Литература

1. Балашова Е.А. Аутсорсинг как эффективная стратегия управления малым бизнесом сферы сервиса в период кризисных явлений // Пути повышения конкурентоспособности специалистов индустрии моды, туризма и сервиса: Сб. трудов конф. Омск, 2014. С. 49–52.
2. Багирова И.Х. Мотивация персонала в условиях кризиса // Вестник Томского государственного университета. Серия «Экономика». 2011. № 4. С. 84–87.
3. Дробышева В.Г. Адаптивная организационная структура как важнейший фактор конкурентных преимуществ организации предпринимательского типа // Социально-экономические явления и процессы. 2014. Т. 9. № 11. С. 63–71.
4. Кирсанова Е.В. Условия устойчивого функционирования предприятий малого и среднего бизнеса в период экономического кризиса // Вестник Томского государственного ун-та. 2010. № 336. С. 141–143.
5. Мау В. Социально-экономическая политика России в 2014 году: выход на новые рубежи? // Вопросы экономики. 2015. № 2. С. 5–31.
6. Самоукина Н. Эффективная мотивация персонала при минимальных финансовых затратах. М.: Вершина, 2008.
7. Hakanson H. Evolution Processes in Industrial Networks. London: Routledge, 1988. P. 135.

Моделирование оттока кадров в крупной компании с применением технологий интеллектуального анализа данных

Modeling the Outflow of Personnel in a Large Company Using Data Mining Technologies (DOI: 10.34773/EU.2021.3.28)

Л. АБЗАЛИЛОВА

Абзалилова Лия Рашитовна, канд. физ.-мат. наук, доцент кафедры цифровой экономики и коммуникаций Института экономики, финансов и бизнеса Башкирского государственного университета. E-mail: abzalilova.liya@gmail.com

В данной статье рассматриваются методы интеллектуального анализа данных, позволяющие определить факторы, которые влияют на причины увольнения (или неувольнения) сотрудников крупной компании. В связи с тем, что увольнение сотрудников может повлечь большие расходы для бизнеса, вопрос об основаниях, побуждающих к этому, является важным и актуальным. С использованием моделей логистической регрессии, случайного леса, k-средних и модели пропорциональных рисков Кокса были выявлены факторы, влияющее на вероятность увольнения, а также показана их значимость и определены отделы с самой большой текучестью кадров.

***Ключевые слова**: методы машинного обучения, логистическая регрессия, метод k-средних, HR-аналитика.*

This article discusses the methods of data mining that allowed us to determine the factors that affect the reasons for the dismissal (or non-dismissal) of employees of a large company. Due to the fact that the reduction of employees can be associated with large costs for the business, the question of the reasons that prompted this is important and relevant. Using logistic regression models, random forest, k-means, and the Cox proportional risk model, the factors that affect the probability of dismissal were identified, their significance was shown, and the departments with the highest staff turnover were identified.

***Key words**: machine learning methods, logistic regression, k-means method, HR analytics.*

Основные положения

Проведен анализ исходных данных, построены модели логистической регрессии, случайного леса, k-средних и модель пропорциональных рисков Кокса, выявлены факторы, объясняющие причины увольнения сотрудников, выполнена интерпретация результатов моделирования.

Введение

Отток персонала определяется как естественный процесс, в результате которого сотрудники покидают компанию – например, увольняются по личным причинам или выходят на пенсию. Подобное явление – это неизбежная часть любого бизнеса. Придет время, когда тот или иной сотрудник захочет покинуть компанию по личным или профессиональным причинам. Но когда отток персонала превышает определенный порог, это становится причиной для беспокойства. Необходимо знать основные причины того, почему сотрудники решают уйти, а затем принять соответствующие меры, чтобы улучшить производительность компании.

В этом контексте использование моделей классификации для прогнозирования вероятности увольнения[1] сотрудника может значительно повысить способность HR вовремя вмешаться и исправить ситуацию, чтобы предотвратить увольнение.

Целью данного исследования являлось моделирование оттока кадров в крупной компании с применением технологий интеллектуального анализа данных. В рамках сформулированной цели были решены задачи моделирования увольнения сотрудников крупной компании и проинтерпретированы факторы, влияющие на этот процесс.

Методы

Информационной базой для исследования являлся набор данных, который представляет собой опрос сотрудников IBM, показывающий, есть ли увольнение или нет. Набор данных содержит около 1500 записей (табл. 1).

