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 01.04.02 Прикладная математика и информатика
Направленность (профиль) образовательной программы Математическое и программное обеспечение вычислительных систем

ДОПУСТИТЬ К ЗАЩИТЕ
И.о. зав. кафедрой
_____ Н.Н. Максимова
«15» 06 2018 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему: Скоринговая модель на основе log-регрессии для оценки кредитоплатёжности заемщиков

Исполнитель
студент группы 652 ом — 10.06.18 (подпись, дата) — В.С.Карелин

Руководитель
доцент, канд. физ.-мат. наук — 14.06.18 (подпись, дата) — Н.Н. Максимова

Руководитель научного содержания программы магистратуры — 14.06.2018 (подпись, дата) — А.Г. Масловская

Нормоконтроль
доцент, канд. техн. наук — 11.06.2018г (подпись, дата) — А.В. Рыженко

Рецензент
доцент, канд. эконом. наук — 13.06.18 (подпись, дата) — Е.А.Самойлова

Благовещенск 2018

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
И.о. зав. кафедрой
_____ Н.Н. Максимова
«_26_» ____02_____ 2018 г.

З А Д А Н И Е

К магистерской диссертации студента Карелина Вячеслава Сергеевича.

1. Тема магистерской диссертации: Скоринговая модель на основе log-регрессии для оценки кредитоплатёжности заёмщиков (утверждена приказом от 13.02.2018 № 319-уч).

2. Срок сдачи студентом законченной работы: 14.06.2018 г.

3. Исходные данные к магистерской диссертации: научные публикации по данной тематике, учебные пособия.

4. Содержание магистерской диссертации (перечень подлежащих разработке вопросов): данная магистерская работа исследует вопросы кредитного скоринга, рассматривает способ построение скоринговой карты в виде модели логистической регрессии;

5. Консультанты по магистерской диссертации: рецензент – Самойлова Е.А., канд. эконом. наук, доцент; нормоконтроль – Рыженко А.В., канд. техн. наук, доцент.

6. Дата выдачи задания: 26.02.2018 г.

Руководитель магистерской диссертации: Максимова Надежда Николаевна, доцент, канд. физ.-мат. наук, доцент.

Задание принял к исполнению (26.02.2018): _____ Карелин В.С.

РЕФЕРАТ

Магистерская диссертация содержит 72 страницы, 16 рисунков, 13 таблиц, 30 источников.

КРЕДИТНЫЙ СКОРИНГ, СКОРИНГОВАЯ МОДЕЛЬ, БАНК, КЛИЕНТ БАНКА, МАТЕМАТИЧЕСКИЙ АНАЛИЗ, ЛОГИСЧТИЕСКАЯ РЕГРЕССИЯ, МЕТОД БАЙЕСОВСКОГО, ВИЗУАЛИЗАЦИЯ МОДЕЛИ, АПРОБАЦИЯ, РЕЙТИНГ, СКОРИНГОВАЯ КАРТА

Актуальность данной работы подтверждается в активном развитии банковского сектора. Современные тенденции развития информационного общества обязывают банки идти в «ногу со временем», что и обуславливает активное продвижение и использование скоринговых моделей, кроме того, скоринговая карта является дополнительным уровнем защиты от мошенников и потенциальных «дефолтных» клиентов.

Целью работы является построение модели скоринговой карты на основе логистической регрессии.

В работе исследована выборка клиентов банковского учреждения и создана модель скоринговой карты в аналитической платформе DEDUCTOR 5.3. Данная модель будет апробирована в реальных условиях на соискателях кредитов банковского учреждения.

Результаты магистерской диссертации обсуждены на вузовских конференциях и региональных научно-практических конференциях.

По результатам работы опубликованы двое материалов в сборниках докладов межвузовских конференциях.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Скоринг – это эвристический способ построения рейтингов и классификации различных объектов на группы. Он основывается на предположении о том, что люди со схожими социальными показателями ведут себя одинаково. Он применяется в банковской сфере, маркетинге, страховом деле.

Основной целью традиционного скоринга является классификация клиентов банка на «хороших» и «плохих», исходя из которой кредитор может выбирать соответствующие действия по отношению к данному клиенту. «Плохого» клиента, к примеру, можно определить как клиента с низкой эмпирической вероятностью возвращения кредита. Но, как правило, такое определение «плохого» клиента расширяется до любого нежелательного банку поведения клиента. Классификация осуществляется на основе скоринговой карты с помощью которой рассчитывается скоринговый балл клиента.

При рассмотрении степени изученности темы можно выделить несколько работ.

«Руководство по кредитному скорингу» под редакцией Элизабет Мейз, 2012 г. – единственная книга о скоринге на русском языке. Описываются общие понятия, разбираются методы построения скоринговой карты, обсуждается применение скоринга на практике. Книга состоит из статей написанных зарубежными специалистами в области финансов.

Диссертационная работа Сэмюэла Глассона «Метод цензурированной выборки для кредитного скоринга», 2015 г. В ней исследуются инструменты анализа выживаемости, применительно к кредитному скорингу, в условиях цензурированных данных. Разбирается применение метода линейной регрессии и в частности метода Бакли-Джеймса. Практическая часть работы содержит в себе применение этих методов к оценке времени кредитного дефолта и времени выплаты очередного платежа.

