УДК 519.23

В. Д. Колесников

ПОСТРОЕНИЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОТТОКА СОТРУДНИКОВ

Для многих производственных предприятий и крупных компаний с большим количеством сотрудников характерна проблема текучести кадров и неопределенность в этом отношении. В результате руководство не способно эффективно контролировать отток ценных кадров, что может привести к снижению производительности. В работе поставлена задача построить прогнозирующую модель оттока сотрудников, которая помогла бы в решении данной проблемы. Рассмотрен набор данных оттока сотрудников, опробованы некоторые алгоритмы машинного обучения, среди которых выбран оптимальный, обучена прогнозирующая модель.

For many manufacturing enterprises and large companies which emply many people, there is a problem of staff turnover and, thus, unreliability. As a result, management can't effectively control the outflow of highly qualified personnel, which can result in reduced productivity. The author sees his task in addressing this issue through building a predictive model of the employee outflow. The research included reviewing a dataset of employee outflow, developing some machine learning algorithms and testing a predictive model.

Ключевые слова: машинное обучение, анализ данных, бинарная классификация, текучесть кадров, отток сотрудников.

Keywords: machine learning, data analysis, binary classification, staff turnover, employee outflow.

Актуальность рассматриваемой проблемы продиктована тем, что одним из важнейших ресурсов предприятия или компании являются сотрудники, которые, помимо простой рабочей силы, представляют собой источник знаний и накопленного опыта. Поэтому HR-менеджеры не только ищут новый персонал, но и прикладывают усилия по удержанию текущего. Существует множество публикаций из раздела экономики и управления, посвященных данному вопросу, например, [2—4].

Применение методов машинного обучения для решения данной проблемы позволит спрогнозировать уход того или иного сотрудника на основе определенных исходных данных, что, в свою очередь, может послужить дополнительной информацией для принятия мер по его удержанию. Готовую обученную модель можно внедрить в имеющеюся HRM-систему как обособленный модуль и пользоваться ею по назначению.

В качестве источника данных взят набор (датасет), находящийся в открытом доступе [8]. Все обработки и обучение произведены на CPU i7-4700HQ с тактовой частотой 2,4 GHz (с возможность разгона до 3,4). Ниже приведены некоторые характеристики набора данных.

Признаки:

— satisfaction_level — уровень удовлетворенности сотрудника;

— last_evaluation — последняя оценка уровня сотрудника (своего рода коэффициент полезного действия);

— number_project — число проектов, на которых отработал сотрудник;

— average_montly_hours — среднее количество отработанных сотрудником часов за месяц;

— time_spend_company — количество лет, проведенных сотрудником на предприятии;

— work_accident — признак, указывающий, происходил ли с сотрудником на работе несчастный случай;

— promotion_last_5years — бинарный признак того, имел ли сотрудник повышение за последние пять лет;

— department — занимаемая сотрудником должность;

— salary — зарплата, объектный признак, который не говорит о конкретных значениях;

— left — целевой бинарный признак, информирующий об оттоке сотрудников.

В таблице 1 приведены некоторые данные о целевом и остальных числовых и бинарных признаках (для них нет смысла приводить другие характеристики помимо минимального и максимального значений). Здесь STD — стандартное отклонение, последние три столбца — 0,25/0,5/0,75-перцентили.

Таблица 1

Характеристики числовых и бинарных признаков

Признак	Mean	STD	Min	Max	25 %	50 %	75 %
satisfaction_level	0,613	0,249	0,09	1	0,44	0,64	0,82
last_evaluation	0,716	0,171	0,36	1	0,56	0,72	0,87
number_project	3,803	1,232	2	7	3	4	5
average_montly_hours	201	49,943	96	310	156	200	245
time_spend_company	3,498	1,460	2	10	3	3	4
work_accident	—	—	0	1	—	—	—
promotion_last_5years	—	—	0	—	—	—	—
left	—	—	0	—	—	—	—

На рисунках 1, 2 для признаков salary и department приведено относительное количество сотрудников.

