Разработка скоринговой модели для оценки вероятности отчисления студентов вуза

Я.И. Славянова, Д.Г. Лагерев
yanchos7@gmail.com | LagerevDG@mail.ru
Брянский государственный технический университет

Работа большинства информационных систем предусматривает обработку данных, их накопление в ходе эксплуатации и последующий анализ. Однако анализ такого большого объема информации человеком невозможен без их предварительной автоматической обработки. С этой целью используется Data Mining, включающий в себя описательное и предсказательное моделирование. Задача классификации является одной из наиболее понятных человеку технологий анализа данных и относится к предсказательному моделированию. Данная задача состоит в разделении множества наблюдений на классы на основе их формального описания. Одним из методов решения задачи классификации является логистическая регрессия, в то время как распространенной областью применения является скоринг. В данной статье рассматривается применение скоринга к решению задачи оценки вероятности отчисления студентов из вуза на основании данных о проявленных ими посещаемости и успеваемости. Решение данной задачи позволит кураторам групп, направлений и другим заинтересованным лицам вовремя идентифицировать тенденцию к отчислению, выделить группу риска среди студентов и принять заблаговременные меры для того, чтобы спрогнозированное построенной моделью событие не стало фактом. Построенная скоринговая модель подлежит публикации в виде веб-сервиса для дальнейшего применения в программном комплексе поддержки работы преподавателя вуза. В данном случае на вход модели поступают агрегированные характеристики, полученные на основе аккумулированных программным комплексом данных об успеваемости и посещаемости студентов, с помощью которых на выходе получается интегрированный показатель вероятности наступления события, а именно отчисления. В результате построения скоринговой модели выполняется последующая оценка её качества.

Ключевые слова: *Data Mining, задача классификации, скоринг, анализ успеваемости и посещаемости студентов, аналитическая платформа.*

Development of the scoring model for assessing the probability of expulsion of university students

Y.I. Slavyanova, D.G. Lagerev
Bryansk State Technical University

The work of most information systems involves the processing of data, its accumulation during operation and subsequent analysis. However, the analysis of such a large amount of information by a person is impossible without its preliminary automatic processing. For this purpose, Data Mining is used, which includes descriptive and predictive modeling. The statistical classification is one of the most understandable data analysis technologies for humans and relates to predictive modeling. This task consists in dividing the set of observations into classes based on their formal description. One of the methods for solving the classification problem is logistic regression, while scoring is a common area of application. This article discusses the application of scoring to the problem of assessing the probability of students' expulsion from the University based on data on their attendance and academic performance. The solution of this problem will allow curators of groups, directions and other interested parties to identify the tendency to expulsion in time, identify a risk group among students and take early measures to prevent the event predicted by the built model from becoming a fact. The built scoring model is subject to publication as a web service for further use in the software package for supporting the work of a University teacher. In this case, the model input receives aggregated characteristics obtained from accumulated data on student performance and attendance by the software package, which results in an integrated indicator of the probability of an event, namely, deductions. As a result of building a scoring model, a subsequent assessment of its quality is performed.

Keywords: *Data Mining, statistical classification, scoring, students' performance and attendance analysis, analytical platform.*

1. Введение

Сложно представить информационную систему, не оперирующую данными. Объемы аккумулируемых данных настолько внушительны, что человек не в состоянии проанализировать «сырые» данные без предварительной обработки, однако такой анализ необходим в связи со знаниями, заключенными в этих данных, владение которыми полезно при принятиях решений. С целью проведения автоматического анализа данных используется Data Mining.

Data Mining («добыча данных», интеллектуальный анализ данных) представляет собой методологию и процесс обнаружения в больших массивах данных, накапливающихся в информационных системах, ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Примером информационной системы, требующей автоматизированный анализ данных, является программный комплекс поддержки работы преподавателя вуза, который позволяет автоматизировать деятельность преподавателя за счет единой системы учета данных об успеваемости и посещаемости студентов вуза, доступной на устройствах различных операционных систем [1].

