

Введение Сложная ситуация, сложившаяся с кредитованием физических лиц, побуждает кредитные организации более ответственно подходить к выдаче кредитов. Для принятия обоснованного решения о выдаче кредитов применяется скоринг (от англ. scoring - подсчет очков), являющийся инструментом классификации клиентской базы на две группы: клиенты, которым можно выдать кредит и клиенты, кредитование которых рискованно [1]. В основе скоринга лежит предположение о наличии связи добросовестности заемщика с его показателями социального статуса и уровня финансовой обеспеченности (наличие детей, уровень образования, место работы, доходы и др.). На практике скоринг-тест состоит из двух основных частей. Первая часть представляет собой опросник, отражающий социальные характеристики клиента, например: семейное положение; постоянство и уровень дохода; наличие финансовых обязательств; качество кредитной истории и т.п. (табл. 1) [2-4]. Показатели каждой характеристики раздельно ранжируются по их значимости для кредитоспособности заемщика. Вторая часть скоринг-теста включает метод классификации заемщиков на два класса: клиенты, которым можно выдать кредит; и клиенты, кредитование которых рискованно. Таким образом, задачу скоринга можно рассматривать как задачу дискриминантного анализа, т.е. построения классификатора, с некоторым уровнем гарантии, на основе имеющейся выборки, который позволяет судить о принадлежности объекта x определенному классам Y_1 или, Y_2 . Построение классификатора При наличии в кредитной организации достаточной информационной базы по выданным кредитам, для построения скоринговой системы необходимо определиться с алгоритмом классификации. Известно, что ни один из описанных методов не может быть признан «самым лучшим» во всех случаях. Только сопоставление предсказания и факта может дать оценку эффективности скоринговых моделей [5]. Одним из наиболее простых классификаторов является классификатор Байеса [6]. Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации. Таблица 1 - Пример опросной анкеты банка

| Параметр (Количество допустимых значений) | Допустимые значения |
|---|---|
| Семейное положение заемщика (2) | Женат (замужем) / одинок(а) |
| Наличие иждивенцев (2) | Есть / нет |
| Карьерный уровень должности клиента (4) | Рабочий / служащий / управленец среднего звена / топ-менеджер |
| Личный доход, тыс. руб. (6) | До 10 / 11-25 / 26-50 / 51-75 / 76-100 / более 100 |
| Семейный доход, тыс. руб. (6) | До 10 / 11-25 / 26-50 / 51-75 / 76-100 / более 100 |
| Наличие у клиента платежных обязательств (2) | Есть / нет |
| Кредитная история клиента в данном банке (3) | Нет / Положительная / С задержками |
| Кредитная история клиента в других банках (3) | Нет / Положительная / С задержками |
| Имущество (4) | Нет / дом / квартира / автомобиль |
| Клиент Банка (2) | Да / нет |
| Возраст (5) | 18-25 / 26-40 / 41-50 / 51-60 / > 60 |
| Своевременный возврат кредита заемщиком (2) | Да / нет |
| Наивный | |

