РАЗДЕЛ XXIII. МЕНЕДЖМЕНТ И УПРАВЛЕНИЕ ОРГАНИЗАЦИЕЙ**Жиеру Ю.С.****Разработка скоринговой модели логистической регрессии при наличии экспертных ограничений***Национальный исследовательский ядерный университет «МИФИ»
(Россия, Москва)**doi: 10.18411/sr-10-04-2021-97***Аннотация**

Кредитный скоринг можно опередить как метод начисления потенциальным заемщикам определенного количества баллов на основе информации из его анкетных данных, кредитной истории, социально-демографическом положении. Цель создания скоринга — предсказание случаев, связанных с отсутствием у заемщика намерения исполнять обязательства по кредиту. В основе оценки кредитоспособности могут лежать различные математические модели, и каждый банк выбирает наиболее оптимальные в зависимости от имеющихся в распоряжении банка ресурсов.

Ключевые слова: кредитный скоринг, логистическая регрессия, машинное обучение, валидация, калибровка модели.

Abstract

Credit scoring can be outlined as a method of assigning a certain number of points to potential borrowers on the basis of information from their profile data, credit history, socio-demographic situation. The purpose of scoring is the prediction of cases when the borrower has no intention to fulfill his obligations on the credit. The assessment of creditworthiness can be based on different mathematical models, and each bank chooses the most optimal ones depending on the resources available to the bank.

Keywords: credit scoring, logistic regression, machine learning, validation, model calibration.

1. Обзор существующих моделей

В основе оценки кредитоспособности могут лежать различные математические модели, и каждый банк выбирает наиболее оптимальные в зависимости от имеющихся в распоряжении банка ресурсов.

Описательные модели уделяют внимание сути зависимостей в наборе данных, взаимному влиянию различных факторов, улучшению понимания анализируемых данных, т.е. построению эмпирических моделей различных систем. Ключевой момент в таких моделях — легкость и прозрачность результатов для восприятия человеком. Возможно, обнаруженные закономерности будут специфической чертой именно конкретных исследуемых данных и больше нигде не встретятся, но это все равно может быть полезно и потому должно быть известно.

К описательным моделям относятся:

- 1) регрессионные модели — описывают функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме, зависимость между непрерывными числовыми параметрами, зависимость между категориальными параметрами;
- 2) модели кластеризации — описывают группы (кластеры), на которые можно разделить объекты, данные о которых подвергаются анализу;

- 3) модели исключений – описывают исключительные ситуации в записях, которые резко отличаются чем-либо от основного множества записей;
- 4) итоговые модели – выявление ограничений на данные анализируемого массива.

Решение предсказательных задач разбивается на два этапа. На первом этапе на основании набора данных с известными результатами строится модель. На втором этапе она используется для предсказания результатов на основании новых наборов данных. При этом, естественно, требуется, чтобы построенные модели работали максимально точно.

Рассмотрим более подробно некоторые из них.

Линейный дискриминант Фишера:

Модель использует метод максимального правдоподобия и основана на предположении о том, что функции правдоподобия классов многомерные нормальные. Основной недостаток в том, что такая гипотеза слишком сильная и не всегда выполняется.

Логистическая регрессия:

Модель напрямую оценивает вероятности принадлежности объектов классам, используя принцип максимума правдоподобия. Строится скоринговая функция на основе логит-функции, веса функции находятся по методу наименьших квадратов. Модель основана на более слабых предположениях, чем предыдущий метод, поэтому часто работает лучше.

Логистическая регрессия является самым популярным методом в задаче кредитного скоринга из-за своей простоты и эффективности.

Метод k ближайших соседей:

Данная непараметрическая модель является одной из самых простых и понятных. Объект относится к классу, к которому относятся k ближайших к нему объектов выборки. Близость определяется по заранее выбранной метрике на пространстве данных об объектах.

Основной недостаток модели – необходимость полного просмотра набора объектов, когда нужно проклассифицировать новый. Кроме того, выбор правильной метрики является достаточно сложной задачей.

