МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ БИННИНГА ДЛЯ ЗАДАЧИ МОДЕЛИРОВАНИЯ КРЕДИТНОГО РИСКА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Шерера Даниила Владимировича

Научный руководитель
доцент, к. ф.-м. н. _____ Н. Ю. Агафонова

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2022

Введение. В современном мире кредиты являются неотъемлемой частью жини многих людей, которым часто приходится брать кредит на различное дорогостоящее имущество и услуги из-за невозможности оплатить их полную стоимость.

Одной из актуальных проблем кредитующих финансовых организаций является оценка кредитоспособности заемщиков. Оценка риска невозврата кредита осуществляется на основе двух подходов - субъективного заключения экспертов или автоматизированных систем скоринга.

Под кредитным скорингом понимается система оценки кредитоспособности (кредитных рисков) потенциального заёмщика, в основе которой заложены численные статистические методы. Кредитный скоринг широко используется как крупными банками, микрофинансовыми организациями, так и в потребительском (магазинном) экспресс-кредитовании на небольшие суммы. Также возможно его использование в бизнесе сотовых операторов, страховых компаний и т. д.

На сегодняшний день кредтный скоринг представляет из себя компьютерную программу, в которую вводятся данные потенциального заёмщика и на основе этих данных генерируется ответ - стоит ли выдавать ему кредит.

Целью данной работы является подробный разбор и описание процедуры биннинга, реализация программного продукта, демонстрирующего использование разлчных методов биннинга при построении скоринговых карт на основе логистической регрессии, сравнение результатов построения скоринговых карт и оценки модели логистической регрессии при использовании метода бининга и без него.

Виды алгоритмов оптимального биннинга. Биннинг - это метод дискретизации значений непрерывных переменных в бины (группы). Метод биннинга может решить распространенные проблемы, возникающие при анализе данных, такие как обработка отсутствующих значений, наличие выбросов и статистического шума, а также масштабирование данных.

Методы биннинга используются в приложениях машинного обучения, разведочном анализе и для ускорения обучающих задач; в последнее время биннинг применяется для ускорения обучения в дереве решений с повышением градиента.

В частности, биннинг широко используется в моделировании кредитного риска, являясь важным инструментом при построении скоринговых карт для максимальной дифференциации наблюдений с высоким и низким риском, а также моделирования ожидаемых кредитных убытков.

Существует несколько методов биннинга, основанные на обучении без учителя и с учителем. Распространенными методами обучения без учителя являются интервалы равной ширины и равного размера или равных частот. Хорошо известными методами с учителем, основанными на слиянии, являются монотонный смежный алгоритм объединения (MAPA), также известный как монотонный грубый классификатор максимального правдоподобия (MLMCC) и ChiMerge, в то время как другие методы, основанные на деревьях решений, такие как CART, принцип минимальной длины описания (MDLP), а в последнее время - деревья условного вывода (CTREE).

Существует ряд алгоритмов биннинга, используемых при моделировании кредитного риска.

Равной ширины (Equal-width)

Это очевидный и простой подход к биннингу. Весь диапазон значений предиктора делится на заранее заданное количество интервалов равной ширины. Каждый интервал определяет границы. Предварительно определено количество бинов.

Равного размера (Equal-size)

Диапазон значений независимой переменной разделен на интервалы таким образом, чтобы сформированные в результате бины содержали приблизительно равное количество наблюдений. Ширина интервалов варьируется в зависимости от плотности наблюдений. Целевое количество бинов заранее определено. Алгоритмы биннинга равной ширины и равного размера не используют отношения с целевой переменной.

Оптимальный биннинг является развитием алгоритмов равного размера и равной ширины. Цель алгоритма - определить бины, которые имеют достаточно разные статистические средние оценки значения предиктора. Алгоритм состоит из следующих шагов:

1. Начинаем с достаточно большого количества маленьких бинов.
2. Для каждой соседней пары сегментов вычисляем p-значение.

3. Находим наибольшее значение p из всех пар; если он выше некоторого порога, объедините соответствующую пару сегментов, затем повторите 1, в противном случае выйдите.

Алгоритм ChiMerge состоит из следующих шагов:

1. Сначала инициализируется входной диапазон, который разбивается на подинтервалы, причем каждой выборке присваивается собственный интервал.
2. Для каждой пары смежных подинтервалов вычисляется значение χ^2.
3. Пара с наименьшим χ^2 объединяется в одну ячейку.
4. Шаги 1 и 2 повтояются, пока максимальное значение χ^2 не станет меньше некоторого предварительно определенного порога.

