МОДЕЛИ «БИНАРНОГО ВЫБОРА» В БАНКОВСКОМ АНАЛИЗЕ

Вобленко И.И.

Воробьев Н.А.

Захаров Д.И.

Шабров Т.С.

АГУ МосПолитех

Москва, Россия

[ilya.voblenko@gmail.com](mailto:ilya.voblenko@gmail.com)

**Аннотация.** В данной статье предлагается алгоритм анализа на основе моделей дискретного выбора, позволяющая прогнозировать использование кредита, выданного банком, с целью оптимизации процесса составления клиентской базы кредитного отдела банка. Предметом исследования является модель «бинарного выбора» для прогнозирования факта использования клиентом выдаваемого банком кредита с целью составить максимально оптимизированную клиентскую базу.В ходе исследования были применены методы многомерного статистического анализа данных и математический аппарат моделей дискретного выбора. Были рассмотрены механизм идентификации, а также оценки параметров logit- и probit-моделей на основе метода максимального правдоподобия. Были предложены методы оценки значимости параметров регрессионных уравнений и методы оценки качества модели. Проведенный анализ единой совокупности статистических данных при помощи logit- и probit- моделей позволил сделать вывод о наибольшем влиянии оценочных показателей, имеющих более высокие коэффициенты значимости. Моделирование показало возможности применения данных моделей в поиске количественных закономерностей развития изучаемого процесса.

**Ключевые слова:** модели бинарного выбора, logit-модель, probit-модель, оценка, метод максимального правдоподобия.

**ВВЕДЕНИЕ**

Современный мир требует от банков принятия быстрого и качественного решения при подборе клиентов для выдачи кредитов. Сам процесс отбора клиентов является весьма трудоемкой задачей, так как требует досконального анализа клиента по многим критериям, ведь одной из основных задач кредитных отделов банков является минимизация рисков при выдаче кредитов и минимизация страховых бюджетов, требуемых государством от банков[1]. Самым простым и дешевым способом автоматизации принятия решения о выдачи клиенту кредита будет создание алгоритма анализа на основе моделей дискретного выбора. Так как в нашем случае рассматривается проблема использования выданного кредита в будущем, то оптимальным выбором являются модели «бинарного выбора», а именно logit- и probit- модели.

**ПОСТАНОВКА ЗАДАЧИ**

Оценка предположения об использовании клиентом банка кредита может быть задана в виде переменной с двумя качественными индикаторами ({“используется”}, {“не используется”}). Для бинарных моделей необходимо использование численной интерпретации этих качественных индикаторов: yi = 1 – “используется”, yi = 0 – “не используется”.

В аналитической деятельности банка по оценке клиента крайне важна степень точности полученного предсказания и понимание степени важности конкретных критериев при принятии решения. Поэтому перед нами стоит задача нахождения максимально точной модели получения прогноза, а также расчет “весов” влияния конкретных критериев на полученное прогнозное значение для выявления главных признаков с их степенью влияния и отсечения лишних признаков.

Для решения данных задач нами предлагается следующий алгоритм оценки использования клиентом кредита, выданного банком.

**АЛГОРИТМ ОЦЕНКИ**

Разработанный нами алгоритм состоит из 7 этапов:

1. Определение значений индикаторной переменной на основе механизмов комплексного оценивания.
2. Анализ данных. Удаление из данных выбросов и линейно-зависимых векторов. Отсеивание статистически незначимых признаков на основе комплексного анализа проблемы, тепловой карте корреляции признаков.
3. Построение моделей бинарного выбора (logit, probit).
4. Оценка неизвестных параметров на основе метода максимального правдоподобия. На основе значений статистической характеристики p-значимость определяется уровень значимости коэффициентов модели(статистика Вальда).
5. Расчет и сравнение показателей качества модели: pseudoR^2, метрика, характеризующая долю правильных ответов алгоритма (Accuracy), метрика точности(precision), метрика полноты(recall) и метрика F-меры (Далее используется их английская версия названия).
6. Построение ROC-кривой для анализа доли корректных прогнозов.
7. Выбор оптимальной модели и подбор для нее границы разделения классов.

Рассмотрим механизм оценки с использованием моделей бинарного выбора. Обозначим использование кредита y, и будем считать, что y = 1 в случае использования выданного кредита и y = 0 в противном случае. Набор множества факторов, оказывающих влияние на значение индикаторной переменной, обозначим вектором x = (x1, x2, …, xn). Тогда модель бинарного выбора в общем случае примет вид [2]:

(1.1)

, где f^() – функция распределения случайной величины, область значений которой лежит в диапазоне от 0 до 1. При использовании функции стандартного нормального распределения модель будет называться probit(1.2), а при использовании стандартной логистической функции распределения – logit(1.3). W – вектор оцениваемых параметров модели.