Диссертационная работа Кристины Болтон «Логистические регрессии и их применение в кредитном скоринге», 2011 г. Разбирается концепция кредитного скоринга применительно к банковскому делу в Южной Африке. Рассматриваются методы построения скоринговой модели с особым акцентом на метод логистической регрессии. Применяется этот метод для создания скоринговой модели.

Диссертационная работа Маттиаса Кремпля «Адаптивные модели и их применение в кредитном скоринге», 2011. Акцент ставится на изучении методов построения предсказывающих моделей в условиях дрейфа и задержки данных. Представлен новый метод для построения скоринговых моделей, базирующийся на методе дерева принятия решений. Представленный метод применяется для оценки дрейфа в двух наборах реальных финансовых данных.

В приведённых выше работах имеется общая проблема они не являются «гибкими» и не способны «адекватно» реагировать на изменения категорий дефолтности клиентов. Отсюда вытекает сомнение в правильности и актуальности полученных данных при исследовании современного скоринга. Задача состоит в создании скоринговой модели на основе логистической регрессии, применение которой отвечало бы современным тенденциям в кредитном скоринге. В данной работе представляется практическое и теоритическое решение данной задачи.

Существует множество подходов к построению скоринговой модели. В главе 1 раскрывается и описывается суть кредитного скоринга, как явления в банковском секторе. Рассматриваются проблемы и вероятные тенденции развития . В главе 2 рассматриваются виды математического моделирования и математического анализа в кредитном скоринге и разбирается эмпирический Байесовский подход (как вспомогательный) и метод логистической регрессии (как основной) к построению скоринговой модели: подход описывается теоретически и практически в рамках математического анализа, а метод логической регрессии осуществляется визуально с помощью аналитической платформы DEDUCTOR 5.3. в главе 3 Исследование строится на основе

обезличенных данных 15 000 клиентов банковского учреждения после применяется к реальным данным банка для построения скоринговой модели. Также в главе 3 проводится анализ полученной модели и строится вывод о её «актуальности и новизне». Данная скоринговая модель пройдет в апробацию в ОАО «АТБ» – Азиатско-Тихоокеанском Банке».

Новизной данного исследования выступает комбинированный анализ скоринговой карты, построенной с помощью логистической регрессии и дополненный выявлением рейтингов заёмщиков с помощью эмпирического метода Байесовского. Что будет подразумевать под собой добавление в скоринговую карту помимо выставляемых баллов – обобщенного рейтингового показателя, просчитанного с помощью метода Байесовского и реализованного на практике. Данное направление, по мнению автора исследования, является малоизученным и практически не рассматривалось в аналогичных научных трудах.

Кроме того данное исследование получит реализацию в рамках обучению студентов вузов АмГУ по направлению «Прикладная Математика и Информатика». Данные и методики проводимого исследования дадут обучающимся возможность наглядно увидеть, закрепить и использовать в дальнейшем навыки в математическом и статистическом анализе, а также в умении построения математических моделей в различных прикладных пакетах программ.

1 ОПИСАТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ КРЕДИТНОГО СКОРИНГА

Наибольшее распространение в банковской сфере получил кредитный скоринг. Кредитный скоринг можно опередить как метод начисления потенциальным заемщикам определенного количества баллов на основе информации о его социально демографическом положении, кредитной истории, параметрах запрашиваемого кредита, и принятие решения о выдаче или об отказе в кредите на основе набранного суммарного количества баллов. На настоящий момент банки предъявляют повышенные требования к риск аналитике в связи с участившимися случаями мошенничества и ростом числа невозвратных кредитов.[1] По данным Национального бюро кредитных историй по состоянию на 1 января 2018 года потери кредиторов от мошенников составили 153 млрд руб., тогда как годом ранее их объем был 67 млрд руб. На практике возникает задача не только принятия решения в отказе или выдачи кредита конкретному заемщику на основе набранного количества баллов, но и задача определения оптимального минимального количества набранных баллов для выдачи кредита. Вторая задача решается на основе анализа распределения баллов «надежных» и «ненадежных» заемщиков на основе полученной скоринговой карты и тесно связана с анализом соотношения риска и доходности во всем кредитном портфеле банка. Таким образом, кредитный скоринг является инструментом снижения рисков невозврата кредитов, а также помогает определить оптимальную структуру кредитного портфеля, корректировать процентные ставки по кредитам в зависимости от уровня риска.

В большинстве коммерческих банков скоринговые модели являются собственными разработками с различными методиками на основе данных о заемщиках конкретного банка прошлых лет, или являются готовыми решениями специализированных фирм на основе данных о заемщиках нескольких банков или финансовых институтов. И в первом и втором случае методики построения скоринговых карт, как правило, составляют коммерческую тайну [1].

1.1 История и виды кредитного скоринга

Повышение доходности кредитных операций непосредственно связано с качеством оценки кредитного риска. В зависимости от классификации клиента по группам риска банк принимает решение, стоит ли выдавать кредит или нет, какой лимит кредитования и проценты следует устанавливать. Для оценки кредитного риска производится анализ кредитоспособности заемщика, под которой в российской банковской практике понимается способность юридического или физического лица полностью и в срок рассчитаться по своим долговым обязательствам. В западной банковской практике кредитоспособность трактуется как желание, соединенное с возможностью своевременно погасить выданное обязательство. В соответствии с таким определением основная задача скоринга заключается не только в том, чтобы выяснить, в состоянии клиент выплатить кредит или нет, но и степень надежности и обязательности клиента. Иными словами, скоринг оценивает насколько клиент «достоин» кредита.