ChiMerge использует статистику χ^2, чтобы определить, являются ли относительные частоты классов соседних интервалов отчетливо разными или они достаточно похожи, чтобы оправдать их объединение в один интервал. Критерий χ^2 используется для проверки гипотезы о том, что два дискретных признака статистически независимы. Применительно к проблеме дискретизации он проверяет гипотезу о том, что атрибут класса не зависит от того, какому из двух соседних интервалов принадлежит пример. Если вывод теста χ^2 заключается в том, что класс не зависит от интервалов, то интервалы следует объединить. С другой стороны, если тест χ^2 заключает, что они не являются независимыми, это указывает на то, что разница в относительных частотах классов является статистически значимой, и поэтому интервалы должны оставаться отдельными.

Критерий χ^2 определяется следующим образом:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

где $m = 2$ (сравниваются 2 интервала), k - количество классов (количество наблюдений в i-м интервале j-го класса). $A_{ij} = \sum_{j=1}^{k} A_{ij}$ - количество примеров в i-м интервале, $R_i = \sum_{i=1}^{m} A_{ij}$ - количество примеров в j-м классе, $N = \sum_{j=1}^{k} C_j$ - общее количество примеров, $E_{ij} = A_{ij} = \frac{R_i C_j}{N}$ - ожидаемая

частота.

Значение порога χ^2 определяется путем выбора желаемого уровня значимости, а затем с помощью таблицы или формулы для получения соответствующего критического значения χ^2 (для получения значения χ^2 также необходимо указать число степеней свободы, которое будет на 1 меньше числа классов). Например, когда имеется 3 класса (2 степени свободы), значение χ^2 на уровне значимости 0,90 равно 4,6. Смысл этого порога в том, что среди случаев, когда класс и атрибут независимы, существует 90% вероятность того, что вычисленное значение χ^2 будет меньше 4,6; таким образом, значения χ^2, превышающие пороговое значение, означают, что атрибут и класс не являются независимыми. В результате, выбор более высоких значений для порога χ^2 заставляет процесс объединения интервалов продолжаться дольше.

Алгоритм деревьев условного вывода (Conditional Inference Trees) основан на исчерпывающем поиске схемы разбиения, которая максимизирует некоторую тестовую статистику, определенную следующим образом, в обобщенной форме:

$$T_j(L_n, \omega) = vec(\sum_{i=1}^{n} \omega_i g_i(X_{ij}) h(Y_i, (Y_1), ..., Y_n))^T).$$

Здесь проверяется независимость ответа Y и независимой переменной X_i, j - индекс переменной в m-мерном X, L_n - длина обучающей выборки, ω - вес, вектор-функция влияния h и векторная функция преобразования g. Полученная векторная статистика отображается в скалярную форму по следующей формуле:

$$c(t, \mu, \sum) = \max_{k=1,...,pq} \left| \frac{(t-\mu)k}{\sqrt{(\sum)_{kk}}} \right|,$$

где t - фактическое значение тестовой статистики T, μ - ожидаемое значение T при независимости, а $\sum_{kk}$ - это k-й диагональный элемент ковариационной матрицы. Значение этого скаляра используется для оценки независимости X и Y. Алгоритм состоит из следующих шагов:

1. Критерий остановки. Проверяеется глобальная нулевая гипотеза H_0 о независимости между Y и всеми ковариатами X_j с $H_0 = \cap_{j=1}^{m} H_0^j$ и

$D(Y|X_j) = D(Y)$ (проверка перестановки для каждой ковариаты X_j). Если H_0 не отклонено (не имеет значения для всех X_j) $\Rightarrow$ остановиться.

2. Выбор переменной. Выбирается ковариата X_{j*} с наиболее сильной ассоциацией (наименьшее значение p).
3. Поиск наилучшей точки разбения. Выполняется поиск наилучшего разбения для X_{j*} (макс. Тестовая статистка).
4. Повтор / повторение. Шаги 1), 2), 3) повторяются для обоих новых разбиений.

Показатели качества биннинга. Хороший алгоритм биннинга с точки зрения моделирования кредитного риска должен соответствовать следующим пунктам:

1. Пропущенные значения группируются отдельно.
2. Каждая группа должна содержать не менее 5% наблюдений.
3. Не должно быть групп с количеством «плохих» или «хороших», равным 0

Первое условие - это очень техническая деталь реализации, и ее довольно легко решить любым способом. Однако два других условия необходимо встроить в алгоритм биннинга. Далее определяется показатель веса категорий предиктора WoE. WoE широко используется в кредитном скоринге для отделения хороших от плохих по конкретной группе:

$$WoE_i = ln\left(\frac{G_i}{\sum\limits_{i=1}^{n} G_i} / \frac{B_i}{\sum\limits_{i=1}^{n} B_i}\right)$$