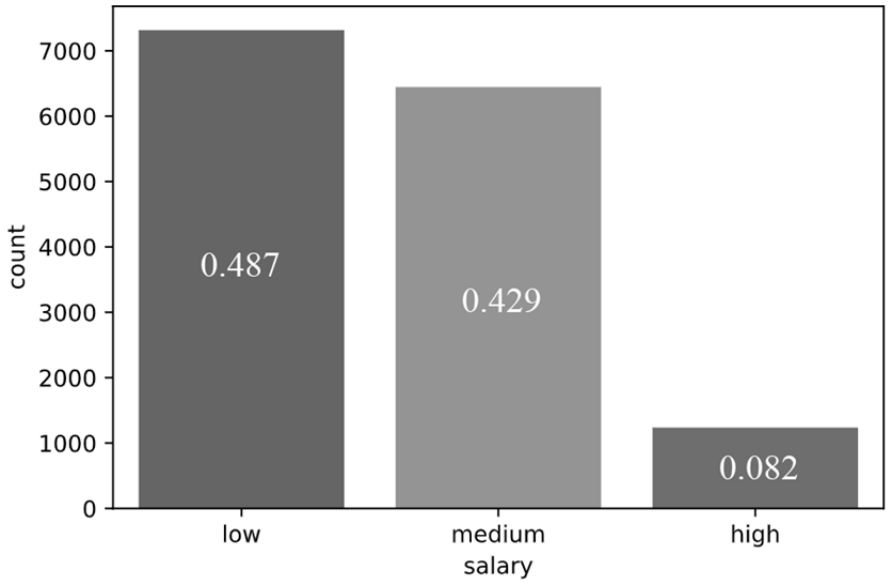

Рис. 1. Распределение количества сотрудников по признаку salary

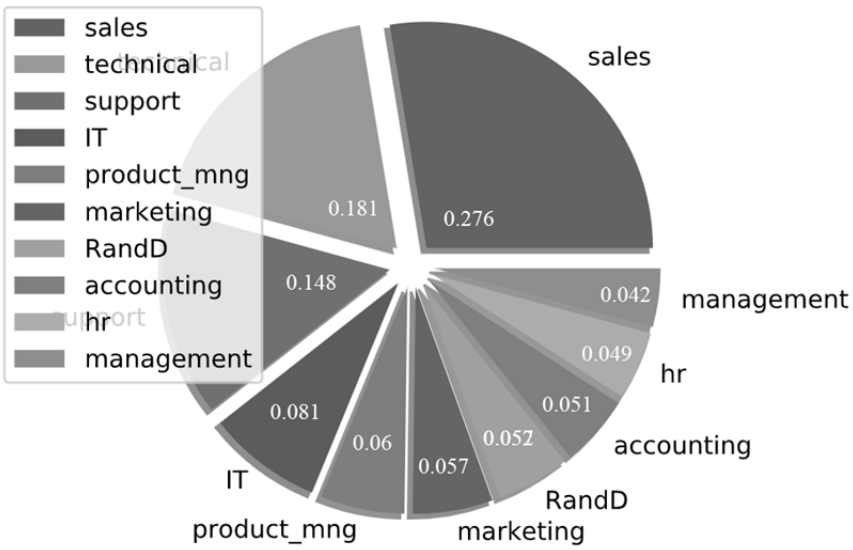

Рис. 2. Распределение количества сотрудников по признаку department

Поскольку признак salary является порядковым по своей природе, его значения заменены на 0, 1, 2. Для признака department проведена бинаризация [5].

Сначала была построена модель методом решающих деревьев. Предварительно было проведено разбиение датасета в отношении $^{3}/_{7}$, где 30 % данных отложены для тестов, а на остальных 70 % проведена 5-крат-

ная кроссвалидация на каждом сочетании параметров (maxDepth, maxFeatures), где maxDepth принимает значения от 4 до 10 и означает максимальную глубину дерева, а maxFeatures — от 6 до 12 и означает максимальное количество используемых признаков при построении. На основе максимизации f-меры после данной процедуры были получены максимальные значения этих параметров: maxDepth — 9, maxFeatures — 11 (рис. 3). В качестве минимального количества примеров, необходимых для создания нового листа (внешнего узла), взято значение, равное 1, а минимальное количество примеров, необходимых для инициации разбиения внутреннего узла, равно 2. В качестве критерия разбиения взята энтропия Шеннона.