Data Mining включает в себя две категории моделирования, а именно описательное и предсказательное. К описательному моделированию относятся задачи ассоциации и кластеризации, в то время как к предсказательному – задачи классификации и регрессии. Описательное моделирование проще в плане построения моделей и требований к знаниям

аналитика, однако предсказательное моделирование является более эффективным.

2. Задача классификации

В искусственном интеллекте и машинном обучении задача классификации представляет собой задачу разделения множества наблюдений на группы, называемые классами, на основе анализа их формального описания с использованием обучения с учителем, поскольку классы определяются заранее и для примеров обучающего множества метки классов заданы. Аналитические модели, решающие задачу классификации, называются классификаторами.

Задача классификации представляет собой одну из базовых задач прикладной статистики и машинного обучения, а также искусственного интеллекта в целом, поскольку классификация является одной из наиболее понятных и простых для интерпретации технологий анализа данных, а классифицирующие правила могут быть сформулированы на естественном языке.

Если число классов ограничено двумя, то имеет место бинарная классификация, к которой могут быть сведены многие более сложные задачи. Бинарные классификационные модели являются более понятными и интерпретируемыми. При их использовании удается упростить модель и снять некоторые ограничения, связанные с большим числом возможных состояний выходной переменной.

3. Скоринг

Скоринг представляет собой процедуру ранжирования объектов в соответствии с их измеренными характеристиками. Скоринг использует статистическую модель, оценивающую вероятность того, что наступит определенное событие. На вход такой модели подаются определенные характеристики, а на выходе формируется интегрированный показатель, который определяет вероятность наступления события.

Скоринговая карта представляет собой набор характеристик рассматриваемого объекта и присваиваемых им весовых коэффициентов, выраженных в некоторой балльной шкале (таблица 1). В результате обработки собранных сведений объекту начисляется определенное количество скоринговых баллов.

Таблица 1: Простейшая скоринговая карта

Характеристика	Атрибут	Балл
Пол	мужской женский	5 9
Возраст	< 22 ≥ 22	2 12
Социальный статус	пенсионер, студент, занятое население	4 10
Трудовой стаж	< 3 ≥ 3	6 8
Состоит ли в браке	да нет	10 2

Основным алгоритмом для расчета баллов скоринговых карт является логистическая регрессия.

3.1. Этапы разработки скоринговой модели

Разработка скоринговой модели включает в себя следующие этапы:

1) подготовка данных;
2) идентификация события;
3) моделирование (включая сэмплинг, двумерный анализ, расчет баллов или весов и оценку качества);
4) сравнение моделей и выбор оптимальной модели;
5) интеграция в систему принятия решений и тестовая эксплуатация;
6) промышленное использование и мониторинг.

3.2. Подготовка данных

Для разработки скоринговых моделей необходима историческая выборка данных – обучающая выборка. От репрезентативности этой выборки зависит точность оценок параметров модели скоринга и, соответственно, предиктивная мощность скорингового алгоритма. Репрезентативность выборки определяется тем, насколько полно в ней присутствуют положительные и отрицательные прецеденты.

Подготовка данных включает в себя:

1) получение данных из внешних и внутренних, свободных и коммерческих источников;
2) расчет производных показателей и агрегатов;
3) аудит и профайлинг данных;
4) решение проблем с качеством данных, выявленных на этапе аудита (пропуски, выбросы, экстремальные значения).

3.3. Идентификация событий

Идентификация события является необязательным этапом и требует ответа на следующие вопросы.

1. Как фиксировать событие?
2. В течение какого времени должно произойти событие?
3. Помогает ли выбранное событие решить поставленную задачу?

Событием при этом считается некоторый факт, например, пациент болен или здоров, заемщик вернул кредит или допустил просрочку, клиент лоялен или не лоялен, страховой случай или не страховой и т.д.

3.4. Моделирование

Подготовив данные для моделирования и определившись с тем, что считать событием, можно перейти непосредственно к моделированию скоринговой модели, которое включает в себя четыре подэтапа.