Таблица 1

Описание переменных

Наименование	Описание
Attrition (целевая переменная)	Сотрудник увольняется из компании (0 = нет, 1 = да)
Age	Возраст (числовая величина)
Business Travel	(1 = Командировки отсутствуют, 2 = Командировки часто, 3 = Командировки редко)
Daily Rate	Уровень заработной платы (Числовое значение)
Department	Отдел (1 = Human Resources, 2 = Research & Development, 3 = Sales)
Distance From Home	Расстояние от работы до дома (Числовое значение)
Education	Образование (1 =Ниже колледжа, 2=Колледж, 3 = Бакалавр, 4 = Магистр, 5 = Доктор)
Education Field	Сфера образования (1 = Human Resources, 2 = Life Sciences, 3 = Marketing, 4 = Medical, 5 = Other, 6 = Technical Degree)
Employee Number	ID сотрудника (Числовое значение)
Environment Satisfaction	Удовлетворённость обстановкой (1 = Низкая, 2 = Средняя, 3 = Высокая, 4 = Очень высокая)
Gender	Пол (1 = Female, 2 = Male)
Hourly Rate	Почасовая оплата (Числовое значение)
Job Involvement	Вовлеченность в работу (1 = Низкая, 2 = Средняя, 3 = Высокая, 4 = Очень высокая)
Job Level	Уровень занятости (1 = Очень низкий, 2 = Низкий, 3 = Средний, 4 = Высокий, 5 = Очень высокая)
Job Role	Должность

[1] В данной работе рассматривается «увольнение» по инициативе работника.

Продолжение таблицы 1

Job Satisfaction	Удовлетворенность работой (1 = Низкая, 2 = Средняя, 3 = Высокая, 4 = Очень высокая)
Marital Status	Семейное положение
Monthly Income	Ежемесячный доход (Числовое значение)
Monthly Rate	Ежемесячная ставка (Числовое значение)
NumCompaniesWorked	Число предыдущих мест работы (Числовое значение)
OverTime	Сверхурочные (0 = нет, 1 = да)
Percent Salary Hike	Увеличение заработной платы (%, Числовое значение)
Performance Rating	Рейтинг эффективности
Total Working Years	Число лет работы в целом
Training Times Last Year	Время обучения в прошлом году
Years At Company	Общее число лет, отработанных в компании
Years In Current Role	Число лет в данной должности

В статье были использованы четыре алгоритма машинного обучения: случайный лес, логистическая регрессия, кластеризация k-средних и модель пропорциональных рисков Кокса.

Случайный лес – это метод обучения, используемый для классификации, регрессии и других задач, который объединяет несколько деревьев решений (ансамблей) для выбора наилучшего результата. Алгоритм случайного леса избегает «ошибки» деревьев решений, которые слишком сильно полагаются на обучающий набор, что повышает точность модели [5].

Логистическая регрессия – используется для оценки вероятности события, а также для анализа факторов, влияющих на это событие. Наиболее часто используется бинарный логистический регрессионный анализ [4]:

$$\text{logit } p = \ln \frac{p(y=1)}{1-p(y=1)} = \beta_0 + \beta_1 \cdot x_1 + ... + \beta_n \cdot x_n$$

В рамках рассмотренной задачи зависимая переменная *y* – это вероятность увольнения сотрудников, где 0 означает «нет», а 1 означает «да». Пол, уровень образования, сверхурочная работа, уровень занятости и остальные факторы выступали в качестве переменных-предикторов: $n=34$, $x_1, x_2, ..., x_{34}$; $\beta_1, \beta_2, ..., \beta_{34}$ – это коэффициенты, которые представляют влияние переменных-предикторов на зависимую переменную.

Кластеризация k-средних. Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга [2].

Модель пропорциональных рисков Кокса в общем случае прогнозирует риск наступления события для рассматриваемого объекта и оценивает влияние на этот риск независимых переменных. Риск – функция, зависящая от времени, в данном случае время – количество месяцев/кварталов/лет с момента попадания респондента в группу риска, то есть до момента увольнения [1].

Результаты

Данные содержали 34 переменные, однако некоторые из них не имели смысла для исследования, например, EmployeeCount, MaritalStatus, а также EmployeeNumber, и были исключены.

Используя алгоритм случайного леса, проведена оценка качества используемых предикторов.

