Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

Факультет математики и информатики
Кафедра математического анализа и моделирования
Направление подготовки 01.03.02 – Прикладная математика и информатика
Профиль: Математическое и информационное обеспечение экономической деятельности

ДОПУСТИТЬ К ЗАЩИТЕ
Зав. кафедрой
_____ Н.Н. Максимова
«_____»_____2017г.

БАКАЛАВРСКАЯ РАБОТА

на тему: Построение скоринговых карт с использованием логистической регрессии

Исполнитель
студент группы 352об _____ Г.М. Шостокас
(подпись, дата)

Руководитель
канд. физ.-мат. наук _____ И.В. Красников
(подпись, дата)

Нормоконтроль
доцент, канд. техн. наук _____ А.В. Рыженко
(подпись, дата)

Благовещенск 2017

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

Факультет математики и информатики
Кафедра математического анализа и моделирования

УТВЕРЖДАЮ
Зав. кафедрой
_____ Н.Н. Максимова
«_____»_____2017г.

З А Д А Н И Е

К бакалаврской работе студента Шостокас Германа Михайловича.

1. Тема бакалаврской работы: Применение логистической регрессии с построением скоринговой карты для оценки заемщиков банка(утверждена приказом от 10.04.2017 № 770-уч).

2. Срок сдачи студента законченной работы:13 июня 2017 г.

3. Исходные данные к выпускной квалификационной работе: сведения из литературных источников, диссертаций, монографий, справочные данные, определяющую предметную область.

4. Содержание выпускной квалификационной работы: анализ и формализация сведений о предметной области – описание среды и концептуальной постановки задачи моделирования; адаптация данных и метода анализа к рабочей среде программного приложения; описание этапов реализации алгоритма в программной среде DeductorAcademicStudio 5.3;анализ полученных результатов.

5. Перечень материалов приложения: отсутствуют.

6. Консультанты по бакалаврской работе – нормоконтроль: Рыженко А.В., канд. техн. наук, доцент.

7. Дата выдачи задания 24.04.2017 г.

Руководитель бакалаврской работы:КрасниковИ.В.,канд. физ.-мат. наук.

Задание принял к исполнению:_____Г.М.Шостокас

РЕФЕРАТ

Бакалаврская работа содержит 36 с., 6 рисунков, 5 таблиц, 18 источников.

ЦЕЛИ СКОРИНГА, РАБОТА СКОРИНГА, ВИДЫ СКОРИНГА, НЕДОСТАТКИ, ДАННЫЕ ДЛЯ ПОСТРОЕНИЯ, ВЫХОДНАЯ ФУНКЦИЯ, ОБУЧАЮЩАЯ И ТЕСТОВАЯ ВЫБОРКИ, ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, ROC-КРИВАЯ, WOE

Приведены сведения об аспектах применения скоринговых карт. Изучены теоретические методы построения логистической регрессии, разбиение на классы WOE, ошибки первого и второго рода, построение ROC- кривой.

Приведена практическая реализация построения модели логистической регрессии и скоринговой карты. Проведена оценка карты.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В банковской системе очень сильно распространен кредитный скоринг. Кредитный скоринг можно определить, как метод начисления потенциальным заемщикам соответствующего количества баллов, опираясь на информацию о его социально-демографическом положении, кредитной истории, параметрах кредита, и принятии решения о выдаче или об отказе в кредите в зависимости от набранного суммарного количества баллов. В современное время банки предъявляют высокие требования к риск-аналитике из-за мошенничества и роста невозвратных кредитов. В жизни возникает задача не только принятия решения об отказе или выдаче кредита заемщику на основе набранных баллов, но и задача определения оптимального количества баллов для выдачи кредита заемщику. Вторая задача решается путем распределения баллов «надежных» и «ненадежных» заемщиков, делая вывод об этом из скоринговой карты и анализа соотношения риска и доходности в кредитном портфеле банка. Следовательно, кредитный скоринг – это инструмент снижения рисков невозвращаемых кредитов, а также хороший помощник в определении оптимальной структуры кредитного портфеля.

Во многих коммерческих банках скоринговые модели являются собственными разработками на основе данных о заемщиках банка прошлых лет или на основе заемщиков нескольких банков специализированные фирмы составляют индивидуальные скоринговые модели.

1 ПРЕДМЕТНАЯ ОБЛАСТЬ, ЦЕЛИ И ЗАДАЧИ РАБОТЫ

За основу скоринговой модели могут браться различные статистические модели. Каждая из моделей может быть получена любым доступным методом: линейной регрессией, логистической регрессией, дискриминантным анализом, деревом решений, нейронной сетью. Мы будем использовать логистическую регрессию, так как данная модель наиболее часто используется на практике для построения скоринговых карт.

В данной выпускной квалификационной работе будет представлен метод построения скоринговой карты, благодаря которому можно снизить индивидуальные кредитные риски банков. Кроме того, рассматриваются такие вопросы как анализ ROC– кривых, анализ распределения скоринговых баллов, коэффициент Джинни, расчет статистики Колмогорова-Смирнова, ошибки первого и второго рода. Рассматривается вопрос применения логистической регрессии для верной классификации заемщиков.

Практическая часть работы выполнена на аналитической платформе DeductorAcademicStudio 5.3, предоставляемая компанией BaseGroupLabs. Данная платформа предназначена для создания законченных аналитических решений. В платформу встроены современные методы извлечения, визуализации данных и анализа данных.

Целью и задачей данной работы является построение наиболее оптимальной скоринговой карты, которая бы верно классифицировала «хороших» и «плохих» заемщиков банка.

2 ОПИСАНИЕ И ИСПОЛЬЗОВАНИЕ СКОРИНГОВОЙ МОДЕЛИ

Тем банкам, которые выдают кредиты нужно оценить нового клиента и принять какое-либо решение о выдаче или невыдаче ему кредита.

В мировой практике существует два основных метода осуществления этой процедуры, которые могут применяться как отдельно, так и в сочетании друг с другом:

1)субъективное заключение экспертов или кредитных инспекторов;

2)автоматизированные системы скоринга.

Оценка кредитного риска в скоринговых системах дает возможность, оценив набор определенных признаков, которые характеризуют заемщика, принять решение выдавать кредит или нет. Данная методика имеет довольно долгий срок использования в оценке насколько способно предприятие или любое другое физическое лицо. Серьезным свойствомскоринговых систем является, то что решение принимается автоматически и специалист в этом не участвует.

Исходя из всей философии скоринга, не нужно думать, почему определенный клиент не вернул выданные ему средства. Скоринг определяет характеристики, которые присущи надежным и ненадежным лицам.

Иными словами скоринг – это такая система оценки надежности заемщика, исходящая из целого ряда параметров. Клиенту поступает предложение заполнить анкету в то время. Как подается заявка на получение кредита. Вопросы, содержащиеся в анкете имеют определенный смысл. В этом и состоит неотъемлемая часть оценки заемщика скоринговой моделью. От того, как ответит клиент по каждому пункту анкеты, ему присваивается определенное количество баллов. Чем это количество больше, темы выше вероятность того, что решение о выдаче кредита будет склоняться в положительную сторону. Если у вас негативная кредитная история, то последующие ответы и число набранных баллов не играют особой роли. Одного такого случая достаточно, чтобы вам отказали.

Суть скоринговой модели состоит в том, чтобы всю информацию о клиенте перевести в баллы, а затем определить комбинацию факторов, которые позволят показать причину невозврата средств заемщиком. Скоринговая модель должна предсказывать высокую вероятность дефолта для неплатежеспособных заемщиков и низкую для тех, кто своевременно погасил кредит.

2.1Цели скоринга в банковской системе

Различная скоринговая модель, которая применяется в кредитовании нужна для получения следующихрезультатов:

1) увеличение кредитного портфеля из-за снижения доли необоснованных отказов по кредитам;

2)ускорение процедуры оценки потенциального заемщика;

3)снижение уровня невозврата кредитных средств;

4) повышение качества и точности оценки заемщика;

5) централизованное накопление данных о клиенте;

6) снижение резерва на сумму вероятных потерь по кредитам;

7) оценка динамики изменений индивидуального кредитного счета и всего портфеля кредитов в целом.