3.4.1. Сэмплинг

Прежде чем передавать на вход скоринговой модели данные, необходимо произвести сэмплинг – процесс отбора из исходной совокупности данных выборки, представляющей интерес для анализа.

При реализации сэмплинга используются специальные методы отбора, которые должны обеспечить репрезентативность выборки с точки зрения решаемой задачи. Сэмплинг можно разделить на два типа:

1) традиционный:
   a) случайный;
   b) равномерный;
   c) стратифицированный;
2) специальный:
   a) удаление мажоритарного класса;
   b) дублирование миноритарного класса.

В исходных совокупностях бизнес-данных часто присутствует несбалансированность классов, то есть классы в целевой переменной представлены неравномерно. В этом состоит проблема редкого класса, который в задачах бинарной классификации назначают событием, а второй – не-событием.

Для решения этой проблемы используются балансировка, при которой происходит либо удаление мажоритарного класса (рис. 1), либо дублирование миноритарного класса (рис. 2).

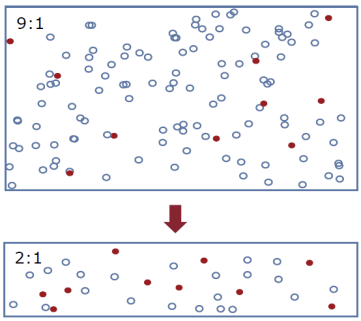

Рис. 1. Удаление мажоритарного класса

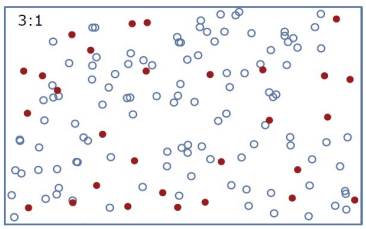

Рис. 2. Дублирование миноритарного класса

Использование данных методов позволяет восстановить равновесие с целью получения более сбалансированного обучающего множества.

3.4.2. Двумерный анализ

Двумерный анализ позволяет одновременно исследовать взаимоотношения двух переменных (входной и выходной), и в той или иной форме проверять гипотезы о причинных связях между ними.

Конечные классы позволяют уменьшить число значений исходного набора данных за счет их объединения в пределах некоторого интервала с использованием информации о бинарной выходной переменной [2].

Конечные классы предназначены для решения следующих задач:

1) снижение разнообразия значения признаков без ущерба для информативности данных;
2) снижение размерности данных за счет исключения признаков с низкой значимостью;
3) восстановление пропусков;
4) борьба с выбросами и экстремальными значениями;
5) упрощение описания исследуемых объектов.

Процедура сокращения уникальных значений признака состоит из следующих шагов:

1) формирование начальных классов;
2) «сжатие» начальных классов в меньшее количество интервалов, называемых конечными классами.

Для формирования конечных классов используется метод WoE-анализа или совокупность доказательства, где каждому наблюдению, содержащему набор признаков, ставится в соответствие бинарная выходная переменная.

Затем производится разбиение всего диапазона изменения того или иного признака на несколько начальных классов, для каждого из которых вычисляется коэффициент WoE:

$$WoE_i = ln\frac{N_i/N}{P_i/P} \quad (1)$$

1) i – индекс начального класса;
2) N_i– число не-событий в группе;
3) N – общее число не-событий;
4) P_i– число событий в группе;
5) P – общее число событий.

Если значение категории совпадает с событием большее число раз, чем с не-событием, то согласно формуле, под знаком логарифма будет значение меньше 1, что делает его отрицательным.

WoE < 0 указывает на большую вероятность появления события, а WoE > 0 – не-события.

Индекс WoE – количественная мера предиктивной силы отдельной категории внутри переменной.

На основе коэффициентов WoE вычисляется величина, определяющая значимость признака в модели бинарной классификации, называемая информационным индексом (IV-индексом):

$$IV = \sum_{i=1}^{K}\left\{\left(\frac{N_i}{N} - \frac{P_i}{P}\right)\cdot WoE_i\right\} \quad (2)$$

Информационный индекс всегда является положительной величиной. На основе IV-индекса определяется значимость признака по следующей методике:

1) IV < 0,02 – отсутствует;
2) 0,02 ≤ IV < 0,1 – низкая;
3) 0,1 ≤ IV < 0,3 – средняя;
4) IV > 0,3 – высокая.