классификатор Байеса (Naive Bayes Classifier) основан на предположении о независимости отдельных компонент вектора x , описывающего социальные характеристики потенциального заемщика (признаки объекта). Для классификации используется формула вычисления апостериорных вероятностей $P(Y_1|x)$, $P(Y_2|x)$. При этом вероятность $P(Y_i|x)$ принадлежности объекта x классу Y_i , в силу предположения независимости компонент, вычисляется по формуле $P(Y_i|x) = P(Y_i|x_1) \cdot P(Y_i|x_2) \cdot \dots \cdot P(Y_i|x_n)$, где x_1, x_2, \dots, x_n - компоненты вектора признаков x . Объект x будет отнесен к тому классу Y_i , для которого выше $P(Y_i|x)$. Однако, при формальном применении этой формулы для целей кредитного скоринга при незначительном отличии вероятностей $P(Y_1|x)$, $P(Y_2|x)$, велика вероятность ошибки ложного отнесения заемщика к категории надежных. Причин этого может быть несколько: отсутствие независимости компонент вектора признаков объекта, статистический характер корреляционной связи, невозможность полного учета всех факторов, влияющих на возврат кредита заемщиком, и др. В связи с этим построим модифицированный классификатор Байеса на основе обучающей выборки, включающей параметры, перечисленные в табл. 1. В рамках предоставления кредитных продуктов для кредитора оказывается предпочтительнее не выдать кредит платежеспособному клиенту, чем предоставить неблагонадежному. В этом случае сравнивается отношение апостериорных вероятностей с некоторой эмпирически заданной величиной (постоянная отсечения) C , то есть рассматривается соотношение: $P(Y_1|x)/P(Y_2|x) > C$. Например, необходимо, чтобы мера того, что заемщик благонадежный, в 2 раза превышала меру того, что он неплатежеспособен. Тогда C принимается равной 2. Качество классификатора зависит от объема выборки. Чем больше объем, тем точнее предсказание. Однако, обработка большой выборки может быть проблематичной с точки зрения объема вычислений. В этой связи встает задача определения минимального объема обучающей выборки и параметра C , при которых достигаются приемлемые результаты классификации. Анализ надёжности классификатора и подбор параметров

Для оценки качества бинарной классификации очень широко применяются ROC-кривые - графики в декартовой системе координат xOy , где по оси Ox откладывается показатель специфичности алгоритма классификации, а по оси Oy - показатель чувствительности. Специфичность алгоритма классификации определяется как доля ошибочных положительных классификаций в общем числе отрицательных событий (ложно отрицательное множество), а чувствительность - как доля верно классифицированных положительных событий в общем количестве положительных классификаций (истинно положительное множество). Очевидно, что ROC-кривая располагается в единичном квадрате (рис. 1). ROC-кривая, в частности, оценивает качество классификации величиной площади под ней. Чем больше площадь под ROC-кривой, тем выше качество классификатора, для которого она построена. C

помощью ROC-кривых можно визуально оценить качество классификатора. Рис. 1 - Пример ROC-кривой, соответствующей Байесовскому классификатору

Построение ROC-кривой производится в программных пакетах, таких, как R, SPSS и др. [7]. В результате этого исследователь лишен понимания механизма влияния параметров классификатора на ROC-кривую. Поэтому для подбора параметров классификатора и определения необходимого для достижения приемлемых результатов классификации объема обучающей выборки, ROC-кривые, на наш взгляд, не являются оптимальным решением. В этой связи построим метод, позволяющий осуществлять подбор параметров классификатора на основании на корреляционной связи между количеством верных предсказаний скорингового алгоритма и общим числом испытаний,.

Рассмотрим признаки: $X = \{\text{рекомендация скоринг-теста}\}$ с возможными значениями «X+» и «X-» соответственно при положительной и отрицательной рекомендации по выдаче кредита; признак $Y = \{\text{реальный исход кредитной истории}\}$ с возможными значениями «Y+» и «Y-» соответственно при положительной и отрицательной ситуации погашения задолженности заемщиком. Таким образом, исходные результаты для анализа надежности скоринг-теста можно представить в виде следующей таблицы: Таблица 2 - Результаты анализа обучающей выборки

| $X \setminus Y$ | Y+ | Y- | Σ |
|-----------------|-----|-----|-----------|
| X+ | a | b | a+b |
| X- | c | d | c+d |
| Σ | a+c | b+d | a+b+c+d=n |

Здесь $a+b+c+d=n$ - объем обучающей выборки, a, d - количество верно проклассифицированных объектов, b, c - количество ложно проклассифицированных объектов. Для анализа корреляционной связи между дихотомическими признаками используются коэффициенты контингенции и ассоциации. Коэффициент контингенции вычисляется по формуле: (1)