(1.2)

(1.2) – функция стандартного нормального распределения.

(1.3)

(1.3) – функция стандартная логистическая распределения.

Функции стандартного логистического и стандартного нормального распределений представлены на рисунке (Рисунок 1):

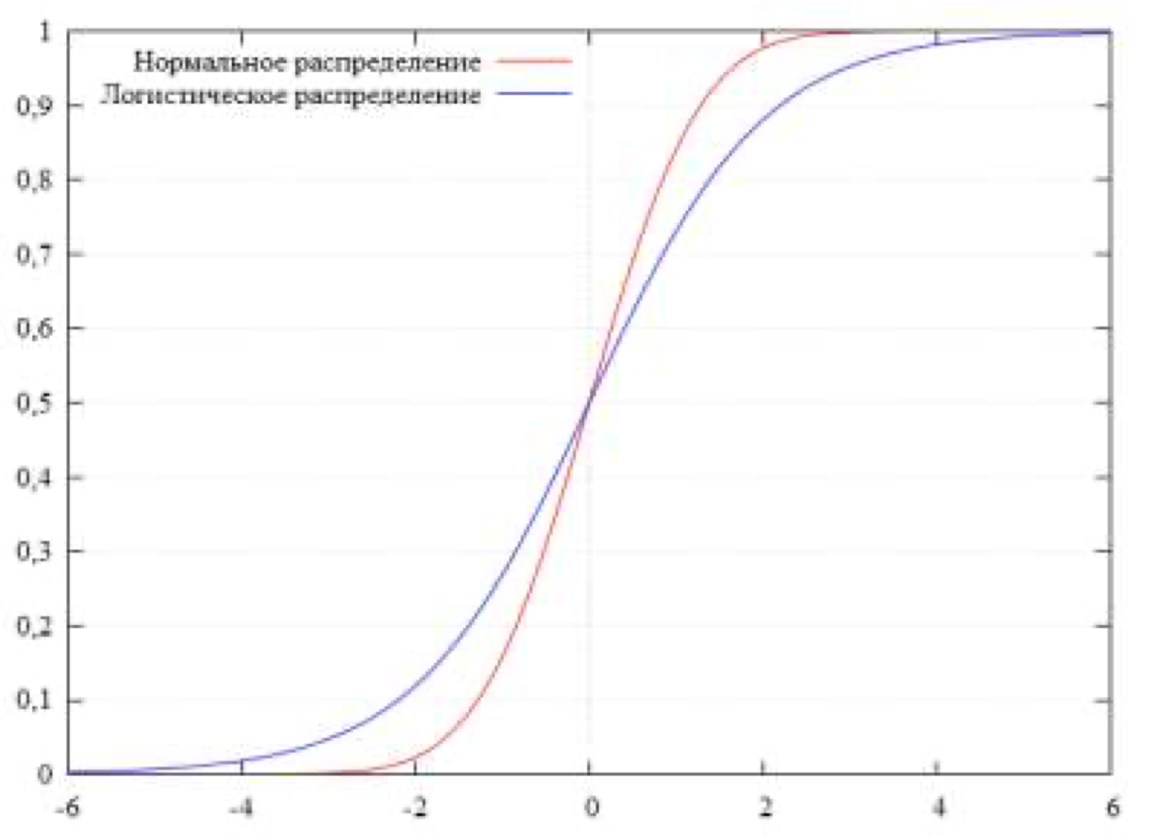


Рис. 1 Нормальное и логистическое распределение.

Для оценки вектора w воспользуемся методом максимального правдоподобия. Для оценки параметров рассчитаем функцию правдоподобия. Вероятность P(yi = Yi|xi1, xi2, … xin) – это вероятность того, что индикаторная переменная равна определенному значению при определенных признаках x. В общем виде данная вероятность запишется следующим образом P(y=Y|X) ,и будет рассчитываться по формуле (1.4) [2]:

Для получения оценок W необходимо максимизировать функцию правдоподобия. Чтобы упростить процесс оценивания неизвестных параметров следует воспользоваться свойством монотонного возрастания функции натурального логарифма и преобразовать с помощью данного свойства уравнение (1.4) в следующее выражение (1.5):

После чего стоит сменить задачу максимизации на задачу минимизации путем добавления “–“ перед суммой сумм логарифмов и воспользоваться методом градиентного спуска для нахождения минимума полученной функции.