2.2 Работа кредитногоскоринга

Многие банки имеют на вооружении скоринговые модели оценки кредитоспособности. Эти модели помогают достичь определенных целей. Дабы минимизировать предвзятое отношение менеджера или сговор сотрудников используют скоринговые модели. Вся информация, котрая вносится в анкету, должна подтверждаться наличием документов. В банке менеджер исполняет роль того, кто вносит данные в программу. По мере полной загрузки данных компьютерная программа выдает результат, то есть определенное количество набранных баллов.

Если вы набрали минимальное количество баллов, то в кредите вам могут отказать. Если же вы набрали количество баллов намного выше среднего, то при средней кредитной сумме вам могут одобрить выдачу кредита сразу. Если же вы хотите получить высокую сумму кредита, то вас оповестят о том, что

первоначальный проверочный этап вы прошли и заявка будет передана в службу безопасности банка.

2.3Виды скоринга

В общем случае скоринговая модель состоит из семи видов оценки, четыре из которых имеют отношение к кредитованию, а три – к маркетингу. Для кредитной практики характерны такие виды скоринга[12]:

1) по заявкам (Application-скоринг). В этой модели проверяют насколько надежен клиент. Такая модель строится на проверке анкеты и начислении балла за каждый данный вами ответ;

2) от мошенничества (Fraud-скоринг). Такая модель хороша для отбора мошенников, прошедших первоначальную проверку.То, какими способами банк проводит тестирование мошенничества является тайной у каждого банка;

3) прогноз поведения (Behavioral-скоринг). В данном виде анализируется отношение между заемщиком и кредитом, оценивается то на сколько заемщик может просрочить выплату по кредиту. По этим результатам корректируют сумму кредита, которую возможно выдать данному клиенту;

4) работа по возвратам (Collection-скоринг). Такая модель больше применима к неудачным кредитам, когда клиенту необходимо вернуть неоплаченную им задолженность. Происходит разработка плана по возврату кредита: от предупреждения до судебного делопроизводства или передачи дела в коллекторскую фирму;

5) предпродажная оценка (Pre-Sale) –программа ищет какие потребности мог бы испытывать заемщик, пытаясь предложить ему определенный продукт;

6) отклик (Response) –происходит оценка мнения клиента,насколько близко программа выбрала подходящий для него кредитный продукт;

7) оценка истощения (Attrition) –программа оценивает возможность прекращения сотрудничества клиента с данным банком в скором времени.

2.4Недостатки скоринговой системы

Каждая система оценки кредитоспособности имеет свои недостатки. Большим недостатком является плохая адаптация под реальные параметры и

недостаточная гибкость. Например, скринговая модель, принятая в США, даст высокий балл клиенту, который сменил не малое количество мест работы. Этот человек будет считаться очень хорошим специалистом, и достаточно востребованным на рынке труда. У нас же такой факт скажется довольно негативно на определении баллов заемщику. Большое количество баллов получит тот, кто имеет одну единственную запись в трудовой книжке. Заемщик часто меняющий место работы, считается неблагонадежным и плохим специалистом. Рейтинг такого заемщика резко падает. Потому что новой работы может и не быть. Следовательно, просрочки в платежах неизбежны. Для адаптации под наши условия жизни, анкеты для оценки кредитоспособности клиента должны разрабатываться специалистами высокой категории и квалификации.Так что любая система скоринга имеет, по крайней мере, два недостатка:

1) дороговизна адаптации под современные реалии;

2) влияние субъективного мнения специалиста на выбор модели оценки клиента.

2.5Данные участвующие в построении модели

В тех случаях, когда проводится оценка кредитоспособности физических лиц, сотрудник банка должен опираться на целый ряд критериев. Все их можно разделить на три большие группы, в каждую из которых входит множество показателей[2].

Личные:

1) паспортные данные;

2) семейное положение;

3) возраст;

4) наличие детей, их возраст и количество.

Финансовые:

1) сумма вашего ежемесячного дохода;

2) место работы и занимаемая должность;

3) количество записей в трудовой книжке;

4) период трудоустройства на последнем предприятии;

5) наличие долгов, непогашенных кредитов;

6) наличие собственного жилья, автомобиля, банковских счетов.

Дополнительные:

1) имеется ли дополнительный источник дохода, не подтвержденный документально;

2) возможность предоставления поручителя.

Если говорить о физических лицах, то тут оценка заемщика также проводится по многим показателям. Существует множество факторов, способных положительно повлиять на рейтинг:

1) высокая зарплата;

2) наличие недвижимости;

3) долгий срок проживания в определенном регионе;

4) наличие вкладов;

5) документальное подтверждение доходов;

6) наличие дорогого сотового телефона;

7) подтверждение официального трудоустройства, особенно на государственных предприятиях и в бюджетной сфере;

8) наличие открытых счетов в банке;

9) наличие высокой суммы аванса при получении ипотеки или автокредита;

10)возможность предоставления рекомендаций от поручителя,отличная кредитная история.

3 ПОДГОТОВКА ДАННЫХ ДЛЯ ПОСТРОЕНИЯ СКОРИНГОВОЙ МОДЕЛИ

Скоринговые карты строятся в основном на статистических моделях. Для того, чтобы ее построить нужна качественная информация о заемщике банка. Качество исходных статистических данных для построения статистической модели определяет ее точность прогнозирования и успех разработки скоринговой системы в целом. Предыдущий кредитный опыт является основной составляющей для разработки скоринговой модели. Очень важен достаточный объем информации. Данные должны удовлетворять требованиям значимости и случайности, варьирование количества данных тоже возможно. Для построения скоринговой модели исходные данные могут содержать внутренние данные анкет заемщиков, а так же данные кредитных историй, записи которых насчитывают сотни тысяч.

В лучшем случае скоринговые модели должны применяться в отношении тех кредитных продуктов, сектора рынка, и экономической ситуации, которые положили основу данных о прошлом кредитном опыте. Например, для разработки скоринговой карты по автокредитованию не могут быть использованы сведения по потребительским кредитам. Для построения достоверной скоринговой модели исходные данные должны иметь определенную историческую давность[11]. То есть определяется период за который собирались данные. Например, данные по потребительским кредитам трехмесячной давности не будут подходящими для построения модели, нежели данные трехгодичной давности, а десятигодичной давности данные будут считаться устаревшими. Исторический период данных, который будет наиболее подходящим для построения модели, чаще всего исходит из вида скоринга и вида кредитования и требований надзорных органов. Данные определенных типов клиентов нужно исключить из информационной базы. Это могут быть сотрудники банка, умершие клиенты, несовершеннолетние, мошенники, двойные заявки, кредиты по украденным картам и др. Также следует исключить кредиты с очень большими суммами

кредитов, необычными условиями погашения и целями займа. Дополнительным звеном критерия отбора может послужить рынок или вид кредитования, для которого нужно построить скоринговую модель. В некоторых случаях использование не одной единицы скоринговых карт для одного портфеля по виду кредитования обеспечивает лучшее распознание риска, чем использование одной единственной скоринговой карты. Для осуществления данного подхода обычно перед построением скоринговой модели исходную базу клиентов делят по группам с помощью многомерных статистических методов, например, кластерного анализа, деревьев решений или эвристическими методами.

3.1 Определение выходной функции

Целью скоринговой модели является определение зависимой переменной. Цели могут быть общими, например, сократить потери по кредитным счетам или сокращение числа дефолтов по одобренным заявкам в течении трех месяцев после принятия положительного решения. Зависимая переменная может определять количественные и качественные значения. Например, та сумма которую погасит заемщик по кредиту который он прострочил можно взять за целевую переменную. В скоринге зависимая переменная может принимать категориальную шкалу измерения. Когда происходит определение зависимойперемнной, заемщиков непосредственно делят на «плохих», «хороших» и «неопределенных». Мошенники, банкроты определяются как «плохие» заемщики. В отношении остальных заемщиков критерий « плохого клиента» определяется количеством дней просрочки платежа. К «неопределенным» клиентам можно отнести клиентов у которых малая кредитная история, те кто имеет небольшую просрочку платежа. При построении скоринговой карты используются клиетны, которые определены, как «плохие» и «хорошие». «Неопределенные» клиенты учитываются при построении прогноза зависисмой переменной по модели логистической регрессии. Иногда, в отдельную категорию вносят заемщиков. Которым отказали в выдаче кредита, определяя их как « отклоненные» клиенты. Для построения генеральной совокупности заемщиков в обучающей выборке учитываются «неопределенные» и «отклоненные» заемщики[11].

