволился.

Ниже приведем таблицу с признаками:

Имя столбца	**Значение**	**Имя столбца**	**Значение**
Age	Возраст	NumCompaniesWorked	Количество компаний, в которых работал сотрудник
BusinessTravel	Частота командировок	PercentSalaryHike	Процент повышения з/п за время работы
Department	Отдел	StandardHours	Стандартная продолжительность рабочего дня
DistanceFromHome	Расстояние от дома в км	StockOptionLevel	Уровень опциона на акции
Education	Уровень образования	TotalWorkingYears	Общий трудовой стаж
EducationField	Сфера образования	TrainingTimesLastYear	Общее время дополнительного обучения
Gender	Пол	YearsAtCompany	Стаж работы в данной компании
JobRole	Должностная роль	YearsSinceLastPromotion	Количество лет с последнего повышения
MaritalStatus	Семейное положение	YearsWithCurrManager	Количество лет работы с текущим менеджером
MonthlyIncome	Ежемесячный доход	**Attrition**	**Целевая переменная: ушел работник или нет**

Вся выборка разбивается нами на две части: для обучения и для тестирования модели. В обучающей выборке 3308 (75%) сотрудников, в тестовой выборке — 1102 (25%).

Далее для обучающей и тестовой выборок мы получим сводные статистики с помощью метода describe() и проверим данные на наличие пропусков с помощью метода info(). У двух признаков имеется незначительно количество пропусков. Для заполнения пропусков мы рассчитаем средние значения признаков в обучающей выборке, и заполним полученными числами пропуски как в тестовом наборе данных, так и в самой обучающей выборке, так как при решении реальной задачи нам будут доступны только данные для обучения.

В исходной выборке имеются также текстовые данные. Для дальнейшей работы с ними закодируем эти данные с помощью метода LabelEncoder() из библиотеки sklearn.

Опишем кратко план анализа данных:

1) загрузить данные для обучения;

2) обработать данные перед обучением модели;

3) обучить модель на обучающей выборке;

4) загрузить и предобработать данные для тестирования;

5) провалидировать модель на тестовой выборке.

2.3. Методика решения

Мы воспользуемся двумя методами для построения моделей классификации и сравним их между собой: методом логистической регрессии и градиентным бустингом над решающими деревьями. Будем применять следующие готовые реализации этих методов: LogisticRegression из библотеки sklearn для логистической регрессии и XGBClassifier из библиотеки xgboost для градиентного бустинга над решающими деревьями.

Метод логистический регрессии — это один из самых старых и основных методов классификации, относительно быстрый и с небольшим количеством настраиваемых параметров, [14].

Метод XGBoost градиентного бустинга над решающими деревьями — наоборот, один из самых последних методов классификации, и считается одним из самых универсальных и сильных методов машинного обучения, известных на сегодняшний день, [11], [13].

3. Результаты

После того, как мы обучили модель на обучающей выборке, провалидируем ее на тестовой выборке и вычислим точность прогноза с помощью метрики accuracy — доли правильных ответов, [8]. Мы будем использовать готовую реализацию для подсчета этой метрики — функцию accuracy_score() из библиотеки sklearn.

Провалидировав модель, мы получили, что точность предсказания модели логистической регрессии равна 0.836, а точность модели градиентного бустинга над решающими деревьями равна 0.876.

Таким образом, модель градиентного бустинга над решающими деревьями работает лучше, поэтому для дальнейшего исследования мы оставим её.

Для модели XGBClassifier мы также составим таблицу сопряженности модели классификации и найдем прогноз вероятности принадлежности к классу.

Чтобы получить более полное представление о качестве модели XGBClassifier, мы исследуем такие метрики, как precision — точность и recall — полнота, [8].

Проанализировав таблицу сопряженности модели, а также значения метрик precision и recall, мы приходим к выводу, что

- классификатор срабатывает достаточно редко, но 94% сотрудников, отнесенных им в класс 1, действительно увольняются;
- при этом классификатор имеет низкую полноту в 25%, то есть достаточно часто ложно бездействует.

Также мы построим две кривые AUC-PR (precision-recall кривую) и AUC-ROC и проанализируем площади под ними, [8].

Площадь под кривой AUC-PR равна 0.71, а под кривой AUC-ROC равна 0.9. Относительно невысокое значение площади под AUC-PR при высоких значениях accuracy и площади под AUC-ROC говорят о том, что положительный класс 1 намного меньше класса 0, то есть классы несбалансированны. В нашей задаче этой означает, что количество сотрудников, которые решают уволиться, намного меньше количества сотрудников, которые остаются.

Дальше мы определим важность признаков для модели. Во-первых, это позволит нам определить причины увольнения сотрудников. Во-вторых, признаки, важность которых для модели очень низкая, можно исключить из модели, тем самым сократив время обучения.

Мы видим, что самые значимые признаки для исследуемого дата-сета — это количество лет, которые сотрудник проработал в данной компании, его семейный статус, количество лет, которые сотрудник проработал вместе с одним и тем же менеджером и возраст сотрудника. То есть, можно сформулировать гипотезу, что если сотрудник проработал в компании больше определенного количества лет и его устраивает непосредственный начальник, то этот сотрудник не склонен менять место работы.

Анализируя значимости признаков для классов 0 и 1, можно, например, сформулировать гипотезы, что причинами увольнения сотрудника является отсутствие повышения заработной платы в течении определенного периода и большое расстояние от дома до работы. Вероятно также, что сотрудники, несколько раз менявшие место работы до устройства в компанию, более склонны к увольнению.