Коэффициенты WoE и вычисленные на их основе значения IV-индекса являются критерием для

формирования конечных классов оптимальным образом.

Логистическая регрессия представляет собой разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (регрессорами или предикторами) и зависимой переменной [3]. С помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого.

Поскольку множественная регрессия не «знает», что переменная отклика бинарная по своей природе, вместо предсказания бинарной переменной, предсказывается непрерывная переменная со значениями на отрезке [0,1] при любых значениях независимых переменных. Это достигается применением следующего регрессионного уравнения (логит-преобразования):

$$p = \frac{1}{1 + e^{-y}} \quad (3)$$

1) p – вероятность того, что произойдет интересующее событие;
2) e – основание натуральных логарифмов 2,71...;
3) y – стандартное уравнение множественной регрессии: $y = F(x_1, x_2, \dots, x_n)$.

Специальные виды сэмплинга нарушают распределение целевой переменной. Смещенная модель может завышать вероятности наступления события. Для использования модели, построенной на сбалансированной выборке, необходимо провести процедуру её калибровки (рис. 3).

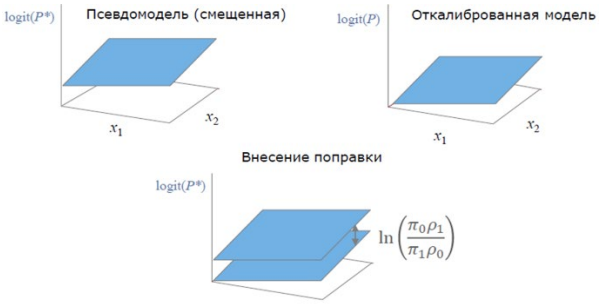

Рис. 3. Графическое представление процесса калибровки

Проверку качества и обобщающей способности модели на тестовом множестве следует проводить на выборке, которая имеет соотношение событий и не-событий, соответствующее условиям, в которых в дальнейшем будет эксплуатироваться модель.

3.4.3. Расчет баллов или весов

Превращение уравнения логистической регрессии в формат балльной скоринговой карты называется масштабированием.

Преимуществами балльной скоринговой карты являются:

1) легкость интерпретации баллов;
2) причины отказа объяснимы бизнес-терминами;
3) прозрачность получения результата;
4) развитые методики оценки качества и мониторинга балльных карт.

По умолчанию масштабирование осуществляется в стандарт FICO® Risk Scores компании «Fair Isaac», диапазон баллов в котором от 0 до 850 и три показателя, влияющие на результат масштабирования, выглядят следующим образом [4]:

1) количество баллов, которое удваивает шансы наступления (не-наступления) события – PDO равно 40 баллам;
2) отношение событий Odds к не-событиям равно 72:1;
3) значение шкалы Score, в котором достигается заданное отношение шансов «не-событий» к «событиям» Odds равно 660.

Чаще всего встречаются следующие параметры масштабирования:

1) Score = 600; Odds = 50:1; PDO = 20;
2) Score = 600; Odds = 30:1; PDO = 20;
3) Score = 500; Odds = 32:1; PDO = 50.

Сейчас FICO широко известен и повсеместно применяется в США и Канаде при принятии решений о выдаче кредитов.

3.4.4. Оценка качества

Существуют различные показатели, позволяющие оценить качество смоделированной скоринговой карты. Одним из них является ROC-кривая и связанный с ней индекс AUC.

ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров.

В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. При этом предполагается, что у классификатора имеется некоторый параметр, варьируя который, получается то или иное разбиение на два класса. Этот параметр часто называют порогом отсечения. В зависимости от него будут получаться различные величины ошибок I и II рода.

На рис. 4 представлена матрица ошибок, которая строится на основе результатов классификации моделью и фактической принадлежностью примеров к классам.