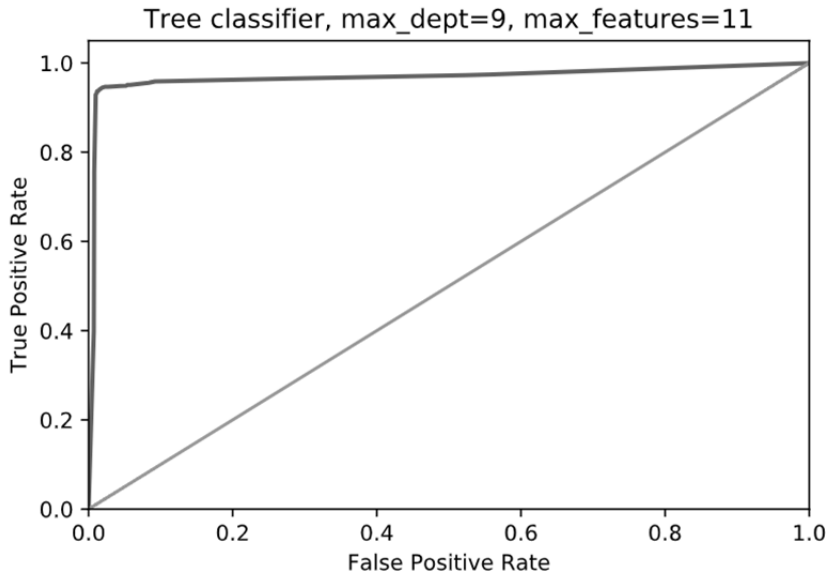

Рис. 3. Roc-кривая для решающего дерева с максимальной глубиной — 9, максимальном количеством используемых признаков — 11

Данная процедура заняла порядка 1,9 с на одном выделенном потоке.

Далее был получен прогноз на отложенной выборке (оставшиеся 30 % данных) и посчитаны метрики [1], результаты которых приведены в таблице 2.

Таблица 2

Метрики в случае решающего дерева

Метрики	precision	recall	f-score	accuracy
Non-Left	0,98	0,99	0,98	0,97
Left	0,96	0,93	0,95	

Уже на данном относительно простом алгоритме были получены достаточно высокие результаты.

Затем была применена логистическая регрессия. Аналогично была проведена 5-кратная кроссвалидация на сетке значений коэффициента регуляризации (C), которая задалась 300 случайными равномерно распределенными числами от 10^{-2} до 10^{2}. Данная процедура заняла значительно больше времени, около 8,2 с на одном потоке. Причина, по которой время обучения в сравнении с методом решающих деревьев оказалось дольше, заключается, с одной стороны, в переборе параметров, а с другой — в необходимости приведения всех числовых признаков к одному масштабу, что увеличивает число необходимых числовых операций. Лучшее значение C получилось равным 1,46. Метрики приведены в таблице 3.

Таблица 3

Метрики в случае логистической регрессии

Метрики	precision	recall	f-score	accuracy
Non-Left	0,83	0,93	0,88	0,80
Left	0,63	0,37	0,47	

Заметно, что данный алгоритм демонстрирует куда более низкий результат. На рисунке 4 показаны коэффициенты регрессии.

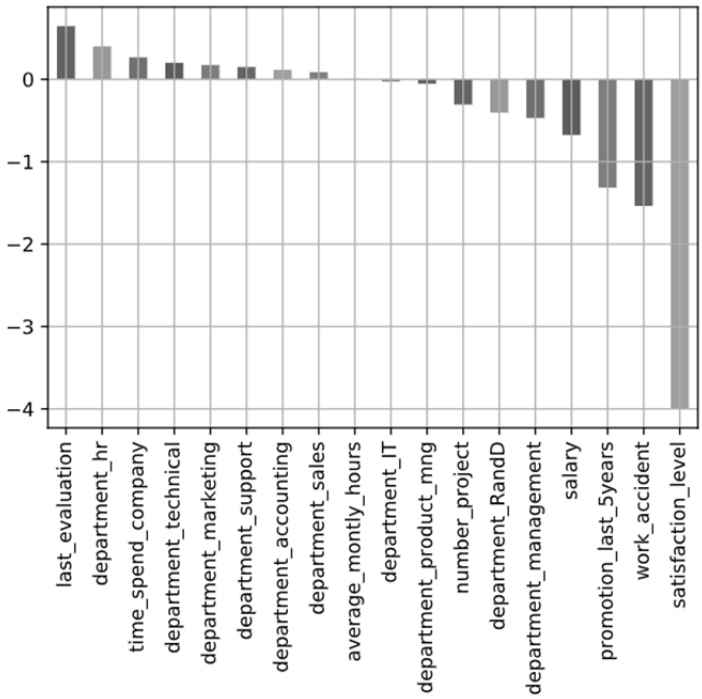

Рис. 4. Коэффициенты логистической регрессии

Коэффициенты показывают интенсивность влияния на результат [6]. Как видно satisfaction_level вносит значительный вклад в отнесении сотрудника к классу Non-Left, что объясняет низкие показатели точности и полноты, то есть модель делает существенное количество ошибок, опираясь на данный признак. Действительно, значение TPR [1] равно 407, TNR — 3198, FPR — 250, FNR — 645 на оставшихся данных из тестовой выборки. Особенно велика ошибка FNR, то есть модель ошибочно полагает большое количество оставшихся сотрудников ушедшими. Для сравнения приведем соответствующие значения для модели, построенной методом решающих деревьев: TPR равно 993, TNR — 3397, FPR — 71, FNR — 39, значения ошибок значительно меньше.