Процесс построения скоринговой модели часто разбивают на два этапа:

1)построение первичной модели по данным «плохих» и «хороших» клиентов без учета «отклоненных» клиентов;

2)построение конечной модели с учетом анализа отклоненных заявок. Многие эксперты в области кредитного скоринга отмечают, что анализ отклоненных заявок клиентов требует больших ресурсов и не всегда приводит к улучшению качества конечной скоринговой модели.

Категориальный с двумя категориями – часто используемый вид измерения зависимой переменной. К категории «плохих» клиентов чаще всего относят клиентов, которые имеют просроченную задолженность в девяносто и более дней. Лучше всего подходит логистическая регрессия для моделирования значений такой переменной. Банки имеют право строить различные виды скоринговых карт с различными значениями зависимой переменной, ставя свои критерии для определения «плохих» и «хороших» заемщиков, так же имеют возможность изменить срок просрочки платежей. Например, зависимой переменной могут быть просрочки более 30 дней, 60 дней, 90 дней и более по определенному кредиту на данный момент или низкий статус за всю кредитную историю, количество просрочек выше заданного числа, и размер задолженности играют немаловажную роль.

3.2 Формирование обучающей и тестовой выборки

Доступные для построения скоринговой модели информационные данные называются исторической выборкой. Историческая выборка должна быть репрезентативной, то есть она должна правильно и наиболее точно отображать исследуемую генеральную совокупность. Чтобы проверить скоринговую модель на адекватность и то, насколько точно она делает предсказания, то но этапе разработки историческую выборку делят на обучающую выборку – это те наблюдения по которым будет построена модель; тестовую выборку – наблюдения которые не войдут в построение модели, а будут использованы для проверки точности предсказания, благодаря им будет известно значение зависимой

переменной. Обучающая и тестовая выборки обычно формируются соотношением(70÷80) % и (30÷20) % из объема исторической выборки.

Чтобы проверить достоверность логистической регрессии используют тестовую выборку непосредственно после построения самой регрессии. В кредитном скоринге это называется способностью отличать «хороших» заемщиков от «плохих». На контрольной и тестовой выборке основывается проверка достоверности модели путем применения и сравнения результатов. Обычно используют стратегию генерализации модели на основе двух выборок. Небольшая погрешность в точности, которая получена в обучающей и тестовой выборке – признак того, что скринговая модель на практике покажет себя идентичным способом. Создание трех или более выборок относится к сложной генерализации модели. Одну выборку берут для оценки параметров модели, следующую используют как проверочную выборку, если же получены отклонения от результатов по обучающей и тестовой выборке. Третья выборка используется для тестирования двух предыдущих выборок из которой убираются выбросы и переменные, имеющие влияние на отклонения[5].

3.3Модель логистической регрессии

Логистическая регрессия – самая распространенная статистическая модель для построения скоринговых карт при бинарной зависимой переменной. Математически модель логистической регрессии выражает зависимость логарифма, шанса (логита) от линейной комбинации независимых переменных:

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_i^{(1)} + b_2 x_i^{(2)} + \ldots + b_k x_i^{(k)} + \varepsilon_i, \qquad (1)$$

где p_i – вероятность наступления дефолта по кредиту для *i*-го заемщика;

b_0 – независимая константа модели;

b_j – параметры модели;

$x_i^{(j)}$ – значение*j*-ой независимой переменной;

ε_i – компонент случайной ошибки.

Уравнение (1) отражает линейную зависимость вероятности наступления просрочки по кредиту в зависимости от значений независимых переменных. Константа в модели отражает естественный уровень риска наступления моделируемого события при равенстве всех независимых переменных нулю. Значения коэффициентов при независимых переменных, отражающих степень их влияния на шанс дефолта в логарифмической шкале, используются для построения скоринговой карты. Значение константы в модели логистической регрессии зависит от распределения в данных по категориям зависимой переменной. В случае перевзвешивания выборки для изменения этого распределения, для более адекватной последующей оценки качествополученной модели константу корректируют и получают следующую модель логистической регрессии:

$$\ln\left(\frac{p_i^*}{1-p_i^*}\right)=\ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right)+b_0+b_1x_i^{(1)}+b_2x_i^{(2)}+...+b_kx_i^{(k)}+\varepsilon_i, \qquad (2)$$

где p_i^* – откорректированная априорная вероятность;

ρ_0 и ρ_1 – доли «хороших» и «плохих» заемщиков в выборке;

π_0 и π_1 – доли «хороших» и «плохих» заемщиков в генеральной совокупности.

Исходя из модели (1) могут быть откорректированы значения прогнозируемых апостериорных вероятностей дефолта и получены, таким образом априорные вероятности для генеральной совокупности[1]. Однако для построения скоринговой карты достаточно значений коэффициентов при независимых переменных, которые при преобразовании (2) остаются неизменными. Для интерпретации коэффициентов модели логистической регрессии обычно используют экспоненциальную форму записи модели:

$$p_i=\frac{1}{1+\exp(-(b_0+b_1x_i^{(1)}+b_2x_i^{(2)}+...+b_kx_i^{(k)}+\varepsilon_i))} \qquad (3)$$

При включении в модель логистической регрессии непрерывных количественных переменных коэффициенты при них будут показывать, на сколько в среднем измениться логарифм шанса наступления просрочки по кредиту при

изменении независимой переменной на единицу своего измерения при неизменности остальных переменных. В экспоненциальной форме коэффициенты будут показывать насколько в среднем изменятся шансы наступления дефолта при изменении независимой переменной на единицу своего измерения при неизменности остальных переменных. Если коэффициент регрессии будет положительный, то его экспонента будет больше единицы и шансы будут возрастать, если коэффициент окажется отрицательным – меньше, шансы будут убывать. При включении в модель бинарной независимой переменной, коэффициент регрессии в экспоненциальной форме при фиктивной переменной будет показывать соотношение шансов проявления дефолтов при наличии фактора, отражаемого бинарной независимой переменной, по сравнению с его отсутствием.

3.4 ROC-кривая(Receiver Operator Characteristic)

ROC-кривая может показать нам классификационную зависимость положительных и отрицательных исходов. Такая кривая, как правило, строится по значениям вероятностей дефолта, по модели логистической регрессии или ссылаясь на результаты скоринговой карты. В первом случае ROC-анализ дает возможность установить порог для разделения «плохих» и «хороших» клиентов для того, чтобы достигнуть нормальный уровень чувствительности и специфичности модели. А во втором найти такой балл, который был бы оптимален для избавления от «плохих» заемщиков и оставлял «хороших».

Высокая чувствительность может позволить максимально предотвратить пропуск «плохих» заемщиков. Перед анализом чувствительности стоит задача минимизации кредитного риска, связанного с выдачей кредита. Модель, обладающая высокой специфичностью, будет являться рискованной, так как не достаточно тщательно будет выбирать «плохих» заемщиков.Для определения оптимального порога классификации существуют множество критериев, например:

1) уровень плохих кредитов в портфеле одобренных заявок;

2) минимизация ошибок классификации и др.

Наилучший выбор порога неотъемлемо зависит от ошибок 1-го и 2-го рода при классификации. Составляемая модель должна как можно лучше определять «плохих» заемщиков, так как ошибки первого рода гораздо выше. Если же произойдет снижение порога, то чувствительность данной модели увеличится. Из этого следует способность модели хорошо определять заемщиков с просрочкой платежа. За более лучший порог отсечения следует взять балансовую точку, которая расположена между чувствительностью и специфичностью.

Для анализа классификационной способности скоринговой модели используют показатель площади под ROC-кривой – AUC (от англ. AreaUnder-Curve). Площадь под кривой AUC (см. таблица 1) изменяется от 0.5 (нет разделения) до 1 (идеальное разделение). 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.