Скоринг представляет собой математическую или статистическую модель, с помощью которой на основе кредитной истории «прошлых» клиентов банк пытается определить, насколько велика вероятность, что конкретный потенциальный заемщик вернет кредит в срок [1]. Скоринг является методом классификации всей интересующей нас популяции на различные группы, когда нам неизвестна характеристика, которая разделяет эти группы (вернет клиент кредит или нет), на зато известны другие характеристики, связанные с интересующей нас.

В статистике идеи классификации популяции на группы были разработаны Фишером в 1936 г. на примере растений. В 1941 г. Дэвид Дюран впервые применил данную методику к классификации кредитов на «плохие» и «хорошие». По времени это совпало со Второй мировой войной, когда почти все кредитные аналитики были призваны на фронт, и банки столкнулись с необходимостью срочной замены этих специалистов. Банки заставили своих аналитиков перед уходом написать свод правил, которыми следовало

руководствоваться при принятии решения о выдаче кредита, чтобы анализ мог проводиться неспециалистами. Это и был как бы прообраз будущих экспертных систем. В начале 50-х гг. в Сан-Франциско образовалась первая консалтинговая фирма в области скоринга – Fair Issac, которая до сих пор является лидером среди разработчиков скоринговых систем. Но широкое применение скоринга началось с распространением кредитных карточек. При том количестве людей, которые ежедневно обращались за кредитными карточками, банкам ничего другого не оставалось, как автоматизировать процесс принятия решений по выдаче кредита. Однако очень скоро они оценили не только быстроту обработки заявлений на выдачу кредита, но и качество оценки риска. По данным некоторых исследований, после внедрения скоринг-систем уровень безнадежного долга сокращался до 50%.

В 1974 г. в США был принят Закон о предоставлении равных возможностей на получение кредита, который запрещал отказывать в выдаче кредита на основании следующих характеристик: раса, цвет кожи, национальное происхождение, возраст, пол, семейное положение, религия, получение социальных пособий, отстаивание прав потребителей. В Великобритании законодательство допускает использование информации о возрасте и семейном положении, но зато запрещает принимать во внимание какие-либо физические увечья и недостатки (инвалидность). Для кредитных организаций использование скоринговых систем стало доказательством исполнения этих антидискриминационных законов – у компьютера нет предубеждений [2].

Помимо установления принципов равноправия в области кредитования, кредитное законодательство США, как и Закон о потребительском кредите, принятый в Великобритании в том же 1974 г., имели важное значение для формирования службы кредитных бюро. В таких бюро записывается кредитная история всех людей, когда-либо обращавшихся за ссудой в любую кредитную организацию страны.

В кредитных бюро содержатся следующие виды данных:

– социально-демографические характеристики;

– судебные решения (в случае передачи дел о востребовании задолженности по кредиту в суд);

– информация о банкротствах;

– данные об индивидуальных заемщиках, получаемые от кредитных организаций по принципу «ты – мне, я – тебе», т. е. банк может получать информацию о клиентах других банков, только если сам поставляет аналогичную информацию.[2]

Объем и характер информации, хранящейся в бюро, строго регулируется законодательством каждой страны. В нашей стране Федеральный закон «О кредитных историях» вступил в силу с 1 июня 2005 года (за исключением части третьей статьи пятой данного закона). Часть третья статьи пятой ФЗ «О кредитных историях» вступила в силу с 1 сентября 2005 года, в соответствии с которой «кредитные организации обязаны представлять всю имеющуюся информацию», определенную статьей четвертой названного Федерального закона, «в отношении всех заемщиков, давших согласие на ее представление, … хотя бы в одно бюро кредитных историй, включенное в государственный реестр бюро кредитных историй». Содержание кредитных историй, порядок их предоставления, хранение и защита содержащейся в них информации регламентируются данным Федеральным законом.[3]

Значение кредитных бюро чрезвычайно велико. Во-первых, их существование позволяет кредитным организациям выдавать ссуды клиентам, которые ранее в этой организации не обслуживались. Во-вторых, добросовестный заемщик получает доступ к более дешевым кредитным ресурсам за счет более эффективной, быстрой и менее дорогостоящей процедуры оценки связанного с ним риска. В-третьих, дисциплина возврата кредитных средств повышается. Кроме того, общепризнанной является ценность предыдущей кредитной истории для прогнозирования вероятности дефолта, то есть благодаря кредитным бюро кредитные организации имеют

возможность гораздо более точно прогнозировать и составлять менее рискованные кредитные портфели. [3]

На сегодняшний день известно несколько видов скоринга:

1) application-scoring (дословный перевод с английского – «скоринг заявки, обращения») – оценка кредитоспособности заемщиков при выделении кредита. Это самый распространенный и известный клиентам вид скоринга. В его основе лежат первичный сбор анкетных данных заемщика, их обработка компьютером и вывод результата: предоставлять заем или нет;