Предсказанный класс	Фактический класс: Плохой	Фактический класс: Хороший
Плохой	TP	FP
Хороший	FN	TN

Рис. 4. Матрица ошибок

Каждое поле данной матрицы отображает полученный исход:

1) TP (True Positives) – верно классифицированные положительные примеры;
2) TN (True Negatives) – верно классифицированные отрицательные примеры;
3) FN (False Negatives) – положительные примеры, классифицированные как отрицательные;

4) FP (False Positives) – отрицательные примеры, классифицированные как положительные.

Ложный пропуск является ошибкой I рода, когда интересующее событие ошибочно не обнаруживается. Это ложное обнаружение, в свою очередь, является ошибкой II рода, так как при отсутствии события ошибочно выносится решение о его присутствии.

Индекс AUC – площадь под ROC-кривой (рис. 5). Его значения варьируются от 0,5 до 1. Чем больше показатель AUC, тем лучшей прогностической силой обладает построенная модель.

Рис. 5. Графическое представление индекса AUC

В литературе иногда приводится следующая экспертная шкала для значений AUC, по которой можно судить о качестве модели:

1) 0,5 – 0,6 – неудовлетворительное;
2) 0,6 – 0,7 – среднее;
3) 0,7 – 0,8 – хорошее;
4) 0,8 – 0,9 – очень хорошее;
5) 0,9 – 1,0 – отличное.

4. Модель анализа успеваемости и посещаемости студентов

Естественным этапом развития программного комплекса поддержки работы преподавателя вуза является обработка и последующий анализ аккумулируемых данных.

Среди целевых пользователей данного программного комплекса возникла необходимость в оценке вероятности отчисления студентов на основании данных о проявленных ими успеваемости и посещаемости, информация о которых хранится в базе данных программного комплекса, и информации об уже отчисленных студентах, а также о студентах, благополучно завершивших свое обучение в вузе. В данном случае событием можно считать отчисление студента, поскольку данное событие является редким. Заблаговременное определение таких событий позволит вовремя идентифицировать группу риска среди студентов и минимизировать процент отчисленных студентов за счет своевременного принятия превентивных мер кураторами групп и направлений.

Для решения данной задачи было принято решение о построение аппликационной скоринговой модели, представленной на рис. 6.

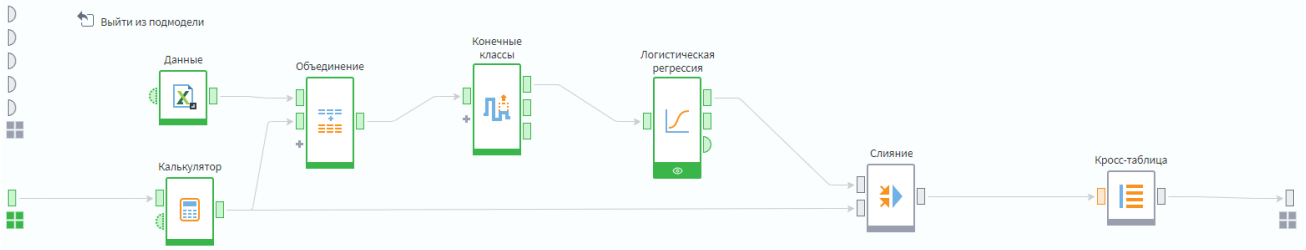

Рис. 6. Построенная скоринговая модель

Построение скоринговой модели было выполнено в аналитической платформе «Loginom», которая позволяет реализовать все аналитические процессы и настроить их логику без какого-либо программирования. Данная платформа поддерживает интеграцию с различными источниками данных, их последующую обработку, моделирование, визуализацию и интерпретацию полученных результатов и последующие развертывание и интеграцию для построенных моделей [5].

В качестве характеристик объектов (студентов) выступают:

1) средняя оценка, полученная за расчетно-графические и курсовые работы (проекты);
2) процент лабораторных и практических работ, оценка за которые ниже «хорошо»;
3) средняя продолжительность выполнения лабораторных и практических работ;
4) средняя продолжительность выполнения расчетно-графических и курсовых работ (проектов);
5) процент пропусков лекционных занятий, совершенных по неуважительной причине;
6) количество групп, в которых побывал студент за время своего обучения в вузе.