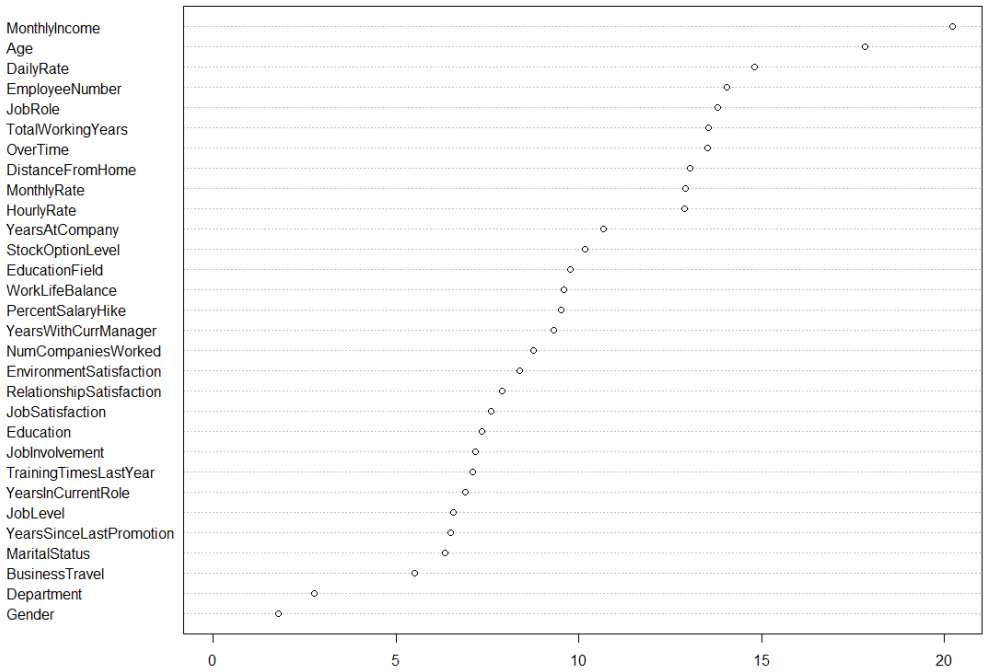

Рис.1. График среднего уменьшения точности

Таблица 2

Результат построения бинарной логистической регрессии

	Оценка коэффициентов
Age	-0,05381 (<0,05)[2]
BusinessTravel:	
BusinessTravelTravel Frequently	2,09331 (<0,05)
BusinessTravelTravel Rarely	0,99154 (<0,05)
DistanceFromHome	0,05798 (<0,05)
JobRoleHuman Resources	1,54693 (<0,05)
JobRoleManager	1,02624 (0,18)
JobRoleManufacturing Director	0,79326 (0,19)
JobRoleResearch Director	-1,04178 (0,20)
JobRoleResearch Scientist	0,71094 (0,20)
JobRoleSales Executive	1,24891 (<0,05)
JobRoleLaboratory Technician	1,72879 (<0,05)
JobRoleSales Representative	2,60624 (<0,05)
JobSatisfaction	-0,4086 (<0,05)
DistanceFromHome	0,05798 (<0,05)
TrainingTimesLastYear	-0,20397 (<0,05)
YearsInCurrentRole	-0,13595 (<0,05)
YearsSinceLastPromotio	0,15742 (<0,05)

[2] В скобках приведены p-значения для проверки гипотезы о значимости факторов.

График среднего уменьшения точности показывает, насколько модель теряет точность при исключении каждой переменной. Чем больше страдает точность, тем важнее переменная для успешной классификации. Переменные представлены по убыванию важности.

Использование логистической регрессии для прогнозирования взаимосвязи между предикторами (характеристиками сотрудников) и прогнозируемой переменной (увольнением сотрудников), где зависимая переменная является бинарной («не уволится» = 0, «уволится» =1), позволило выявить различия между категориями, чтобы понять, какой тип людей с большей вероятностью уволится.

Чтобы определить, какие люди уволятся с большей вероятностью, использовалась кластеризация k-средних, чтобы разделить набор данных на две категории. Первый тип склонен к тому, чтобы уйти, вероятность ухода второго ниже.