Заметим, что коэффициент площади под ROC-кривой нужен только для того, чтобы провести сравнительную характеристику между моделями. Значение, полученное от площади под кривой, не влияет на чувствительность и специфичность модели. Во время анализа на качество модели по полученному значению площади под ROC-кривой чаще всего вычисляют индекс Джини (таблица 1). Данный индекс формирует значение площади под кривой в интервал от нуля до единицы. И чем выше его показатель, тем выше ограничивающая способность модели. Рассчитываетсяиндекс Gini[14] по формуле:

$$Gini = 2 \cdot (AUC - 0.5). \quad (4)$$

Для оценки качества прогнозирования модели логистической регрессии на основе этого распределения рассчитывают тест Колмогорова-Смирнова. В тесте Колмогорова-Смирнова[14] сравниваются два кумулятивных распределения скоринговых баллов «хороших» и «плохих» заемщиков. Статистика Колмогорова-Смирнова (см. таблица 1) вычисляется как максимальная разница между кумулятивными функциями этих распределений.

$$KS = \max\left|F_m(x) - G_n(x)\right| \cdot 100, \quad (5)$$

где F_m и G_n – эмпирические кумулятивные распределения скорингового балла для «плохих» и «хороших» заемщиков;

n, m – количество «плохих» и «хороших» заемщиков.

Таблица 1 –Показатели индексов AUC, Gini, KS

Метка столбца	Минимум	Максимум	Среднее	Стандартное отклонение	Сумма	Сумма кв-ов
Индекс AUC	0,764	0,767	0,766	0,002	1,532	1,17
Индекс Gini, %	52,922	53,552	53,237	0,445	106,474	5668,5
Индекс KS, %	45,482	48,618	47,05	2,217	94,101	4432,4

Диапазон значений статистики Колмогорова-Смирнова измеряется от 0 до 100. Чем выше значение статистики Колмогорова-Смирнова, тем лучше работает модель.

Для вычисления статистики Колмогорова-Смирнова и проверки гипотезы о равенстве двух функций распределений необходимо использовать следующий алгоритм. Заемщиков проранжировать в порядке увеличения скорингового балла и провести группировку. Группировочным признаком выступает набранный скоринговый балл. Затем в каждой полученной группе заемщиков необходимо рассчитать следующие показатели:

1) количество «плохих» заемщиков;

2) отношение шансов «плохих» к «хорошим» заемщикам;

3) процент «плохих» и «хороших» кредитов;

4) кумулятивную сумму «плохих» и «хороших» кредитов;

5) кумулятивный процент «плохих» и «хороших» кредитов;

6) общий кумулятивный процент плохих кредитов от их общего числа;

7) разницу между кумулятивными процентам плохих и хороших кредитов.

После чего нужно найти максимальную разность между кумулятивным процентом«хороших» и «плохих» кредитов и вычислить по формуле (4) стати-

стику Колмогорова-Смирнова.Полученное значение статистики необходимо сравнить стабличным значением по таблице распределения Колмогорова-Смирнова с выбраннымуровнем значимости или при числе «плохих» и «хороших» заемщиков соответственно больше80 можно взять приближенное пороговое значение, вычисляемое по формуле:

$$z(\alpha) = \sqrt{\frac{m+n}{mn}}, \tag{5}$$

где $z(\alpha)$ – значение соответствующее выбранному уровню значимости.

Если расчетное значение статистики по формуле (4) меньше порогового по таблице или по формуле (5), то гипотезу о равенстве двух функций распределений отвергают.

Альтернативная мера оценки валидации модели является коэффициент дивергенции. Коэффициент дивергенции представляет собой оценку разницы-математических ожиданий распределений скоринговых баллов для «плохих» и «хороших» заемщиков, нормализованную дисперсиями этих распределений, и рассчитывается по формуле:

$$D = \frac{2 \cdot (\overline{x_1} - \overline{x_2})^2}{(s_1^2 - s_2^2)^2}, \tag{6}$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения скорингового балла для «плохих» и «хороших»заемщиков;

s_1^2 и s_2^2 – дисперсии скорингового балла для «плохих» и «хороших» заемщиков.

Чем выше коэффициент дивергенции, тем лучше классификационная способность качества модели. В большинстве случаев на практике используют все же статистику Колмогорова-Смирнова. Все же этот показатель не всегда может давать адекватную оценку обоснованности и пригодности модели. Возможность применения статистики Колмогорова-Смирнова, а также коэффициента дивергенции, как правило, можно увидеть в таблице распределения скоринговых баллов, а именно их симметричность и наложение друг на дру-

га.Значение коэффициента дивергенции чувствительно к асимметрии распределенийскоринговых баллов и может давать слишком низкую или слишком высокую оценку в зависимости отнаправления асимметрии. Статистика Колмогорова-Смирнова, наоборот, устойчива касимметрии распределений скоринговых баллов. Однако, статистика Колмогорова-Смирноваможет ошибочно определить оптимальную оценку при наложении кривых распределенийскоринговых баллов друг на друга. Когда обе кривые распределения скоринговых баллов для«плохих» и «хороших» заемщиков нормально или приблизительно нормально распределены– можно использовать и статистику Колмогорова-Смирнова, и коэффициент дивергенции[16].

3.5 Ошибка первого и второго рода

Вероятности ошибки первого рода, т.е. вероятности ошибки, возникающей при отнесении надежного заемщика к классу ненадежных заемщиков, данные ошибки ведут к потерям типа упущенной выгоды.

Вероятности ошибки второго рода, т.е. вероятности ошибки, возникающей при отнесении ненадежного заемщика из класса к классу надежных заемщиков, данные ошибки ведут к прямым потерям банка.

Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности (confusionmatrix, см. таблица 2), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам[4].

Таблица 2 – Таблица сопряженности

	Фактически	
Модель	положительно	отрицательно
положительно	TP	FP
отрицательно	FN	TN

В таблицы 2 введены обозначения:

TP (TruePositives) – верно классифицированные «хорошие» клиенты;

TN (TrueNegatives) – верно классифицированные «плохие» клиенты;

FN (FalseNegatives) – «хорошие» клиенты, классифицированные как «плохие» (ошибка первого рода). Это так называемый «*ложный пропуск*» – когда интересующее нас событие ошибочно не обнаруживается;

FP (FalsePositives) – «плохие» клиенты, классифицированные как «хорошие» (ошибка второго рода);

Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии.

Доля истинно положительных примеров (TruePositivesRate):

$$TPR = \frac{"хорошие" клиенты \cdot 100\%}{"хорошие" клиенты" + ложноотрицательные" результаты}. \quad (7)$$

Доля ложно положительных примеров (FalsePositivesRate):

$$FPR = \frac{"ошибочный" результат}{"плохие" клменты" + ошибочный" результат} \cdot 100\%. \quad (8)$$

Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (Sensitivity) – это и есть доля истинно положительных случаев:

$$S_e = TPR = \frac{"хорошие" клиенты \cdot 100\%}{"хорошие" клиенты" + ложноотрицательные" результаты}. \quad (9)$$

Специфичность (Specificity) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$S_p = \frac{"плохие" клиенты}{"плохие" клиенты" + ошибочный" результат} \cdot 100\%. \quad (10)$$

3.6 Разбиение на классы WOE

Обработчик Конечные классы позволяет уменьшить число значений исходного набора данных за счет их объединения в пределах некоторого интервала с использованием информации о бинарной выходной переменной.

Назначение:

1) понижение большого значения признаков без ущерба информационной базе;

2) исключение признаков с низкой значимостью для понижения размерности данных;

3) восстановление пропусков;

4) борьба с выбросами и экстремальными значениями;

5) более упрощенное описания исследуемых объектов.

Описание алгоритма. Процедура сокращения уникальных значений состоит из двух шагов:

1) формирование исходного множества уникальных значений поля до обработки, или начальных классов;

2) «сжатие» начальных классов в меньшее количество интервалов, называемыхконечными классами;

Предварительный анализ взаимосвязи скоринговых переменных на вероятность дефолта по кредиту помогает ограничить количество рассматриваемых для построения модели логистической регрессии переменных. Основными методами оценки наличия связи между зависимой бинарной переменой и независимыми категориальными переменными является расчет критерия хи-квадрат и показателя информационного значения.