Деревья решений:

Модель легко интерпретируема и может строиться на основе небольших выборок. Здесь не строится линейная скоринговая функция, вместо этого берется функция одного аргумента (обычно этот аргумент – это значения определенного признака) и на ее основе объекты последовательно разделяются так, чтобы группы максимально отличались по уровню риска. Дерево строится до тех пор, пока новое разделение не приведет к статистически значимому различию в величине кредитного риска. Каждый лист построенного дерева в итоге соответствует классу.

После этого множество деревьев, построенных различными разделяющими функциями, тестируются на выборке и выбирается лучшее.

Генетические алгоритмы:

Эта модель не накладывает стандартных ограничений на целевую функцию, таких как гладкость, выпуклость и так далее. В рассматриваемой задаче генетические алгоритмы генерируют начальное множество скоринговых функций, после чего к ним применяются «мутации», «скрещивания», отбрасывание непригодных функций.

Нейронные сети:

Данная модель берет за основу принцип работы биологических нейронных сетей и их строение. Нейронные сети способны моделировать сложные нелинейные зависимости, хотя для задачи кредитного скоринга это не всегда необходимо и чаще применяется в скоринге компаний. Модели имеют достаточно сложную топологию

(множество слоев разного типа, различные функции активации) и веса связей, которые получаются в результате обучения модели, не имеют никакой интерпретации для задачи кредитного скоринга. Значит, нет возможности объяснить полученное предсказание и провести анализ значимости факторов

2. Математическая постановка задачи

Пусть Ω - множество кандидатов на получение кредита в банке. Каждому потенциальному заемщику $\omega \in \Omega$ ставится в соответствие его признаковое описание $x = (x_1(\omega), \dots, x_n(\omega)) \in X = R^n$, а так же значение зависимой переменной $Y = \{-1; +1\}$. Зависимая переменная описывает класс заемщика, метка +1 соответствует классу «хороших», а метка -1 – классу «плохих» клиентов, из-за которых банк понесет убытки.

В задаче кредитного скоринга требуется построить такую модель, которая по признаковому описанию x потенциального заемщика будет предсказывать значение зависимой переменной y . Предполагаем, что на множестве $X \times Y$ существует какое-то вероятностное распределение и его плотность обозначим $p(x, y)$. Из этого множества выбирается так называемая обучающая выборка $\Omega^* = \{(x(\omega_i), y(\omega_i))\}_{i=1}^m$ – это набор случайно и независимо выбранных наблюдений, т.е. описания заемщиков. Именно на основе этой выборки будет строиться модель. Для решения поставленной задачи обычно используется принцип максимума апостериорной вероятности. Апостериорная вероятность того, что заемщик принадлежит «хорошему» или «плохому» классу, может быть посчитана с использованием формулы Байеса:

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} \quad (1)$$

При решении задачи бинарной классификации условное распределение зависимой переменной представляет собой распределение Бернулли:

$$p(y|x, w) = \text{Ber}(y|\text{sigm}(w^T x)), \quad (2)$$

где $\text{sigm}(w^T x) = \frac{1}{1+e^{-w^T x}}$ – логистическая функция (сигмоида), а $w^T x = \sum_{i=1}^n w_i x_i = 1$ – линейная разделяющая гиперплоскость.

Тогда вероятности того, что заемщик принадлежит к классам «плохих» и «хороших» равны соответственно:

$$P(y = +1|x, w) = \text{sigm}(w^T x) \quad (3)$$

$$P(y = -1|x, w) = 1 - \text{sigm}(w^T x) \quad (4)$$

Можно записать это одним выражением:

$$p(y|x, w) = \text{sigm}(yw^T x) = \frac{1}{1+e^{-yw^T x}} \quad (5)$$

Поскольку наблюдения в обучающей выборке независимы, поэтому функция правдоподобия будет выглядеть следующим образом:

$$L(w|\Omega^*) = \prod_{i=1}^m p(y_i|x_i, w) = \prod_{i=1}^m \text{sigm}(y_i \sum_{j=1}^n w_j x_{ij}) \quad (6)$$