Таблица 3

Результат кластеризации методом k-средних

Наименование переменной	1 («уволится»)	0 («не уволится»)
Age	36,59	37,26
Attrition	1,15	1,16
BusinessTravel	2,61	2,60
DailyRate	810,57	794,15
Department	2,25	2,26
DistanceFromHome	9,21	9,16
Education	2,94	2,88
EducatioField	3,29	3,20
EmployeeNumber	1023,66	1026,11
EnvironmentSatisfaction	2,67	2,76
Gender	1,61	1,58
HourlyRate	66,43	65,33
JobInvolvement	2,74	2,72
JobLevel	2,01	2,11
JobRole	5,43	5,47
JobSatisfaction	2,73	2,73
MaritalStatus	2,07	2,11
MonthlyInCome	6282,03	6730,55
MonthlyRate	8217,25	20594,193
NumCompaniesWorked	2,691	0.694
OverTime	1,27	1,30
PercentSalaryHike	15,33	15,08
PerformanceRating	1,67	1,14
TotalWorkingYears	11,01	11,55
TrainingTimesLastYear	2,76	2,83
YearsAtCompany	7,04	6,97
YearsInCurrentRole	2,15	2,23

Рассматривалась зависимость функции выживаемости от времени жизни, то есть зависимость вероятности оттока от продолжительности работы сотрудника [3]. В модели присутствуют две переменные – время работы сотрудника в компании и событие «увольнение»: пока событие не наступило, значение переменной равно нулю, при наступлении – единице. В вычислении коэффициентов регрессии они не участвовали.

Обсуждение

Главным фактором выбытия сотрудников, как показали все модели, является денежный, поскольку наверх вышли переменные «Сверхурочные» и «Ежемесячный доход». Также факторами,

влияющими на вероятность увольнения, являются возраст, доход, удаленность от компании и дома. Обнаружено, что люди, которые работали в 3–4 компаниях, вероятнее всего не уволятся. Кроме того, люди, занимающие более высокие должности, получают более высокий доход, поэтому они с меньшей вероятностью покинут компанию. Более того, удовлетворенность работой также является одной из основных причин, влияющих на коэффициент увольнения сотрудников.

Чтобы оценить производительность моделей, данные были разделены на обучающую и тестовые выборки (табл. 4). Точность для моделей случайного леса, логистической регрессии и пропорциональных рисков Кокса находится на одном уровне, но все они смещены в сторону прогнозирования неувольнения. В то же время метод k-средних лучше предсказывает те случаи, в которых сотрудник планирует уволиться. Существует противоречие между порогом вероятности и количеством сотрудников, которые точно прогнозируются как потенциальные участники оттока. Порог высокой вероятности приведет к большому количеству ошибок. Релевантность для бизнеса заключается в том, чтобы хорошо предсказать сокращение сотрудников, а не его отсутствие, поэтому выбирается более низкий порог вероятности. Модель пропорциональных рисков Кокса представляется наиболее удачной, поскольку она имеет высокий AUC и лучшую матрицу ошибок.

Таблица 4

Сравнение моделей

	AUC	Оценка (recall)	Точность (accuracy)	Специфичность (specificity)
Случайный лес	0,79	0,99	0,85	0,84
Логистическая регрессия	0,78	0,87	0,87	0,86
Кластеризация методом k-средних	0,50	0,16	0,49	0,87
Модель пропорциональных рисков Кокса	0,81	0,93	0,88	0,87

Заключение

Приведенные результаты моделирования позволяют заключить, что итоги исследования соответствуют поведению людей в реальном мире и предыдущим исследованиям [6]. Используя случайные леса решений и кластеризацию методом k-средних, были определены наиболее значимые факторы, влияющие на увольнение. Модели логистической регрессии и пропорциональных рисков Кокса дали схожий результат – значимы «Сверхурочные», «Ежемесячный доход» и «Удовлетворенность работой». Качество построенных моделей проверено специфическими тестами и ROC-анализом.

Литература

1. Груздев А.В. Регрессия Кокса или модель пропорциональных рисков // Независимый проект «Корпоративный менеджмент». 2012 URL: http://www.cfin.ru/management/strategy/plan/cox_regression.shtml

2. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. URL: http://www.machinelearning.ru

3. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. Of 23 International Conference on Machine Learning, Pittsburgh, PA, 2006.

4. David W. Hosmer, Lemeshow S. Applied logistic regression [M]. New York: Wiley, 2000.

5. Pal M. Random forest classifier for remote sensing classification [J]. International journal of remote sensing, 2005, 26(1): 217-222.

6. Shenghuan Y., Pradeep R., Timothy S. IBM Employee Attrition Analysis [Electronic resource]. URL: https://www.researchgate.net/publication/346578445_IBM_Employee_Attrition_ Analysis