Также видно, что порожденные из признака department бинарные признаки вносят незначительный вклад в классификацию (значение коэффициентов близко к нулю), как и некоторые другие. Однако признак last_evaluation вносит наибольший вклад при отнесении к классу Left.

В финале был использован метод градиентного бустинга на решающих деревьях. В качестве элементарных классификаторов выбраны 100 деревьев с параметрами как при построении модели на одном дереве, но максимальная глубина деревьев ограничена значением 3. Обучение на 5 потоках осуществлено за 2 с, на 1 — приблизительно за 4 с (зависимость времени от количества используемых потоков нелинейная), результаты приведены в таблице 4.

Таблица 4

Метрики в случае градиентного бустинга

Метрики	precision	recall	f-score	accuracy
Non-Left	0,98	0,99	0,99	0,97
Left	0,97	0,93	0,95	

Этот алгоритм на данном датасете показал наилучшие показатели метрик, однако превосходство над результатами в случае использования решающего дерева незначительно. Кроме того, обучение на деревьях куда менее затратно, и на одном выделенном системном потоке транзакция подбора оптимальных параметров оказалась быстрее, чем непосредственное обучение на изначально заданных параметрах градиентного бустинга. Но при использовании деревьев больший риск получить переобученную модель, поэтому подбор параметров все же нужно осуществлять. Получить хорошо работающую модель на зашумленных данных довольно трудно. Градиентный бустинг более гибкий, значительно меньше подвержен переобучению и хорошо работает «из коробки». Соответственно, наиболее оптимальный вариант — использовать его.

Выводы. Была построена модель прогнозирования оттока клиентов, применение которой может облегчить работу в области управления персоналом. Формально данная модель является примером клас-

сической задачи бинарной классификации, и подобный подход может быть распространен и на другие области — например, решение о приеме на работу сотрудника, классификацию страниц резюме на специализированных сайтах (например, таких как *LinkedIn*), вычисление вероятности ответа соискателя с возможность последующего перехода к задаче ранжирования и многие другие.

На исходном наборе данных были опробованы несколько алгоритмов и выбран оптимальный. Стоит отметить, что построенная модель позволяет не только отнести сотрудника к классу ушедших или оставшихся, но и дать вероятность ухода. Хранить обученную модель можно в разных форматах (он зависит от фреймворков машинного обучения или собственных реализаций: есть, например, *Joblib*, для которого существует библиотека на *Python* с одноименным названием [7]), что в свою очередь позволяет поместить ее на удаленном сервере, где будут проводиться вычисления прогноза, а на клиенте выдаваться лишь результаты. Можно адаптировать модель как дополнение к любому имеющемуся настольному приложению (вплоть до «1С: Зарплата и управление персоналом») для HR-менеджеров. Все это дает возможность использовать ее как дополнительный модуль к имеющейся HRM-системе предприятия или компании.

Таким образом, поставленная задача решена.

Список литературы

1. *Гайдышев И.П.* Оценка качества бинарных признаков // Вестник Омского университета. 2016. №1. С. 14—17.

2. *Кузьмичев С.М.* Текучесть кадров — положительный или отрицательный фактор развития организации? // Молодой ученый. 2018. №13. С. 235—238.

3. *Никулин А.А.* Текучесть кадров и ее минимизации // Международный научно-исследовательский журнал. 2015. №1. С. 88—90.

4. *Синяева Л.П.* Текучесть кадров как индикатор адекватности управления предприятием // Концепт. 2013. Спецвыпуск №4. С. 1—7.

5. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных : учеб. пособие. М., 2015.

6. *Шеремет А.Д.* Управленческий учет : учеб. пособие. М., 2009.

7. *Joblib*. URL: https://joblib.readthedocs.io/en/latest/ (дата обращения: 18.03.2019).

8. *Kaggle*. Human Resource. URL: https://www.kaggle.com/kuniowu/human-resource (дата обращения: 09.03.2019).

Об авторе

Валерий Дмитриевич Колесников — магистрант, Балтийский федеральный университет им. И. Канта, Россия.

E-mail: irrmm9000@gmail.com

The author

Valery D. Kolesnikov, Master's Student, I. Kant Baltic Federal University, Russia.

E-mail: irrmm9000@gmail.com