2) collection-scoring – система скоринга на стадии работы с невозвращенными займами. Определяет приоритетные действия сотрудников банка для возврата «плохих» кредитов. Фактически программа позволяет предпринять ряд шагов по работе с невозвращенными долгами, например от первичного предупреждения до передачи дела коллекторскому агентству. Считается, что в процессе такой обработки порядка 40% клиентов ссылаются на забывчивость и возвращают кредит;

3) behavioral-scoring, «скоринг поведения» – оценка наиболее вероятных финансовых действий заемщика. Такая система дает возможность прогнозировать изменение платежеспособности заемщика, корректировать установленные для него лимиты. Основой анализа могут служить действия клиента за определенный период, например операции по кредитной карте;

4) fraud-scoring – статистическая оценка вероятности мошеннических действий со стороны потенциального заемщика. Такой скоринг, как правило, используется совместно с другими видами исследования клиентов. При этом считается, что до 10% невозвратов по кредитам связаны в России с откровенным мошенничеством и этот показатель растет.

Многие скоринговые системы не только обрабатывают введенные данные, но и способны к самообучению: они учитывают модель поведения уже принятых на обслуживание клиентов, чтобы корректировать свою оценку будущих заемщиков. Следует отметить, что на рынке программного обеспечения для банков существуют готовые решения. Самые известные

западные программы – SAS Credit Scoring, EGAR Scoring, Transact SM (Experian-Scorex), K4Loans (KXEN), Clementine (SPSS). Среди российских разработчиков выделяются Basegroup Labs, «Диасофт», известна украинская компания «Бизнес Нейро-Системы». В то же время многие банки разрабатывает свои собственные системы.

1.2 Современное состояние кредитного скоринга и тенденции развития

Важным моментом при предоставлении кредитных средств является анализ платежеспособности заемщика и рисков, сопровождающих каждую сделку. Оценка заемщика является сложной процедурой, которая требует централизованного и согласованного подхода, учитывающего бизнес-стратегию банка в целом и специфику самих кредитных продуктов. Грамотная оценка рисков, возникающих в работе банков при выдаче кредитов, и статистически обоснованные правила не избавят полностью от проблемных ситуаций, но позволят значительно сократить их число [4].

Актуальность создания, внедрения и использования скоринговых систем для управления кредитными рисками сегодня не вызывает сомнения. Современные информационные технологии кредитного скоринга позволяют аналитикам банка сконцентрироваться на решении бизнес-проблем, в то время как всю техническую работу по сбору исходных данных и реализации скоринговых алгоритмов берет на себя программное обеспечение. Объемы банковской информации неизбежно будут возрастать, а требования к качеству ее обработки ужесточаться. Реальным выходом из сложной ситуации является не копирование чужих методик принятия решений, а использование методов интеллектуального анализа данных для создания собственных. После их применения уже в ближайшей перспективе можно рассчитывать на увеличение точности и обоснованности принимаемых решений, ускорения работы, рационального использования человеческих ресурсов. Кредитный скоринг является одной из известных методик оценки кредитного риска, основанной на математической модели. Эта модель соотносит уровень кредитного риска с

параметрами, характеризующими заемщика. Как бы ни была сложна модель, она всегда разделяет потенциальных заемщиков на два класса – тех, кому кредит выдать можно, и тех, кому он «противопоказан». Таким образом, любая задача оценки рисков сводится к решению двух задач: задача классификации: отнесение объекта к одному из априори заданных классов (например, «низкий», «средний», «высокий» риск) и задача регрессии: численная оценка вероятности возникновения неблагоприятного события.

Для решения каждой из этих задач существуют соответствующий математический аппарат, возможность применения которого зависит от данных, используемых для анализа.

Следовательно, кредитный скоринг представляет собой модель, с помощью которой банк определяет, насколько велика вероятность, что потенциальный кредитополучатель исполнит свои обязательства в полном объеме в установленный срок. Механизм модели кредитного скоринга заключается в выделении главных факторов, обусловливающих кредитоспособность заемщика, определении удельных весов данных факторов, разработке шкалы оценки, выведении формулы конечного интегрального показателя, конкретизации границ диапазонов возможных значений интегрального показателя, качественно характеризующих должника. Наиболее сложной задачей является определение характеристик, которые следует включить в модель, и весовых коэффициентов, которые следует им присвоить. Значимость финансовых, экономических и мотивационных характеристик, влияющих на возвратность кредитов, приобретает особую важность. Подобный подход к анализу данных реализован в аналитической платформе Deductor Studio Academic, разработанной российской компанией BaseGroup Labs.[4]

Платформа активно используется для создания решений в области анализа рисков. С использованием аналитических возможностей платформы Deductor разработан алгоритм построения скоринговой модели оценки кредитного риска при кредитовании малого и среднего бизнеса.