Разбиение начальных классов на конечные в компоненте «Конечные классы» выполнено таким образом, чтобы обеспечивалась высокая значимость каждой характеристики.

Построенная скоринговая модель была опубликована в виде веб-сервиса для последующего осуществления взаимодействия с программным комплексом посредством архитектурного стиля взаимодействия компонентов распределенного приложения в сети – REST. Так, на вход опубликованной модели поступает уникальный идентификатор студента и рассчитанные характеристики, содержащиеся в теле POST-запроса, в то время как в тело ответа на данный запрос приходит уникальный идентификатор студента, предсказанное моделью событие и вероятность его наступления.

Визуализатор «Отчет по регрессии» отображает статистические параметры и результаты статистических тестов для анализа регрессионных моделей. Также данный визуализатор содержит данные о значимости характеристик, рассчитанной в ходе проведения логистической регрессии, где:

1) характеристики №1 и №2 – средняя значимость;
2) характеристики №4 и №5 – низкая значимость;
3) у характеристик №3 и №6 значимость отсутствует.

Полученные значения значимости свидетельствуют о том, что требуется скорректировать используемые в модели характеристики.

Визуализатор «Качество бинарной классификации» позволяет проанализировать показатели характеристик качества для обучающей и тестовой выборок построенной модели. Визуализатор также содержит индекс AUC ROC > 0,7, статистика KS > 40; индекс Джинни > 40, которые свидетельствуют о хорошем качестве модели.

5. Заключение:

В данной статье были рассмотрены задача классификации и процедура скоринга, позволяющая ранжировать объекты в соответствии с их измеренными признаками.

Исходя из полученных сведений и потребности в автоматизированном анализе данных программного комплекса поддержки работы преподавателя вуза, была построена модель аппликационного скоринга, позволяющая осуществлять классификацию студентов на основании обучающих и тестовых выборок, сформированных из собранных системой данных.

Рассчитанные характеристики качества показывают, что построенная модель обладает хорошим качеством. В дальнейшем она будет использована в рамках программного комплекса для решения поставленной задачи, однако требуется доработка передаваемых на вход модели характеристик.

Поскольку данные в программном комплексе постоянно обновляются и дополняются, в качестве перспективы развития можно рассматривать работу скоринговой модели в фоне, а также получение кураторами групп и направлений уведомлений или рассылки для постоянного мониторинга текущей ситуации для курируемых студентов.

Литература:

[1] Калевко В.В., Лагерев Д.Г., Подвесовский А.Г. Программный комплекс «Автоматизированное рабочее место преподавателя» // Сборник науч. трудов II Международной науч. конференции и XII Международной науч.-практ. конф. «Современные информационные технологии и ИТ-образование» 24-26 ноября 2017 г. М.: Лаборатория открытых информационных технологий факультета ВМК МГУ им. М. В. Ломоносова, 2017. С. 197-205. [Электронный ресурс]. – Режим доступа: https://www.elibrary.ru/item.asp?id=32661960.

[2] Паклин Н.Б. Оптимальное квантование для повышения качества бинарных классификаторов // Искусственный интеллект. – 2013. – В 4. – С. 392-399.

[3] Hosmer D. W., Lemeshow S. Applied Logistic Regression (2nd Edition) // Wiley Publishing, Inc., 2000.

[4] Кочеткова В.В., Ефремова К.Д. Обзор методов кредитного скоринга // Juvenis Scientia. – 2017. – № 6. – С.22-25.

[5] Аналитическая платформа «Loginom» [Электронный ресурс]. – Режим доступа: https://loginom.ru/.

Об авторах:

Славянова Я.И. – магистрант по направлению «Программная инженерия» ФГБОУ ВО БГТУ.

Лагерев Д.Г. – к.т.н., доц. кафедры «ИиПО» ФГБОУ ВО БГТУ. E-mail: LagerevDG@mail.ru.