При использовании критерия хи-квадрат выдвигают гипотезу H_0 об одинаковом распределении «плохих» и «хороших» заемщиков по категориям независимой переменной.

Расчет статистики хи-критерия между зависимой и каждой анализируемой независимой переменной:

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}, \qquad (11)$$

где f_{ij} – фактические частоты;

e_{ij} – ожидаемые частоты;

m и k– число строк и столбцов в таблице сопряженности.

Альтернативой формулы (11) является расчет критерия хи-квадрат на основе формулы логарифма правдоподобия:

$$\chi^2_{LR} = 2\sum_{i=1}^{m}\sum_{j=1}^{k} f_{ij} \ln\left(\frac{f_{ij}}{e_{ij}}\right) \quad (12)$$

При достаточном числе наблюдений значение по альтернативной формуле (12) будет мало отличаться от значения по классической формуле (11). Если расчетное значение критерия будет превышать критическое значение по таблице распределения хи-квадрат с заданным уровнем значимости и числом степеней свободы $\chi^2_{кр} = \chi^2(\alpha; \nu = (m-1)(k-1))$, то проверяемая гипотеза будет отвергаться, будет доказано наличие взаимосвязи между анализируемой независимой переменной и вероятностью дефолта по кредиту[9].

Для формирования конечных классов используется метод WOE-анализа (weightsofevidence)[9], где каждому наблюдению, содержащему набор признаков, ставится в соответствие бинарная выходная переменная (событие или несобытие в зависимости от логики решения задачи).

Затем производится разбиение всего диапазона изменения того или иного признака на несколько начальных классов, для каждого из которых вычисляется коэффициент WOE:

$$WOE_i = \ln\frac{F^-}{F^+}, \quad (13)$$

где i– индекс начального класса;

F^-– относительная частота появления не событий в классе;

F^+– относительная частота появления событий в классе.

На основе коэффициентов WOE вычисляется величина, определяющая значимость признака в модели бинарной классификации, называемая информационным индексом (informationvalue, IV) по формуле:

$$IV = \sum_{i=1}^{k}\left\{\left(\frac{N_i}{N} - \frac{P_i}{P}\right) \times WOE_i\right\}. \quad (14)$$

Информационный индекс всегда является положительной величиной. На основе информационного индекса определяется значимость признака по следующей методике:

$IV < 0{,}02$ – отсутствует;

$0{,}02 \le IV < 0{,}1$ – низкая;

$0{,}1 \le IV < 0{,}3$ – средняя;

$IV > 0{,}3$ – высокая;

Коэффициенты WOE и вычисленные на их основе значения информационного индекса являются критерием для формирования конечных классов оптимальным образом:

1) максимизируя значимость признака в бинарной классификационной модели;

2) максимизируя равномерность заполнения интервалов, что обеспечивает наилучшую репрезентативность результатов;

3) компромисс между этими вариантами.

3.7 Оценка качества модели

Оценить модель логистической регрессии можно методом максимального правдоподобия. Значение функции правдоподобия оценивает качество модели логистической регрессии. Как правило, значение функции правдоподобия преобразуется через $-2 \cdot \log$ правдоподобия, так как в нем есть распределения χ^2, благодаря которому можно проверить гипотезу значимости модели в целом. В достоверной модели функция правдоподобия близка к единице, а $-2 \cdot \log$ функции правдоподобия близок к нулю.

После того, как проверили значимость уравнения логистической регрессии, проверяется значимость отдельных коэффициентов. При довольно хорошей значимости уравнения регрессии будет хотя бы один предиктор, который объясняет изменение зависимой переменной. В конечном итоге важно получить значимое уравнение со всеми значимыми коэффициентами. Если же какой-то коэффициент при определенном предикторе не имеет значимости, то его следует убрать из всех независимых переменныхпеременных и заново пересчитать уравнение. В таких случаях можно использовать алгоритмы пошагового отбора переменных для построения модели. Для того, чтобы проверить гипотезу о зна-

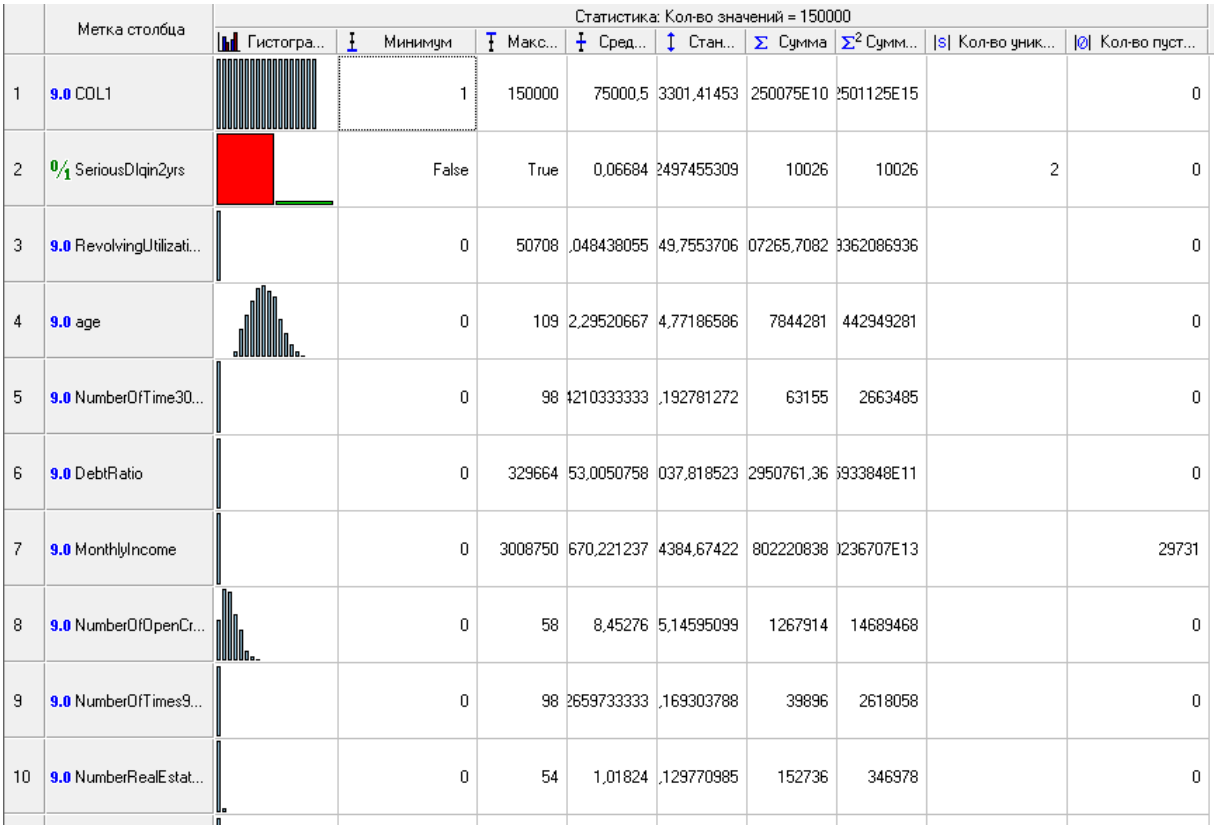
чимости определенных коэффициентов, используют статистический показатель Вальда, которыйимеет распределение χ^2. Во время проверки на значимость коэффициентов следует обратить внимание на конкретные значения стандартных ошибок коэффициентов и доверительные интервалы. Чем меньше стандартная ошибка и уже доверительный интервал – тем модель сама по себе лучше.

Иногда, для оценки качества подогнанной модели используют значение коэффициента детерминации. Но все же данный коэффициент не является основным значением, определяющий точность модели логистической регрессии. Такой коэффициент больше применим в модели линейной регрессии. Такие коэффициенты используют для того, чтобы оценить меру зависимости между переменными на первоначальном этапе построения логистической регрессии и отбора предикторов. Псевдо коэффициенты детерминации не следует рассматривать, как главные меры качества модели. Псевдо коэффициенты детерминации путают качество подгонки модели. Чаще всего такое поведение можно заметить на малых выборках, когда значение коэффициента детерминации может быть достаточно высоким при плохой подгонке. Относительно низкие значения коэффициентов детерминации в модели логистической регрессии – то это нормальное явление. В модели логистической регрессии оценка коэффициента детерминации может быть построена на основе логарифма функции правдоподобия, сравнивая фактические значения, зависимую переменную и расчет значений вероятности зависимой переменной.