Управление проблемными активами потенциально может рассматриваться как выгодный бизнес также и на уровне российских кредитных организаций. Средний показатель сомнительной задолженности по потребительским кредитам сегодня находится на уровне 5.8% от кредитного портфеля российских банков. Однако, по мнению экспертов, это далеко не «мертвые» кредиты, ведь некоторые заемщики просто забывают или не успевают вовремя сделать погашение. Проблемный кредит не всегда однозначно отрицательное явление, так как теоретически самым выгодным клиентом является тот, который платит весь долг с просрочкой, погашая еще и штрафные начисления. Кроме того, важно учитывать диверсификацию по типам кредитования – задолженность по потребительским кредитам и по ипотеке отличается в разы, впрочем, как и средняя доходность по данным видам продуктов. Единственная проблема, которая сегодня является весомой преградой в деле секьюритизации и торговле банковскими активами – это отсутствие в российских банках существенного опыта реструктуризации кредитной задолженности. Согласно доступной статистике, профессиональных мошенников в общей структуре задолженности отечественных банков пока около 1,5%, но реального положения дел не знает никто. Таким образом, создание фондов, которые могли бы управлять сомнительной и безнадежной задолженностью российских заемщиков, может оказаться бизнесом с большой долей риска. В этих условиях наиболее вероятными покупателями «плохих кредитов» могут выступить скорее коллекторские компании, чем специальные фонды. [5]

На данный момент, за некоторыми исключениями, ни у одного российского банка нет действующей классической системы скоринга, а кредитоспособность заемщиков обеспечивается преимущественно собственными программными продуктами. Такая ситуация объективна – намерение внедрить полное скоринговое решение зачастую наталкивается на отсутствие необходимых исторических данных и возможности применить какой-либо статический пакет. Поскольку вся доступная статистика содержится

на бумаге и в кредитных делах экспертов, то для создания необходимого «кредитного кладбища» из 10 тысяч заемщиков может понадобиться мобилизация значительных ресурсов, на что сегодня готово далеко не каждое финансовое учреждение. В ближайшее время в условиях финансового кризиса развитие скоринга будет определяться такими тенденциями [5]:

– рост интереса к «collection» скорингу;

– крайне актуальной будет становиться задача улучшения качества кредитного портфеля;

– все более важным станет вопрос управления бизнес-процессом скоринга;

– увеличение объемов продажи портфелей проблемных активов коллекторским агентствам.

2 ВИДЫ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ В КРЕДИТНОМ СКОРИНГЕ

В настоящее время для кредитного скоринга используются методы статистики (дискриминантный анализ, линейная регрессия, логистическая регрессия, деревья классификации), исследования операций (линейное программирование, нелинейная оптимизация) и искусственного интеллекта (нейронные сети, экспертные системы, генетические алгоритмы, методы ближайших соседей, байесовские сети, логико-вероятностные методы) [6].

Указанные методы могут применяться как по отдельности, так и в различных комбинациях. Обилие совершенно разных методов для решения одной и той же задачи объясняется чисто прагматическим подходом: использовать то, что работает, а не пытаться объяснить причину дефолтов или зависимость от макроэкономических показателей. Для построения модели берётся выборка данных по существующим заёмщикам, обычно не менее нескольких тысяч записей. По каждому заёмщика необходимы анкетные данные и кредитная история за определённый период, как правило от одного года до двух лет. Относительно каждой кредитной истории решается, является ли она «хорошей» или «плохой»; обычно признаком «плохой» кредитной истории является факт задержки платежей на три и более месяцев. По части клиентов такое решение принять трудно (например, по тем, кто пропустил платежи за два месяца). Таких клиентов помещают в группу «средние». В зависимости от используемого метода скоринга, при построении модели могут все три указанные группы, либо «средние» клиенты просто удаляются из выборки. На качество модели сильно влияет выбор периода, за который рассматривается кредитная история (т.е. промежуток времени между заполнением анкеты и классификацией клиентов на «плохих» и «хороших»). Считается, что вероятность дефолта как функция времени от прихода в кредитную организацию вначале сильно колеблется, и только после года начинает стабилизироваться, поэтому использование меньшего периода

приведёт к недооцениванию вероятности дефолта. Напротив, использование периода более двух лет не желательно в силу того, что за это время могут произойти определённые социально-экономические изменения, так что вновь приходящие клиенты будут слишком сильно отличаться по своим характеристикам от тех, на которых была построена модель. Соотношение числа «плохих» и «хороших» клиентов в обучающей выборке также влияет на качество системы скоринга. Для одних методов (например, линейная регрессия) предпочтительно использование равных долей «плохих» и «хороших», тогда как другие методы (деревья классификации, байесовские сети) требуют, чтобы выборка отражала реальное соотношение «плохих» и «хороших» клиентов.[6]

Проблему кредитного скоринга можно рассматривать как задачу классификации: зная ответы на вопросы анкеты $x \in A$, определить, к какой группе относится заёмщик: $x \in$ AG для «хороших» клиентов, и $x \in$ AB для плохих. При этом необходимо понимать, что абсолютно точная классификация принципиально невозможна. хотя бы потому, что один и тот же набор ответов может быть дан как «хорошим», так и «плохим» клиентом (напомним, что вопросы в анкете представляют собой вторичные характеристики заёмщиков). Хотелось бы, однако, построить такую модель, которая производила бы правильную классификацию как можно в большем числе случаев. Большинство статистических методов приводят к построению правила классификации, основанного на линейной скоринговой функции. Оно может быть получено использованием различных подходов.

2.1 Скоринговая карта как модель анализа данных

В зависимости от используемой модели система скоринга может выдавать на выходе следующие данные.