Самыми незначимыми признаками для модели являются пол сотрудника и продолжительность рабочего дня.

4. Выводы и заключение

Мы построили модель прогнозирования оттока сотрудников на основе метода XGBoost градиентного бустинга над решающими деревьями и на дата-сете с искусственными данными, предоставленными IBM, выяснили, что модель с достаточно высокой точностью предсказывает увольнение сотрудников. С помощью проведенного анализа значимости признаков можно выдвигать гипотезы о том, какие из факторов влияют на увольнение сотрудника, а какие, наоборот, удерживают его. Построенную модель можно модифицировать и под другие дата-сеты с данными о сотрудниках и целевой переменной, в которой хранится информация, уволился сотрудник или нет.

В качестве заключения отметим, что существуют непрогнозируемые обстоятельства, такие как пандемия этого года. Вот что говорит о моделях машинного обучения для предсказания оттока сотрудников и клиентов руководитель направления «Машинное обучение» компании «Норбит» Дмитрий Тимаков: «Если тех или иных значимых событий, меняющих конъюнктуру рынка, не было в обучающей выборке, то модель постепенно начинает деградировать, снижая свою точность и увеличивая разрыв между модельными представлениями и реальной ситуацией. Для выхода из такой ситуации можно применять дообучение модели на свежих данных», [6].

Список источников

[1] О. Вильде, Предиктивная аналитика в HR: модно или просто-напросто необходимо. [Электронный ресурс]: *IBS*. Режим доступа: https://ibs.ru/media/media/prediktivnaya-analitika-v-hr-modno-ili-prosto-naprosto-neobkhodimo/, свободный, дата обращения 23.04.2021.

[2] А. Вичугова, Как снизить текучку кадров с помощью Big Data и Machine Learning: реальный опыт 5 крупных компаний. [Электронный ресурс]: *Big Data. Школа больших данных.*. Режим доступа: https://www.bigdataschool.ru/blog/ml-for-hr-churn-rate-use-cases.html, свободный, дата обращения 23.04.2021.

[3] А. Елхин, Предиктивная аналитика в HR — модный тренд или жизненная необходимость? [Электронный ресурс]: *FINASSESSMENT — проект Финансовой академии «Актив»*. Режим доступа: https://finassessment.net/blog/predictiv-analitika-hr, свободный, дата обращения 23.04.2021.

[4] А. Луценко, Как предсказать увольнение сотрудников: 4 успешных бизнес-кейса. [Электронный ресурс]: *ARTSEARCH*. Режим доступа: https://atsearch.ru/kak-predskazat-uvolnenie-sotrudnikov-4-uspeshnyh-biznes-keysa, свободный, дата обращения 23.04.2021.

[5] О. Рыбакова, Предиктивная аналитика в HR. [Электронный ресурс]: *HR-Академия*. Режим доступа: https://hr-academy.ru/hrarticle/prediktivnaya-analitika-v-hr.html, свободный, дата обращения 23.04.2021.

[6] Д. Тимаков, Добровольное удержание. Как сохранить клиентов с помощью машинного обучения? [Электронный ресурс]: *СБЕР Про*. Режим доступа: https://sber.pro/publication/dobrovolnoe-uderzhanie-kak-sokhranit-klientov-s-pomoshchiu-mashinnogo-obucheniia, свободный, дата обращения 23.04.2021.

[7] Н. Черкасенко, Прогнозирование увольнений сотрудников — кейс компании «Ростелеком». [Электронный ресурс]: *HR-TV*. Режим доступа: https://hr-tv.ru/articles/prognozirovanie-uvolnenij-sotrudnikov-kejs-kompanii-rostelekom.html, свободный, дата обращения 23.04.2021.

[8] Classification metrics. [Электронный ресурс]: *Scikit-Learn. Machine Learning in Python*. Режим доступа: https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics, свободный, дата обращения 23.04.2021.

[9] Н. Bendemra, Building an Employee Churn Model in Python to Develop a Strategic Retention Plan. [Электронный ресурс]: *Medium. Towards Data Science*. Режим доступа: https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d, свободный, дата обращения 23.04.2021.

[10] D. Faggella, Machine Learning in Human Resources — Applications and Trends. [Электронный ресурс]: *Emerj Artificial Intelligence Research*. Режим доступа: https://emerj.com/ai-sector-overviews/machine-learning-in-human-resources/, свободный, дата обращения 23.04.2021.

[11] J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine. // *Annals of Statistics*. 29(5), 2001, p. 1189–1232.

[12] A. Navlani Predicting Employee Churn in Python [Электронный ресурс]: *DATACAMP*. Режим доступа: https://www.datacamp.com/community/tutorials/predicting-employee-churn-python, свободный, дата обращения 23.04.2021.

[13] V. Morde, V.A. Setty XGBoost Algorithm: Long May She Reign! [Электронный ресурс]: *Medium. Towards Data Science*. Режим доступа: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d, свободный, дата обращения 23.04.2021.

[14] M. Stojiljkovi?, Logistic Regression in Python. [Электронный ресурс]: *Real Python*. Режим доступа: https://realpython.com/logistic-regression-python/, свободный, дата обращения 23.04.2021.

[15] E. van Vulpen, 15 HR Analytics Case Studies with Business Impact. [Электронный ресурс]: *AIHR. ACADEMY TO INNOVATE HR*. Режим доступа: https://www.analyticsinhr.com/blog/hr-analytics-case-studies/, свободный, дата обращения 23.04.2021.