Достоверность является немаловажным качеством любой регрессионной модели. За достоверность модели можно считать способность отличать «плохих» заемщиков от «хороших». Такую способность модели можно оценить, анализируя таблицу классификации. Очень важно построить такую модель, которая бы одинаково хорошо отличала «хороших» заемщиков от «плохих». Для верной оценки качества классификации модели строят ROC-кривую, которая способна показать зависимость количества верно классифицированных положительных исходов. Для того, чтобы сравнить две и более моделей между со-

бой, следует сравнить их площади под ROC-кривыми – это показатель называется AUC измерения которого имеет интервал от 0.5 до 1.

Если рассмотреть все вышесказанное на практическом примере аналитической платформы, построенной нами скоринговой карты, то это будет выглядетьпримерно, как показано на рисунке 1.



	Метка столбца	Статистика: Кол-во значений = 150000								
		Гистогра...	Минимум	Макс...	Сред...	Стан...	Сумма	Σ² Сумм...	|s| Кол-во уник...	|Ø| Кол-во пуст...
1	9.0 COL1		1	150000	75000,5	3301,41453	250075E10	2501125E15		0
2	0/1 SeriousDlqin2yrs		False	True	0,06684	2497455309	10026	10026	2	0
3	9.0 RevolvingUtilizati...		0	50708	,048438055	49,7553706	07265,7082	3362086936		0
4	9.0 age		0	109	2,29520667	4,77186586	7844281	442949281		0
5	9.0 NumberOfTime30...		0	98	4210333333	,192781272	63155	2663485		0
6	9.0 DebtRatio		0	329664	53,0050758	037,818523	2950761,36	6933848E11		0
7	9.0 MonthlyIncome		0	3008750	670,221237	4384,67422	802220838	0236707E13		29731
8	9.0 NumberOfOpenCr...		0	58	8,45276	5,14595099	1267914	14689468		0
9	9.0 NumberOfTimes9...		0	98	2659733333	,169303788	39896	2618058		0
10	9.0 NumberRealEstat...		0	54	1,01824	,129770985	152736	346978		0

Рисунок 1 – Исходная статистика базы данных клиентов

На рисунке 1 мы видим статистические данные клиентов банка, такие как возраст(age), ежемесячный доход(MonthlyIncome), просрочки по кредиту в 30,60,90 дней, открытые кредиты и др. Эти данные представлены в первоначальном виде и считаются недостаточно корректными для построения скоринговой карты.

Рисунок 2 показывает разбиение исходной базы на обучающую и тестовую выборки в соотношении 70% на 30%. Это делается для того, чтобы проверить скоринговую модель на адекватность и то, насколько точно она делает предсказания.После разбиения на тестовую и обучающую выборки происходит процесс фильтрации данных после которого дело переходит к следующему обработчику.

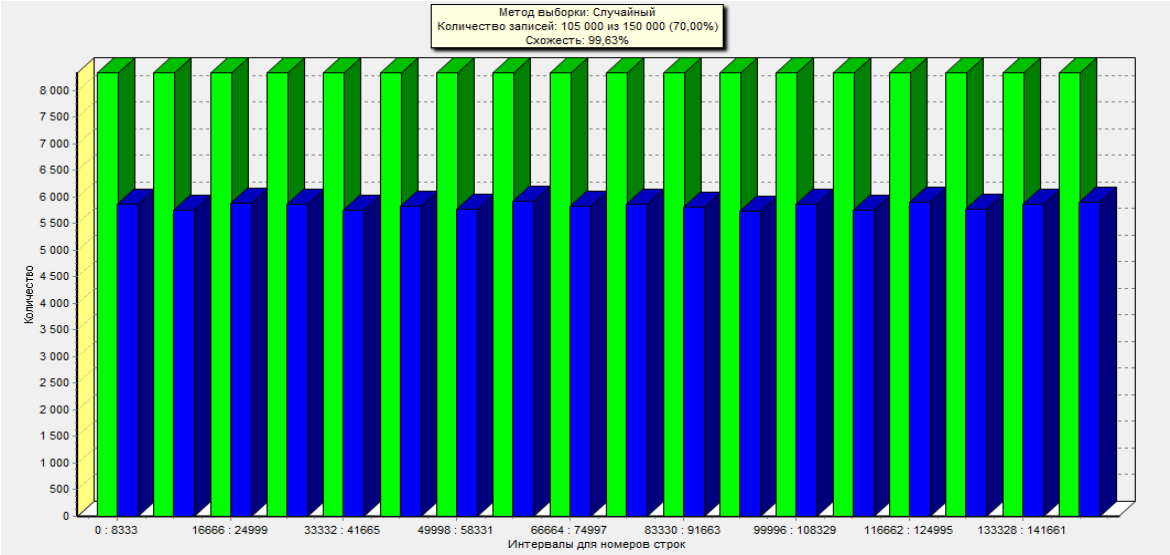


Рисунок 2 – Разбиение на множества

	Метка столбца	Статистика: Кол-во значений = 45000								
		Гистогра...	Мини...	Макс...	Сред...	Стан...	Сумма	Σ² Сумм...	|s| Кол-в...	|∅| Кол-в...
1	9.0 COL1		1	149996	4905,85853	3221,11467	3370763634	6509898E14		0
2	0/1 SeriousDlqin2yrs		False	True	0,0662	2486341363	2979	2979	2	0
3	9.0 RevolvingUtilizationOfUnsecuredLines		0	3,31130592	3434733442	5340408956	5456,30049	3398,77879		0
4	9.0 age		,026769327	6,41078902	2,23056371	4,74914923	350375,367	132550396		0
5	9.0 NumberOfTime30-59DaysPastDueNotW...		0	2,32694996	2200446091	5488853711	9902,00741	5735,96385		0
6	9.0 DebtRatio		0	302,442437	48,6038831	39,0409382	1187174,74	6756614E10		0
7	9.0 MonthlyIncome		0	20165	273,896983	107,833751	226249273	7968696E12	8090	8938
8	9.0 NumberOfOpenCreditLinesAndLoans		0	2,09509543	,353913074	,851318338	75926,0883	199518,361		0
9	9.0 NumberOfTimes90DaysLate		0	,525906758	3657489973	2815388374	995,870488	766,255777		0
10	9.0 NumberRealEstateLoansOrLines		0	,776793353	9799448219	3647969812	4097,51699	5099,69727		0

Рисунок 3 –Редактирование выбросов и экстремальных значений

Обработчик «Редактирование выбросов и экстремальных значений», показанный на рисунке 3, предназначен для корректировки неопределенных значений в наборе данных, отклонений от нормального поведения. Для улучшения обработки такого поведения данных в узле предусмотрена возможность их разделения. Выбросы – это события вызванные исключительными условиями. Экстремальные значения – это ошибки или фиктивные значения. Для того, что-

бы избежать нелогичного построения скоринговой карты, следует провести на первоначальном этапе редактирование выбросов и экстремальных значений.

Каждому типу отклонений присваивается определенный порог обнаружения, это позволяет провести очистку данных более соответствующей логике решаемой задачи.

Для каждого столбца исходных данных пользователь выбирает удобный метод редактирования выбросов и экстремальных значений. В таблице 3 представлены несколько методов и алгоритмов для борьбы с выбросами.

Таблица 3 – Алгоритмы поиска и набор методов

Метод	Неупорядоченное поле		Упорядоченное поле	
	непрерывное	дискретное	непрерывное	дискретное
Оставить без изменения	+	+	+	+
Удалять записи	+	+		
Ограничивать	+	+	+	+
Заменять наиболее вероятным	+	+	+	+
Заменять средним	+		+	
Заменять медианой	+		+	
Заменять заданным значением		+		+
Сглаживать			+	

В данном обработчике был использован метод «Ограничивать», который приводит экстремальное значение или выброс к значению, превышение которого определяется как выброс.