Коэффициент ассоциации вычисляется по формуле: (2) Пусть k - выборочное, а K - генеральное значение коэффициента ассоциации или контингенции. Статистическая значимость коэффициента k при нулевой ($H_0: K=0$) и альтернативной ($H_1: K \neq 0$) гипотезах проверяется на основе t -статистики Стьюдента с фактическим значением: (3) Под устойчивостью коэффициентов k к соотношению пропорций a, b, c, d будем понимать независимость коэффициента от значений a и d при фиксированной их сумме. Исследуем устойчивость коэффициентов k_a, k при изменении параметра a от 0 до 80 и фиксированных значениях $n=100, b=c=10, a+d=80$. Из графика зависимостей k_a, k и $T_{k,a}$ от параметра a (рис. 2), видно что: - при фиксированной сумме a+d изменение соотношения между параметрами a и d приводит к значительным колебаниям коэффициента контингенции - от -0.1 до 0.6 и коэффициента ассоциации - от -1 до 0.88; - экспериментальное значение статистики T_k изменяется от 1 до 7.4 и T_a - от $-\infty$ (при $k_a=-1$) до 18.56. Для уровня значимости $\alpha=0.05$ в интервале $5 \leq a \leq 75$ справедливо неравенство $|T_{k,a}| > T_{\text{crit}} = 1.99$, следовательно, значение генерального коэффициента $K_{k,a} > 0$ и нулевая гипотеза отвергается. Таким образом, можно сделать вывод, что коэффициенты ассоциации и контингенции

являются недостаточно устойчивыми и в реальных задачах затруднительна однозначная их интерпретация и, следовательно, возможны неадекватные выводы об изучаемом явлении. Рассмотрим коэффициент согласованности, который имеет вид [8, 9]: (4) и равен доле совпадений показаний признаков X и Y. Для генерального коэффициента согласованности K_s доверительный интервал определяется соотношением: (5) где предельная ошибка выборки (6) Здесь k_s - выборочный коэффициент согласованности, α доверительная вероятность, $\Phi(\tau)$ - значение функции Лапласа. В равенстве (6) аргумент τ находится по таблице значений функции Лапласа из условия $\Phi(\tau) = \alpha / 2$. (7) В отличие от коэффициентов контингении и ассоциации, коэффициент согласованности является устойчивым к изменению пропорций параметров a и d . Действительно, нетрудно видеть, что выражение $z = k_s(1 - k_s)$ определяет график квадратичной параболы с максимальным значением $z = 0,25$ при $k_s = 0,5$ и, следовательно, из равенства (6) получается оценка $n \tau^2 / 4\Delta^2$ (8) для необходимого объема выборки n , которое обеспечивает доверительный интервал (5) для генерального коэффициента согласованности K_s при заданных доверительной вероятности α и точности Δ . Рис. 2 - Зависимость коэффициента k_a, k и статистик T_a, k от параметра a Параметры a, b, c, d связаны с C через алгоритм классификации. Максимизация коэффициента согласованности дает условие для подбора оптимального параметра C . Таким образом, задача отыскания оптимального параметра C имеет вид: (9) Параметр C может быть определен с помощью компьютерного эксперимента, заключающегося в отыскании значений a, b, c, d и расчете коэффициента согласованности при различных C из интервала (C_{\min}, C_{\max}) . За оптимальное принимается то значение параметра C , при котором коэффициент согласованности достигает своего максимального значения.

Выводы По сравнению с существующими методами анализа надежности скоринговых тестов, метод, основанный на применении коэффициента согласованности в качестве доли (процента) совпадений рекомендаций скорингового алгоритма с фактическим результатом погашения кредита, является универсальным и математически обоснованным методом исследования надежности скоринговых алгоритмов. Данный метод обладает рядом существенных преимуществ: 1) простотой вычисления и наглядностью интерпретации надежности k_s в качестве доли (процента) совпадений показаний признаков X и Y; 2) устойчивостью при изменении пропорций между параметрами a и d при постоянстве суммы $a+d$; 3) наличием доверительного интервала (5) для оценки надежности K_s на генеральной совокупности и, следовательно, его оценки с необходимыми значениями доверительной вероятности α и точности Δ . 4) возможностью однозначного численного сравнения надежности различных скоринговых алгоритмов.