Используя принцип максимума правдоподобия, получаем оценку вектора параметров:

$$\hat{w}_{ML} = \arg \max_w \left\{ \prod_{i=1}^m \text{sigm}(y_i \sum_{j=1}^n w_j x_{ij}) \right\} \quad (7)$$

Прологарифмируем функцию правдоподобия (15) и будем решать задачу минимизации:

$$\hat{w}_{ML} = \arg \min_w \left\{ -\sum_{i=1}^m \ln(\text{sigm}(y_i \sum_{j=1}^n w_j x_{ij})) \right\} \quad (8)$$

В соответствии с (8) оценивают коэффициенты в классической нерегуляризованной логистической регрессии. Предположим, что априорной плотностью распределения параметра является нормальное распределение с нулевым матожиданием и дисперсией σ . В модели дисперсия будет являться случайной величиной.

Тогда совместное распределение вектора параметров будет иметь вид:

$$p(w|r) \propto \prod_{i=1}^n \left[\left(\frac{1}{r_i} \right)^{\frac{1}{2}} \exp \left(- \sum_{i=1}^n \frac{w_i^2}{2r_i} \right) \right] \quad (9)$$

Параметры с малым значением r_i могут быть удалены из модели, а остальные параметры будут называться релевантными. Здесь r_i – является гиперпараметром модели.

Попробуем величины, обратные дисперсиям, использовать в качестве штрафных функций. Тогда предполагаем, что априорная плотность распределения величин обратных дисперсиям является гамма-распределением:

$$p\left(\frac{1}{r} | a, \beta\right) \propto \prod_{i=1}^n \left[\left(\frac{1}{r_i} \right)^{a-1} \exp \left(- \frac{\beta}{r_i} \right) \right] \quad (10)$$

Из (18) видно, что обратная дисперсия зависит от двух параметров гамма-распределения α, β . Для облегчения процесса подбора параметров предположим, что они являются функциями от одного и того же параметра μ .

Для случайной величины, имеющей гамма-распределение, известно:

1. $E\left(\frac{1}{r_i}\right) = \frac{a}{\beta}$ – математическое ожидание;
2. $Var\left(\frac{1}{r_i}\right) = \frac{a}{\beta^2}$ – дисперсия.

Рассмотрим отношение $\frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} = \frac{1}{\sqrt{\alpha}}$:

Если $\frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \rightarrow 0$, то значит все распределения дисперсий r_i сконцентрированы возле математического ожидания.

Тогда можно сказать, что оцененные дисперсии практически фиксированы и равны единице при $\alpha \cong \beta$.

Если $\frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \rightarrow 1$, то априорные распределения практически равномерны.

При $r_i \rightarrow 0$: $\ln r_i \rightarrow -\infty$ и критерию выгодно уменьшать все дисперсии. Но в этом случае невозможно выполнить ограничения, предписывающие достаточно хорошо приближать обучающую совокупность. Из-за этого противоречия критерий проявляет ярко выраженную склонность к чрезмерной селективности отбора признаков, подавляя большинство из них, в том числе и релевантные.

Получается, что необходимо выполнение следующих требований:

$$\mu \rightarrow 0 \Rightarrow \begin{cases} E\left(\frac{1}{r_i}\right) \rightarrow 1 \\ Var\left(\frac{1}{r_i}\right) \rightarrow 0 \\ \frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \rightarrow 0 \end{cases} \quad \mu \rightarrow \infty \Rightarrow \begin{cases} E\left(\frac{1}{r_i}\right) \rightarrow \infty \\ Var\left(\frac{1}{r_i}\right) \rightarrow \infty \\ \frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \rightarrow 1 \end{cases} \quad (11)$$