Здесь G_i - количество хороших в корзине i, B_i - количество плохих в корзине i. Для оценки степени взаимосвязи между независимыми переменными и бинарной зависимой в кредитном скоринге принято использовать расчет показателя информационного значения или IV по формуле:

$$IV = \sum_{i=1}^{n}\left[WoE_i\left(\frac{G_i}{\sum\limits_{i=1}^{n} G_i} - \frac{B_i}{\sum\limits_{i=1}^{n} B_i}\right)\right]$$

На основе практического опыта значения статистики IV можно интерпретировать следующим образом. Если статистика IV:

1. менее 0.02 — независимая переменная не обладает прогностической способностью
2. от 0.02 до 0.1 — низкая прогностическая способность
3. от 0.1 до 0.3 — средняя прогностическая способность
4. от 0.3 до 0.5 — высокая прогностическая способность
5. более 0.5 — превосходная прогностическая способность.

IV достигает максимума, если биннинг почти оптимален. IV используется как мера, чтобы решить, лучше ли одна схема разбиения по сравнению с другой.

Другой важный показатель для оценки качества биннинга - это индекс Херфиндаля-Хиршмана:

$$HHI = n \sum_{j=1}^{n} (\frac{N_i}{N_{total}})^2$$

Здесь Ni - номер выборки в бине i. Чем ниже HHI, тем более равномерно распределен набор данных по результирующим бинам. Большой HHI явный признак недостатка схемы бинирования.

Логистическая регрессия. Для построения скоринговой карты будет использоваться модель логистической регрессии, что является стандартным подходом для данной задачи. Логистическая регрессия представляет собой статистическую модель, используемую для прогнозирования наступления вероятности некоторого события. Логистическая функция способна принимать в качестве входных данных любые значения от отрицательной до положительной бесконечности и выдавать значения от 0 до 1, что можно интерпретировать как вероятность.

Формула логистической регрессии имеет следующий вид:

$$f(y) = \frac{1}{1 + e^{-y}}$$

, где e - экспонента, y - стандартное уравнение регрессии, $f(y)$ - уравнение логистической регрессии.

Математически модель логистической регрессии отражает зависимость логарифма шанса от линейной комбинации предикторов:

$$ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_i + b_2 x_i + ... + b_k x_i + \varepsilon_i$$

где p_i — вероятность наступления просрочки по кредиту для i-го заемщика; x_i - значение j-ой независимой переменной; b_0 — константа, b_j — коэффициенты модели; ε_i — случайная ошибка.

Уравнение отражает линейную зависимость вероятности дефолта по кредиту в зависимости от значений предикторов. Константа отражает уровень риска наступления дефолта при равенстве всех предикторов нулю. Значения коэффициентов при предикторах, отражающих степень их влияния на вероятность дефолта в логарифмической шкале, используются для построения скоринговой карты.

Для интерпретации коэффициентов модели логистической регрессии обычно используют экспоненциальную форму записи модели:

$$p_i = \frac{1}{1+\exp(-(b_0+b_1 x_i+b_2 x_i+...+b_k x_i+\varepsilon_i))}$$

Описание процедуры построения скоринговых карт. Как уже отмечалось ранее, биннинг широко используется в моделировании кредитного риска, являясь важным инструментом при построении скоринговых карт.

Скоринговую карту можно определить как набор характеристик (социальнодемографическое положение, кредитная история, параметры запрашиваемого кредита и пр.) потенциального заемщика и присваивание ему определённого количества баллов на основе этой информации.

Для построения скоринговых карт могут использоваться методы линейной регрессии, логистической регрессии, дискриминантного анализа, деревьев решений, нейронных сетей и др. На практике чаще всего используется логистическая регрессия.

Для построения модели логистической регрессии используются только категориальные переменные. Если имеются количественные переменные их предварительно категоризуют с помощью процедуры биннинга. Затем для всех без исключения градаций категориальных переменных рассчитывают показатели WOE и заменяют ими фактические значения независимых переменных. По значениям WOE строят модель логистической регрессии.

Заключительным этапом разработки скоринговой модели является перевод коэффициентов логистической регрессии в скоринговые баллы. Если взять оценки коэффициентов логистической регрессии и умножить их на значения независимых переменных, то получиться итоговый скоринговый балл в шкале натуральных логарифмов:

$$finalscore = \hat{b_1}x_1 + \hat{b_2}x_2 + ... + \hat{b_k}x_k,$$

где x_j — значение предикторов для оцениваемого заемщика,$\hat{b_j}$ — оценки коэффициентов логистической регрессии.