Класс клиента. В самом простом случае ¾хороший или ¾плохой; первым кредит выдается, вторым нет. В более сложных случаях может быть ещё несколько промежуточных классов кредит выдается, но на других условиях (больший процент, меньший срок, или дополнительные гарантии), либо кредитоспособность оценивается вручную, с использованием дополнительной

информации. Апостериорное распределение класса клиента. Для каждого класса указывается вероятность, с которой данный клиент принадлежит этому классу. Можно выбрать класс с наибольшей вероятностью, либо усреднить какие-либо показатели по каждому классу. Данный вариант предпочтительнее первого, поскольку в распределении содержится существенно больше информации. Например, если классов всего два, то в первом варианте клиент будет считаться «хорошим» как при распределении 90%/10%, так и при 55%/45%. Очевидно, однако, что во втором случае степень уверенности в кредитоспособности существенно ниже, чем в первом [7].

Вероятность дефолта. Для принятия решения остаётся только сравнить эту вероятность с пороговым значением допустимой вероятности дефолта. Последняя определяется так, чтобы с учетом этой вероятности и процентной ставки банк в среднем не терпел убытков в случае выдачи кредита.

Счёт – количественная оценка кредитоспособности потенциального заёмщика (чем больше счёт, тем последняя выше). Обычно счёт пропорционален вероятности или шансам успешного возврата кредита, поэтому либо по счёту определяется вероятность дефолта и на этом основании принимается решение, либо, в зависимости от того, в какой промежуток попал счёт, определяется класс клиента и на основании этого условия выдачи кредита. Если скоринговая система так или иначе определяет вероятность дефолта заёмщика, то необходимо установить пороговую (допустимую) вероятность дефолта. Это можно сделать, вычислив средние потери от дефолта, и введя ограничение, что эти потери не могут быть больше дохода в случае успешного возврата кредита. Потери в случае выдачи кредита неплатежеспособному заёмщику, потери в случае отказа «хорошему» клиенту образуют так называемую матрицу штрафов (misclass ification costs matrix). В этой матрице столбцам соответствуют фактические (реальные) классы клиентов, а строкам предсказанные. На диагонали находятся нули (т.е.за правильно предсказанный класс штраф отсутствует), а во всех остальных ячейках указывается штраф за

предсказание одного класса (строка), если на самом деле клиент принадлежит другому классу (столбец).

Элементы матрицы могут быть определены исходя из порогового значения вероятности дефолта p^* (выше которого кредиты выдаваться не должны). Если посчитать, что штраф за отказ хорошему клиенту равен 1, а n штраф за приём плохого клиента. Матрица штрафов используется в ходе оценки качества используемых моделей, а также является частью входных данных некоторых алгоритмов (например, при построении деревьев классификации).

Процесс построения скоринговой модели можно условно разбить на три этапа:

Формулировка задачи и подготовка данных. С помощью экспертов в конкретной области формулируется задача скоринга, производится сбор и предварительная обработка данных.

Анализ данных и построение модели. Производится поиск оптимальной модели для решения поставленной задачи. Необходимо оценить точность работы различных моделей и выбрать наилучшую из них.

Модель применяется для реального принятия решений, при этом производится оценка её точности на фактических данных. По прошествии времени модель должна перестраиваться, чтобы отражать произошедшие изменения. Несмотря на то, что кредитный скоринг предназначен для автоматического принятия решения по выдаче кредитов, сам процесс построения модели для скоринга не может обходиться без непосредственного участия человека на каждом из этапов. [7]

Наиболее очевидным критерием точности является процент неверной классификации, или интенсивность ошибок, который выступает, как отношение числа неверных случаев классификации к общему числу случаев.

Для задачи классификации с двумя классами это число должно быть между нулём (все случаи классифицированы корректно) и интенсивностью ошибок классификации по умолчанию (присваивающей во всех случаях класс,

которому принадлежит (большинство клиентов). По ряду причин построенная модель должна иметь меньшую интенсивность ошибок, чем классификация по умолчанию, при этом в реальных приложениях не существует моделей с нулевой интенсивностью ошибок. Реальная интенсивность ошибок (true error rate) определяется тестированием модели на настоящих данных. Она не может быть определена до тех пор, пока модель не будет протестирована на большом количестве реальных случаев. Следовательно, в ходе построения модели этот показатель необходимо как-либо оценить. Собственная интенсивность ошибок (apparent error rate) определяется как интенсивность ошибок на наборе данных, который был использован для обучения модели. Однако она не является надёжной оценкой реальной интенсивности, поскольку низкое её значение может означать, что модель является просто-напросто «подгонкой» (overfitting) результата классификации под данные в обучающем наборе (например, у метода ближайших соседей всегда будет нулевая собственная интенсивность ошибок). В этом случае можно ожидать весьма посредственных результатов при применении модели к реальным данным [8].

Для оценки собственной интенсивности ошибок применяется метод «удержания» тестовых данных: исходный набор данных разделяется на «обучающий» (использующийся для построения модели) и «тестовый» (используемый для оценки точности) наборы. Предполагается, что тестовый набор выделяется случайным образом, независимо от самих данных. Определённая таким образом интенсивность ошибок называется тестовой интенсивностью ошибок. Обычно величина тестового набора составляет около 30% от всех данных. При величине тестового набора в 1000 записей тестовая интенсивность ошибок уже является статистически точной оценкой реальной интенсивности.