Оценку значимости коэффициентов построенной модели произведем на основе P-значения. Для этого рассчитаем стандартную ошибку и статистику Вальда. Стандартные ошибки коэффициентов равны квадратным корням диагонали ковариационной матрицы, на основе которых и рассчитывается статистика Вальда. Расчет осуществляется по формуле(1.6):

(1.6)

, где w – оцениваемые коэффициенты модели, SE – стандартная ошибка. Полученные значения в статистике Вальда выше 0,05 указывает статистическую незначимость коэффициентов. [3]

На следующем этапе произведем оценку качества модели на основе следующих метрик: pseudoR^2, accuracy, precision, recall и F-меры.

Псевдо коэффициент детерминации показывает качество подобранной модели и рассчитывается по формуле (1.7):

(1.7)

Для расчета этих метрик зададим обозначения используемых величин: TP – правильно предсказанные моделью истинные значения индикаторной переменной, FP – неправильно предсказанные истинные значения индикаторной переменной, FN – неправильно предсказанные ложные значения индикаторной переменной, TN – правильно предсказанные ложные значения индикаторной переменной.

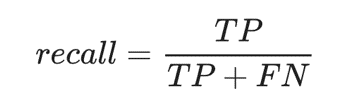
Accuracy характеризует долю правильных ответов алгоритма и рассчитывается по формуле (1.8):

(1.8)

Precision характеризует долю объектов, которые модель отнесла к положительным и при этом действительно являющиеся положительными. Рассчитывается данная метрика по следующей формуле (1.9):

(1.9)

Recall характеризует долю определенных моделью положительных объектов в массе всех положительных объектов. Нахождение данной метрики производится по формуле (1.10):

 (1.10)

F-мера есть среднее гармоническое precision и recall, которое достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю. Вычисляется данная метрика по формуле (1.10):

(1.11)

В нашем случае вес точности метрики ß был выбран равным единице. [4]

Далее для анализа доли корректных прогнозов построим ROC-кривую и рассчитаем значение ROC-AUC по формуле (1.12):

(1.12)

Чем выше значение ROC-AUC, тем качественнее полученная модель. [5]

**ОСНОВНЫЕ РЕЗУЛЬТАТЫ**

В результате было получено, что наилучшей моделью оказалась probit. Итоговая модели имеет вид:

P = -6.6512 + 0.0278\*Income + 0.3645 \* Family + 0.0612 \* CCAvg + 0.8123 \* Education – 0.503 \* SecAcc + 2.0188 \* CDAcc – 0.3725 \* Online – 0.6031 \* CrCard

, где Incom – годовой доход клиента(тыс. $), Family – размер семьи клиента, CCAvg – средние расходы по кредитным картам (тыс. $), Education – уровень образования(1- студент, 2 – выпускник, 3-продвинутый/профессионал),  SecAcc – наличие у клиента счета в ценных бумагах в банке, CDAcc – наличие у клиента депозитного счета в банке, Online – использование клиентом услуг интернет-банка, CrCard – использование клиентом кредитной карты.

На основе полученных метрик pseudoR^2 = 0.584, Accuracy = 0.96, Precision= 0.83, Recall = 0.66, F1-мера = 0.74 можно характеризовать полученную модель как хорошую.

**ЗАКЛЮЧЕНИЕ**

Важным итогом данной работы является разработка алгоритма получения модели оценки использования кредита клиентом в будущем, что позволит оптимизировать полученную в данной работе модель под конкретные потребности банка при составлении клиентской базы. Также стоит отметить, что в ходе выполнения работы было установлено, что probit-модели дают более качественное предсказание. Помимо этого, было установлено, что такие показатели как индекс домашнего адреса, возраст и профессиональный опыт являются не влияющими на результат прогнозирования признаками.

**ЛИТЕРАТУРА**

1. Е.А. Гришина, Е.А. Киреева, С.Б. Коваленко, Г.Ж. Курдюмова, Е.В. Травкина Организация банковского кредитования : учебное пособие для О64 студентов направления подготовки 38.03.01 Экономика (направленность (профиль) программы «Банковское дело») – Саратов : Саратовский социально-экономический институт (филиал) РЭУ им. Г.В. Плеханова, 2018. – 152 с.
2. Носко В.П. Эконометрика для начинающих (Дополнительные главы). **–** М**.:** ИЭПП**,** 2005. С. 379.
3. Горошко И.В Разработка алгоритма оценки результатов деятельности органов внутренних дел с использованием моделей бинарного выбора / И.В. Горошко, Ш.Х. Гонов // Моделирование, оптимизация и информационные технологии.. – 2018 год. – Т. Том 6, №2. – С. 368 стр.- 378 стр.
4. Tom Fawcett An introduction to ROC analysis/ Tom Fawcett // Pattern Recognition Letters – 2006 year. – Volume 27, №8 – С. 861 p.- 874 стр.
5. <https://habr.com/ru/company/ods/blog/328372/>