Одним из наборов функций, удовлетворяющих требованиям, является:

$$\alpha = 1 + \frac{1}{2\mu} \text{ и } \beta = \frac{1}{2\mu} \quad (12)$$

С учетом всех предположений об априорных распределениях вектора параметров и гиперпараметров получаем следующую оценку вектора параметров:

$$\hat{w}_{MAP} = \arg \max_w \{p(y|w, x)p(w|r)p(r|\alpha, \beta)\} \quad (13)$$

$$\hat{w}_{MAP} = \arg \min_w \{-\ln p(y|w, x) - \ln p(w|r) - \ln p(r|\alpha, \beta)\} \quad (14)$$

Тогда получим следующий критерий обучения:

$$J(w, r) = - \sum_{i=1}^m \ln \left(\text{sigm} \left(y_i \sum_{j=1}^n w_j x_{ij} \right) \right) + \frac{1}{2} \sum_{i=1}^n \frac{w_i^2}{r_i} + (1 + \frac{1}{\mu}) \sum_{i=1}^n \ln r_i + \frac{1}{2\mu} \sum_{i=1}^n \ln r_i \rightarrow \min \quad (15)$$

Этот критерий будем называть моделью логистической регрессии.

1. Sorokin, A.S. *Primenenie zakonov raspredelenija sluchajnyh velichin dlja modelirovanija jekonomicheskikh javlenij i processov* [Tekst] : monografija. / N.Ja. Bambaeva, A.C. Sorokin – M.: MJeSI, 2010. – 156 с. – ISBN 978-5-7764-0612-6 2.
2. Kovalev, M., Korzhenevskaja, V. *Metodika postroenija bankovskoj skoringovoj modeli dlja ocenki kreditosposobnosti fizicheskikh lic* [Tekst] // Banki Kazahstana. – 1. –2008. – с. 43–48. 3. Nivorozhkina, L.I. *Jekonometricheskoe modelirovanie riska nevyplat po potrebitel'skim kreditam*. [Tekst] // Prikladnaja jekonometrika. –30 (2). – 2013. s. 65– 76. 4.
3. Sorokin, A.S. *K voprosu ocenki soglasovannosti mnenij jekspertov pri ispol'zovanii metodov jekspertnogo ocenivaniya v kreditnom skoringe*. [Tekst] /A.C. Sorokin // Rol' biznesa v transformacii obshhestva – 2014: Sb. st. po mat. IX mezhdunar. nauchn. kongr. – M.: «Jeditus», 2014. – с. 281-283. – ISBN 978-5-00058-089-9 5.
4. Ulitina, E.V. *Statistika: uchebnoe posobie* [Tekst] / E.V. Ulitina, O.V. Ledneva, O.L. Zhirnova – M.: Moskovskij finansovo-promyshlennyj universitet «Sinergija», 2013. – 320 с. – ISBN: 978-5-4257-0107-7

Носкова Н.Ю.

Децентрализованное бюджетирование

ФНПЦ АО «НПО «Марс»

(Россия, Ульяновск)

doi: 10.18411/sr-10-04-2021-98

Аннотация

В статье установлены принципы децентрализованного бюджетирования. Определены исходные единицы бюджетного процесса при реализации данной модели бюджетирования на предприятии. Исследованы особенности построения модели и на основе полученных результатов определен эффект от ее внедрения на предприятии.

Ключевые слова: децентрализованное бюджетирование, бизнес-единица, центр затрат, бюджетный процесс, бюджет

В настоящее время одним из эффективных инструментов управления деятельностью предприятия является система бюджетирования, которая позволяет:

- реализовывать стратегические планы;
- моделировать возможные состояния в будущем и планировать действия по достижению поставленных целей;
- контролировать соответствие производимых хозяйственных операций целям организации;
- контролировать исполнение принятых стратегических и оперативных решений;
- координировать деятельность различных подразделений и предприятия в целом [1, С. 137].

Процесс формирования и исполнения бюджета организации носит комплексный характер, взаимосвязанными частями которого являются:

- технология составления бюджетных планов, определяющая структуру и состав бюджетов, порядок составления бюджетов отдельных участников группы и консолидированного бюджета предприятия в целом;