Представляет также интерес точность классификации при условии изменений, происходящих в населении с течением времени. Для этого имеющиеся данные упорядочиваются по дате заполнения анкеты, и затем в качестве обучающего набора используется первая часть списка, а в качестве

тестового вторая. Это также отчасти решает проблему излишней «подгонки» под обучающие данные. Сравнение результатов точности классификации различными методами показывает, что они практически совпадают. Это можно объяснить эффектом «плоского максимума»: существенные изменения весов в окрестности оптимальной скоринговой модели приводят к незначительным отклонениям в точности прогнозов.

Хотя точность классификации и является важным критерием выбора скоринговой модели, необходимо также принимать во внимание ряд других качеств:

Скорость работы. Необходимо оценить время, требуемое для обучения и для принятия решения в соответствие с моделью. Приведём два крайних примера. С одной стороны, метод ближайших соседей исключительно быстр в обучении (просто добавляется ещё одна запись), но для принятия решения требуется полный перебор случаев в базе данных, что может занять много времени. Напротив, нейронные сети требуют минимальных вычислений для классификации одного случая, но при этом их обучение является NP-трудной задачей, поэтому алгоритмы обучения требуют экспоненциального времени.

Прозрачность и интерпретируемость. Прозрачность моделей становится важной, когда модель необходимо объяснить кредитным аналитикам. Часто счёт, выдаваемый системой, используется как один из критериев принятия решения квалифицированным кредитным офицером, поэтому модель должна быть в достаточной мере понятной. Наиболее прозрачными моделями являются, по всей видимости, основанные на линейной скоринговой функции. Напротив, нейронные сети действуют как «чёрный ящик» и не предоставляют никаких объяснений результатов классификации, что препятствует использованию таких систем на практике, когда кредиторам требуется объяснять, почему они не выдали тот или иной кредит. [8]

Простота модели. Следует предпочитать наиболее простую модель при одном и том же уровне точности. Это имеет влияние как на скорость работы

модели, так и на её понятность. Кроме того, более простые модели как правило являются более работоспособными.

2.2 Виды математического анализа для обработки кредитных данных

Линейный дискриминантный анализ – метод для классификации объектов на заранее определённые категории. Идея в том, чтобы найти такую линейную комбинацию объясняющих переменных, которая наилучшим образом разделила бы объекты на категории. Под разделением наилучшим образом имеется ввиду такое, при котором обеспечивается максимальная дистанция между средними данных категорий. Скоринговый балл рассчитывается как линейная функция от значений атрибутов клиента [9]:

$$Z = B^T x = B_i x_i + ... + B_k x_k , \quad (1)$$

где $x = (x_i ,..., x_k)$ – значения атрибутов клиента, $B = (B_i ,..., B_k)$ – параметры модели, которые максимизируют отношение

$$M = \frac{B^T (m_G - m_B)}{\sqrt{B^T \Sigma B}}, \quad (2)$$

где $(m_G - m_B)$ – вектор средних для хороших и плохих клиентов;

ΣB – общая ковариационная матрица.

Линейный дискриминантный метод предполагает выполнение двух условий. Во-первых, ковариационные матрицы независимых переменных для обеих групп должны совпадать. Во-вторых, независимые переменные должны быть распределены нормально. Часто, в скоринге, независимые переменные дискретные или распределены не нормально. Отсюда, возникают проблемы в применении этого метода. Однако было показано, что даже в случае нарушения нормальности, данный метод широко применим. Его преимуществом можно назвать простоту применения. [9]

Схожий метод линейной регрессии, также используется для формирования скоринговой модели. В случае двух категорий, он эквивалентен методу линейного дискриминантного анализа и выражает зависимость одной

переменный (зависимой) от других (независимых). В общем виде представляется так:

$$Y = B_0 + B_1 X_1 + ... + B_n X_n + \varepsilon, \quad (3)$$

где Y – зависимая переменная;

X_n – объясняющие независимые переменные;

B_n – неизвестные коэффициента регрессии, которые находятся методом наименьших квадратов;

ε – ошибка.

Для применения модели линейного скоринга требуется выполнение следующего предположения: связь между зависимой и независимыми переменными должна быть линейной. В противном случае, точность оценки значительно ухудшается. Ошибки же должны быть независимы и распределены нормально.

Как и в случае дискриминантного анализа, в условиях кредитного скоринга, предположения, требуемые для применения линейной регрессии, нередко нарушаются. Линейная регрессия может дать оценку вероятности вне диапазона [0,1], что является неприемлемым. К примеру, логистическая регрессия лишена этого недостатка.

Логистическая регрессия и пробит-регрессия больше подходят для построения скоринговой модели, так как допускают категорийное представление данных. Модель логистической регрессии задаётся следующим образом:

$$\log \frac{p}{1-p} = B^T x = B_0 + B_1 x_1 + ... + B_k x_1 + ... + B_k X_k, \quad (4)$$

где p – оценка вероятности того, что клиент «плохой»;

B – вектор неизвестных параметров регрессии, который вычисляется через условие максимизации отношения правдоподобия.