Далее, переходим к обработчику «Конечные классы». Он позволяет уменьшить количество значений набора данных, объединяя их в пределах некоторого интервала и используя информацию о бинарной выходной переменной.

Целью данного обработчика является подготовка выборки для повышения качества логистической регрессии. Поставленная задача выполнима за счет объединения значений в пределах некоторого интервала. Обработчик «Конечные классы», представленный на рисунке 4, решает такие задачи как: понижение значения признаков без ущерба информативности данных; понижение раз-

мерности данных, исключая признаки с низкой значимостью. Ограничениями являются – минимальное число наблюдений, выраженное в процентах и максимальное количество конечных классов.

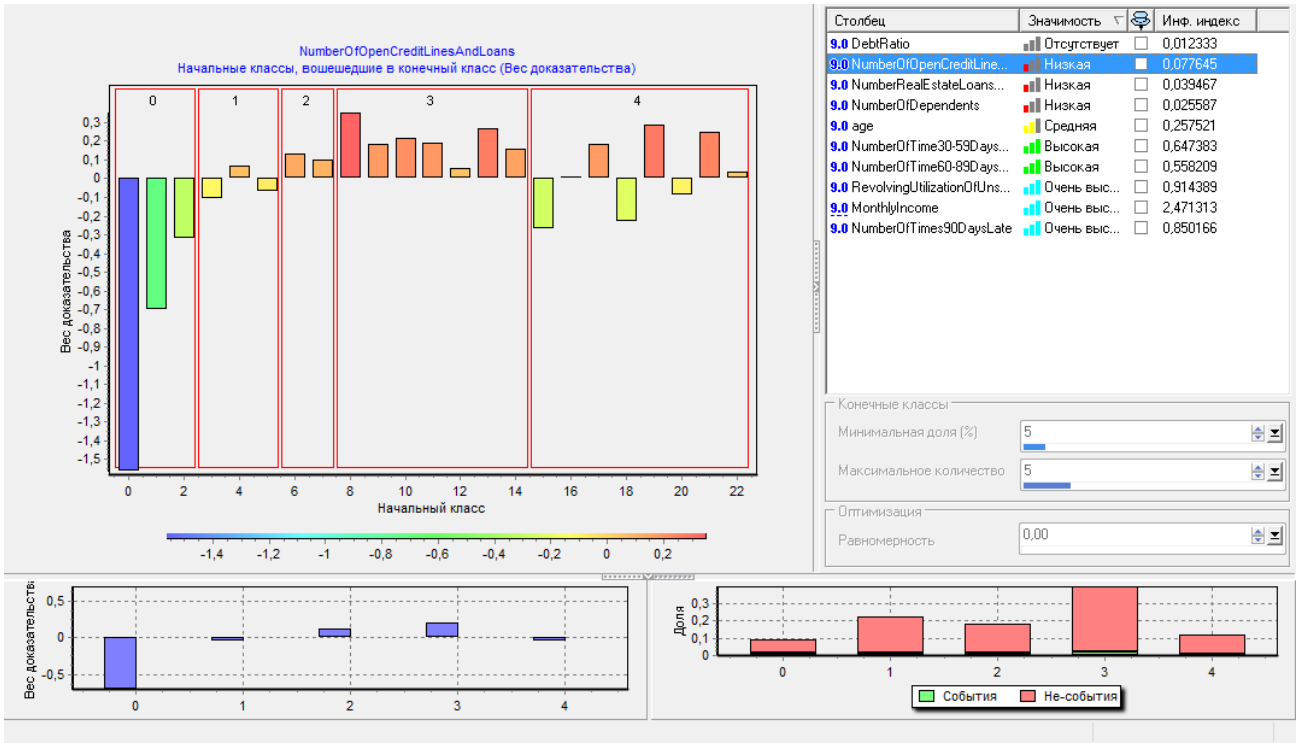


Рисунок 4 – Разбиение на конечные классы

Пользователь может по-своему усмотрению задавать параметры минимальной доли и максимального количества в интерактивном режиме. Так же в данном режиме можно изменять границы конечных классов. Внесенные изменения пользователем автоматически пересчитываются, и происходит изменение показаний целевой функции, веса доказательства и информационного индекса.

После разбиения на классы всех полей информационной базы можно детально рассмотреть полученные результаты из таблицы 4.

Поле «Метка столбца» содержит признаки, по которым происходит оценка их значимости. Значимость признаков в модели классификации рассчитывается на основекоэффициентов WOE и называется информационным индексом(IV).Информационный индекс – величина положительная. Поэтому, значимость столбцов определяется по следующей методике:

1) IV<0,02 – отсутствует;

2) 0,02≤IV<0,1– низкая;

3) 0,1≤IV<0,3– средняя;

4) IV>0,3– высокая.

Таблица 4 – Значимость столбцов

Метка столбца	События	Не события	Всего	Информационный индекс	Значимость столбцов
RevulvingUtil	2979	42021	45000	0,91389024645255	Очень высокая
Age	2979	42021	45000	0,25752135999619	Средняя
PustDue30-59	2979	42021	45000	0,6473831692798	Высокая
DebtRatio	2979	42021	45000	0,1233281207382	Отсутствует
MonthlyIncome	2462	33600	36062	2,4713131693482	Очень высокая
OpenCreditLine	2979	42021	45000	0,0776449135415	Низкая
90DaysLate	2979	42021	45000	0,8501656544926	Очень высокая
RealEstate	2979	42021	45000	0,0394671379883	Низкая
PustDue60-80	2979	42021	45000	0,5582087535282	Высокая
Dependents	2923	40921	43844	0,0255868680361	Низкая

Поле «События» определяет количество человек, которые выполнят данный признак из всех заявленных. Соответственно ему рассчитываются все «Не события».

На рисунке 5 показан график ROC-кривой, на котором по умолчанию отображаются:

1) положение текущего порога отсечения равное 0.0993;

2) значения чувствительности равное 72.22;

3) значение специфичности равное 71.97;

4) индекс AUC равный 0.77 на обучающем и тестовом множествах;

5) индекс KS равный 45.42 на обучающем и 48.76 на тестовом множествах.

Рассчитав индекс Gini, который равен 52.92% на обучающем и 53.55% на тестовом множествах следует, что чем ближе ROC-кривая к идеальной линии, тем ближе карта подходит к идеалу прогнозируемого дефолта. Обратив внима-

ние на индекс KS, заметим, что оба множества достаточно близки друг к другу, что так же говорит нам о хорошей предсказательной способности построенной модели, как и индекс AUC.