Для приведения скоринговых баллов в линейную шкалу пользуются масштабированием. Масштабирование не изменяет прогностическую способность скоринговой карты, а лишь переводит скоринговые баллы в новую шкалу, удобную для использования. Скоринговый балл в линейной шкале представляет собой отношение шансов «хороших» заемщиков к «плохим». Для масштабирования необходимо прежде всего задать диапазон числовой шкалы с минимум и максимум (например, от 0 до 1000). На результат масштабирования также влияют два показателя: количество баллов, которое удваивает шансы стать «хорошим» заемщиком и значение шкалы, в котором достигается заданное отношение шансов «хороших» к «плохим». Наиболее часто используют скоринговые карты, в которых каждые 20 баллов удваивают шансы стать «хорошим». Другой стандарт — каждые 40 баллов удваивают шансы стать «хорошим» заемщиком. Для приведения коэффициента логистической регрессии в скоринговый балл в линейной шкале применяют следующее преобразование:

$$score = A + R \cdot b_j,$$

где, R — множитель; A — смещение.

Множитель определяют по формуле:

$$R = \frac{D}{ln(2)},$$

где D — количество баллов, удваивающее шансы.

Смещение определяют по формуле:

$$A = B - R \cdot ln(C),$$

где B — значение на шкале баллов, в которой соотношение шансов составляет C:1.

Если уравнение логистической регрессии строится по значениям WOE, то формула расчета скорингового балла в линейном масштабе будет следующая:

$$score = -(WOE_j \cdot b_i + \frac{(b_0)}{n}) \cdot R + \frac{A}{n},$$

где WOE_j — значение WOE для каждой j-ой категории сгруппированной переменной, n — количество независимых переменных в уравнении регрессии, $b0$ — константа, bi — коэффициент регрессии для i-ой переменной.

Предобработка данных и построение модели логистической регрессии. Написание программного кода будет осуществляться на языке Python. Для анализа был выбран набор данных, который содержит информацию о дефолтах, демографических факторах, кредитных данных, историях платежей и выписках по счетам клиентов кредитных карт в Тайване с апреля 2005 года по сентябрь 2005 года. Данные взяты с сайта Kaggle: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.

Обычно перед построением модели необходимо провести анализ и предобработку данных. В данном случае была произведена проверка пропущенных значений в данных, замена значений переменных, которые не были описаны. В ходе алгоритма биннинга производится замена значений переменных на значения WoE. Для биннинга, основанного на деревьях решений используются вероятности возвращаемые деревом. Набор данных разделяется на тестовый и обучающий наборы, отбираются наиболее информативные признаки по критерию IV и по тепловой карте.

Далее строится модель логистической регрессии для дальнейшего построения скоринговой карты и оценки вероятности дефолта.

Сравнение результатов применения алгоритмов биннинга для моделрования кредитного риска. Для оценки предсказательной способности логистической регрессии используется ROC-кривая — график, харак-

теризующий качество бинарной классификации, которая показывает зависимость количества верно классифицированных положительных примеров (TPR) от количества неверно классифицированных отрицательных примеров (FPR) при варьировании порога решающего правила.

Достоверность модели логистической регрессии характеризуется ее способностью отличать «хороших» заемщиков от «плохих». Для оценки качества модели используется характеристика площади под кривой ROC AUC (Area under ROC). Обычно считают, что значение площади от 0.9 до 1 соответствует отличному качеству модели, от 0.8–0.9 — очень хорошему, 0.7–0.8 — хорошему, 0.6–0.7 — среднему, 0.5–0.6 — неудовлетворительному.

При использовании различных алгоритмов биннинга получились следующие результаты:

1. Модель с использованием оптимального биннинга - 0.764
2. Биннинг с использованием алгоритма равной ширины - 0.756
3. Биннинг с использованием алгоритма равного размера - 0.756
4. Биннинг с использованем алгоритма Chi-Merge - 0.754
5. Биннинг с использованем алгоритма деревьев решений - 0.763

Модель без биннинга с предобработкой данных показала результат - 0.756.

Модель без какой-либо предобработки данных показала результат - 0.627.

Заключение. В результате выполнения бакалаврской работы был подробно описан и реализован подход построения скоринговых карт на основе логистической регрессии при помощи процедуры биннинга. Был разработан программный продукт, посредством которого проведена оценка качества и достоверности модели логистической регрессии при использовании подхода биннинга и без него. В результате наилучшее качество показала модель, при построении которой использовалась процедура биннинга, что может говорить о большей предпочтительности такого подхода. Был проведен сравнительный анализ эффективности различных алгоритмов биннинга, наиболее оптимальными из них оказались оптимальный биннинг и биннинг, основанный на деревьях решений.

Таким образом реализованный программный продукт способен опре-

делять платежеспособных и неплатежеспособных заёмщиков, что позволяет оптимизировать работу банка и минимизировать возможные финансовые риски.