УДК 003.26+004.57

К.А. Ахраменок, маг.;
Н.А. Жиляк, доц., канд. техн. наук
(БГТУ, г. Минск)

ПОСТРОЕНИЕ СКОРИНГОВОЙ МОДЕЛИ С ИСПОЛЬЗОВАНИЕМ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Скоринг (от англ. Scoring — подсчет очков в игре) — это модель классификации клиентской базы на различные группы, если неизвестна характеристика, которая разделяет эти группы, но известны другие факторы, связанные с интересующей нас характеристикой [1].

Логистическая регрессия — статистическая модель для по-

строения скоринговых моделей при бинарной классификации. Математическая модель логистической регрессии выражает зависимость логарифма шанса (логита) от линейной комбинации независимых переменных [2].

Цель работы: описать алгоритм построение скоринговой модели с использованием логистической регресии.

Математическая модель логистической регресии выглядит следующим образом:

$$\ln\left(\frac{p_i}{1-p_i}\right)=b_0+b_1x_i^{(1)}+b_2x_i^{(2)}+\ldots+b_kx_i^{(k)}+\varepsilon_i,$$

где p_i — вероятность наступления искомого события на основе известных параметров i-го исследуемого объекта; $x_i^{(j)}$ — значение j-ой независимой переменной ; b_0 — независимая константа модели, b_j — параметры модели; ε_i — компонент случайной ошибки [3].

Для построениея скоринговой модели с использованием логистической регресии требуется выполнить следующие действия:

1. подготовить набор данных для построения модели;
2. преобразовать количественные и категориальные переменные с использованием WoE;
3. отсеять не значимые независимые перенные используя IV;
4. построить логистическую регрессию используя подготовленные данные.

Для преобразования количественных и категориальных данных используется метод весомости признака (каждый признак будет заменяться его весом) [2]. Метод весомости признака:

$$WoE=[\ln(\frac{\mathrm{Re}\,lativeFrequencyOfGoods}{\mathrm{Re}\,lativeFrequencyOfBads})],$$

где RelativeFrequencyOfGoods — количество «положительных» исходов для выбранной группы; RelativeFrequencyOfBads — количество «отрицательных» исходов для выбранной группы.

Для вычисления WoE количественных переменных следует выполнить следующие шаги:

1. разбить переменные на 10 групп (или меньше);
2. вычислить количество «положительных» и «отрицательных» исходов для каждой группы;
3. вычислить процент «положительных» и «отрицательных» исходов для каждой группы;
4. вычислить WoE.

Для категориальных переменных, не требуется разбирать данные на группы, можно сразу перейти к шагу 2.

IV(information value) — мера определения значимости переменных и измерения разницы в распределении «отрицательных» и «положительных» исходов:

$$IV = \sum (nonEvents\% - events\%) * WOE$$

Значения данного коэффициента трактуют следующим образом:

- менее 0,02 — статистически незначимая переменная;
- 0,02 - 0,1 — статистически малозначимая переменная;
- 0,1 - 0,3 — статистически значимая переменная;
- 0,3 и более — статистически сильная переменная.

На основе полученных значений IV следует удалить из выборки незначимые переменные [3].

После выполнения всех вышеперечисленных шагов, мы получим подготовленный набор данных.

Для построения логистической регресии можно возьпользоваться языком Python и библиотекой sklearn.liner_model. Потребуется разбить набор данных на обучающий и тестовый, обучить модель и проверить ее на тестовых данных [2]. В случае получения удовлетворительных результатов, остается извлеч коэффициенты регресии и встроить их в существующую систему или использовать построенную регрессию для дальнейших предсказаний.

Актуальность работы заключается в необходимости исследования новых методов построения скоринговых моделей.

Научная новизна заключается в том, что в текущий реалях огромное количество банковский скоринговых моделей строиться вручную и имеет довольно слабую предсказательную способность.

В результате работы был описан способ построения скоринговой модели с использованием логистической регресии, а также были описаны способы обработки данных и выявления их значимости.

ЛИТЕРАТУРА

1. The credit scoring toolkit: Theory and Practice for retail Credit Risk Management and Decision Automation/ Anderson R. [et al] // England: Oxford University Press. – 2007. – P. 8–13, 16–17.

2. Practical Guide to Logistic Regression / Josef M. Hibe [et al] // Chapman and Hall/CRC, USA. – 2015. – P. 30–35.

3. Applied logistic regression / David W. Hosmer // Wiley-Interscience Publication , USA. – 2000. – P. 302–305.