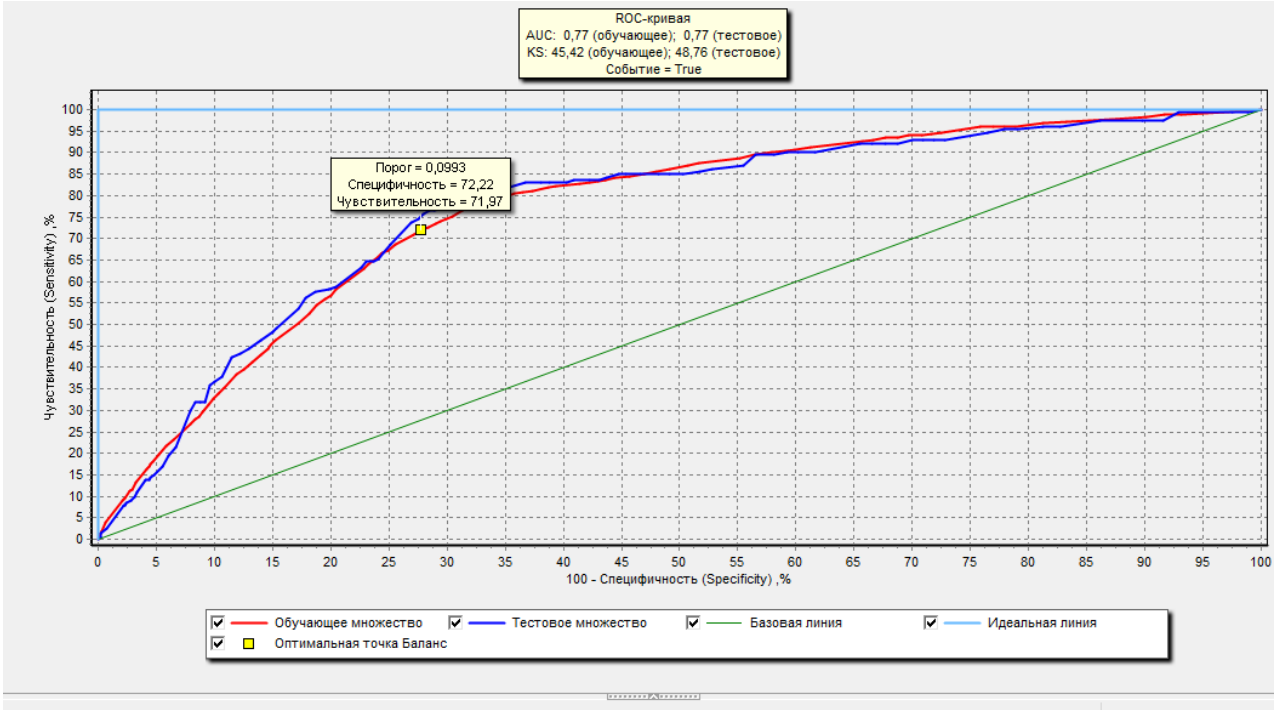


Рисунок 5 –ROC-кривая скоринговой модели

В общем случае можно сделать вывод, что данная скоринговая модель с высоким показателем специфичности соответствует консервативной кредитной политике. То есть чаще происходит отказ в выдаче кредита. А высокая чувствительность ведет к политике рискованных кредитов. В случае со специфичностью минимизируется кредитный риск, связанный с потерей процентов и с дополнительными расходами на возвращение кредита. А в случае с чувствительностью минимизируется риск, связанный с упущенной выгодой. Это хорошо показывает таблица сопряженности (см. рисунок 6).

	Классифицировано		
Фактически	False	True	Итого
False	68,75%	31,25%	100,00%
True	23,26%	76,74%	100,00%
Итого	65,74%	34,26%	100,00%

Рисунок 6 – Таблица сопряженности

Из рисунка 6 можно заметить, что модель достаточно неплохо определяет «хороших» заемщиков, коим реже отказывала в выдаче кредита, чем выдавала кредит «плохим». Так же, смотря на таблицу, можно понять, что данная модель на 23.26% имеет ошибку первого рода, упуская «хороших» клиентов, считая их за «плохих». А так же допускает ошибку второго рода на 31.25%, ложно допуская, что «плохие» клиенты окажутся «хорошими».

Таблица 5– Скоринговая карта

Показатель	Значение	Балл
Коэффициент задолженности	От 863.41	0
	До 863.41	4
Количество иждивенцев	От 2	5
	До 2	1
Открытые кредиты	До 2	8
	От 2-5	1
	От 5-7	1
	От 7-14	0
	От 14	11
Займы под залог недвижимости	От 1	0
	До 1	5
Беззалоговые кредиты	От 1	53
	До 1	0
Возраст	До 36	24
	От 36-52	21
	От 52-61	15
	От 61	0

Таблица 5 представляет собой результирующую скоринговую карту, по которой будут оцениваться заемщики по кредиту. Она содержит в себе «Показатель», по которому будут оцениваться клиенты, «Значение», которое показывает интервал распределения баллов и само поле «Балл». Данная карта имеет минимальный балл равный 355, с которого начинается суммирование баллов, исходя из признаков заемщика.

ЗАКЛЮЧЕНИЕ

Один из основных инструментов снижения рисков – использованиеавтоматизированных систем скоринга. В основе работы скоринговых систем лежитавтоматический расчет баллов в зависимости от параметров запрашиваемого кредита,кредитной истории, социально-демографических характеристик заемщика. В зависимости отколичества набранных баллов скоринговая система выдает решение: выдавать или не выдаватькредит клиенту банка. Большинство скоринговых систем строится на основе моделилогистической регрессии. Коэффициенты полученного уравнения логистической регрессиимасштабируются в скоринговые баллы. В работе рассматривалась методика построенияскоринговых карт на базе модели логистической регрессии.

На практическом примере мы разобрали пошаговое построение скоринговой карты. Использовав исходную базу данных клиентов, мы разбили ее на обучающее и тестовое множество. Затем,для корректировки неопределенностей, произвели редактирование выбросов и экстремальных значений. Следующим шагом было разбиение полей на конечные классы, путем установки допустимого интервала, максимального количества классов и минимальной доли событий. Конечным итогом работы стало построение логистической регрессии, которая детально показала ход проделанной работы и пригодность скоринговой карты.

В конечном итоге из таблицы сопряженности можно сделать вывод о том, что данная модель более ориентирована на определение «хорошего» заемщика. Так же следует заметить, что изъяном модели является ложное обнаружение «хороших» клиентов в 31,25% принимая их за «плохих».

БИБЛИОГРАФИЧЕСКИЙСПИСОК

1 Allison, P.D. Logistic regression using the SAS system: theory and application. [Text] – Cary, NC: SAS Institute, 1999. – 303 p.

2 Anderson R. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. [Text] – New York: Oxford University press, 2007. – 790 p.

3 BaseGroupLabsOOO «Аналитическиетехнологии» [Электронныйресурс].– Режимдоступа: https://basegroup.ru/deductor, свободный.

4 BaseGroupLabs ООО «Аналитические технологии» [Электронный ресурс]. – Режим доступасвободный:https://basegroup.ru/community/articles/logistic.

5 Harrell, Frank.Regression modeling strategies. [Text] – NY: Springer, 2001 – 608 p.

6 Jaccard, J. Interaction effects in logistic regression. [Text] – Thousand Oaks: Sage Publications, 2001. – 70 p.

7 Lyn C. Thomas. Consumer credit models: pricing, profit, and portfolios. [Text] –New York: Oxford University press, 2009. – 400 p.

8 Mays E. (ed.) Handbook of credit scoring. [Text] – Chicago: Glenlake Publishing Company Ltd/Fitzroy Dearborn Publishers, 2001. – 382 p.

9 Naeem, S. Credit risk scorecards: developing and implementing intelligent credit scoring. [Text] – NewJersey: JohnWileyandSons, 2006. – 208 p.

10Журавель,Ю.Ю.Чтоможетичегонеможетскорингвпотребитель-ском кредитовании[Текст] / Ю.Ю. Журавель // Банковский ритейл.2006.№4. Справочно-правовая система Гарант.

11Ковалев,М.Методикапостроениябанковскойскоринговой модели для оценки кредитоспособности физических лиц [Текст] / М. Ковалем, В. Корженевская // Банки Казахстана. – № 1. – 2008. – С. 43-48.

12 Национальные особенности кредитногоскоринга /А.С. Пищулин, «Банковское кредитование», № 1, январь-февраль 2008.

13 Ниворожкина, Л.И. Эконометрическое моделирование риска невыплат по потребительским кредитам. [Текст] // Прикладная эконометрика. –30 (2). – 2013. С. 65–76.

14 Руководство по кредитному скорингу [Текст] / под.ред. Элизабет Мэйз ; пер. с англ. И. М. Тикота ; науч. ред. Д. И. Вороненко. – Минск: ГревцовПаблишер, 2008. – 464 с.

15 Сорокин, А.С. К вопросу оценки согласованности мнений экспертов при использовании методов экспертного оценивания в кредитном скоринге. [Текст] /А.С. Сорокин // Роль бизнеса в трансформации общества – 2014: Сб. ст. по мат. IX междунар. научн. конгр. – М.: «Эдитус», 2014. – с. 281-283.

16 Сорокин, А.С. К вопросу валидации модели логистической регрессии в кредитномскоринге / Интернет-журнал \"Науковедение\", Вып. 2 (21), 2014.

17 Сорокин, А.С. Применение законов распределения случайных величин для моделирования экономических явлений и процессов [Текст]: монография. / Н.Я. Бамбаева, А.С. Сорокин – М.: МЭСИ, 2010. – 156 с.

18 Улитина, Е.В. Статистика: учебное пособие [Текст] / Е.В. Улитина, О.В. Леднева, О.Л. Жирнова – М.: Московский финансово-промышленный университет «Синергия», 2013. – 320 с.