Модель логистической регрессии базируется на функции логарифм. В свою очередь, пробит-регрессия базируется на нормальном распределении и задаётся следующим образом:

$$N(p) = B^T x = B_0 + B_1 x_1 + ... + B_k x_k . \quad (5)$$

Так как логистическая регрессия и пробит-регрессия используют схожие по форме распределения, результаты применения данных моделей также схожи. Логистическая регрессия пользуется большим предпочтением, так как вычисления проще, чем в пробит-регрессии и имеется больше инструментов для работы с ней. За счёт своей бинарной природы, логистическая регрессия предпочтительней линейной регрессии в использовании для построения скоринговых моделей. На практике же было выяснено, что разница в точности предсказываемых результатов незначительна. Тем не менее, наблюдается преобладание логистической регрессии в скоринговых системах.

Искусственные нейронные сети являются симуляцией нейронных сетей имеющихся в природе. Возникло это понятие при попытке смоделировать процессы, происходящие в мозге человека .

Нейронные сети, также называемые многослойным перцептроном, особенно подходят для решения задачи классификации. Они широко используются в различных сферах: финансах, компьютерных науках, физике и медицине. Популярность нейронных сетей отчасти обуславливается возможностью моделировать сложные ситуации без особых затрат со стороны использующего этот метод. По своей природе нейронные сети автоматически обнаруживают любую нелинейную ситуацию в данных и подстраиваются под неё. Также многослойные нейронные сети являются универсальными аппроксиматорами, то есть могут аппроксимировать любую функцию сколь угодно точно. [5]

Нейронные сети состоят из слоев которые, в свою очередь, состоят из узлов. Есть 3 типа слоёв в сетях: входной, скрытые, выходной. Входной слой образуют атрибуты клиента, такие как пол, возраст и т.п.

Выход y_k для *k*-го узла с *m* входами представляется так:[9]

$$y_k = \phi(V_k) = \phi(\sum_{j-0}^{m} \omega_i x_i) = \phi(w^T x), \quad (6)$$

где φ – активационная функция;

x – вектор входных данных;

ω – весовой вектор который обозначает силу связи между узлами.

Основным недостатком является то, что несмотря на возможность добиться высокой точности прогноза, понять причины, по которым было принято то или иное решение, невозможно.

В контексте кредитного скоринга было показано, что нейронные сети работают не хуже традиционных методов.

Данный метод отлично подходит для нахождения связей между данными, особенно если связи нелинейные. Он применяется для построения деревьев принятия решений, и имеет много общего с классическими методами, такими как дискриминантный анализ и линейная регрессия. Аббревиатура CHAID расшифровывается как Chi-squared Automated Interaction Detector.

Гибкость данного метода делает его привлекательным для использования, но это не означает, что его стоит использовать вместо традиционных методов. В случае, когда встречаются строгие теоретические предположения о распределении, традиционные методы предпочтительней. Как техника исследования или в случае, когда традиционные методы не срабатывают, CHAID анализ является непревзойдённым инструментом.

CHAID строит не бинарные деревья (т.е. деревья у которых может быть более двух ветвей) на основе относительно простого алгоритма, который особенно хорошо подходит для анализа больших массивов данных. Алгоритм основывается на применении теста Х-квадрат. [9]

Дерево принятия решений как метод разделяет данные на подмножества, каждое из которых более однородно в своем поведении, нежели исходное множество данных. Каждое из этих подмножеств делится далее, по такому же алгоритму. Результат деления именуется «листом» это дерева. Имеются и другие методы, работающие по схожему принципу.

Достоинства этого метода – простота и интуитивность. Метод способен работать с отсутствующими наблюдениями. Особенно он применим в случае, когда о данных до их исследования практически ничего неизвестно и нельзя построить какие-либо догадки или гипотезы.

Главный недостаток этого метода – сложность компьютерных расчетов. Вследствие громоздкости получаемых деревьев, процесс изучения модели трудоёмкий. Изменения в ситуации может привести к пересмотру всего дерева решений. В основном метод используется как вспомогательный. К примеру для определения переменных, которые наиболее сильно объясняют поведение зависимой переменной. Метод k ближайших соседей. Непараметрический метод классификации объектов. Основывается на метрике, определяющей схожесть между данными.

Первоначально вводятся тренировочные данные, разделенные на классы. Затем вводятся оцениваемые данные и определяется схожесть между введёнными и тренировочными данными. На основе метрики выбирается *k* ближайших соседей. Новый элемент относят к тому классу, к которому принадлежит большинство его соседей [10].

Количество соседей *k* определяется компромиссом между компенсацией и дисперсией. Чем меньше класс, тем меньше выбирается *k*. При этом необязательно, что при больших *k* результат будет лучше. Одно из преимуществ данного метода – легко добавить новые данные, не изменяя при этом модель. Непараметрическая сущность этого метода позволяет работать с иррациональностями в функциях риска на пространстве признаков. Отсутствие формального метода для выбора *k* и невозможность вероятностной интерпретации результата, так как результатом являются ожидаемые частоты, являются главными недостатками метода. Данные сложности могут быть решены использованием методом Байесовской аппроксимации. Данный метод мало используется в скоринге. Одной из причин этого является то, что для классификации одного объекта необходимо иметь базу по всем объектам.

Более новый метод опорных векторов, построенный на машинном обучении, показал себя не хуже традиционных скоринговых методов. Он состоит из двух процессов: первый преобразует входные данные к данным высокой размерности в пространстве признаков; второй классифицирует данные с помощью линейного классификатора. Классификатором может выступать, например, линейный дискриминантный анализ [10].

Был проведено сравнительно исследование для скоринговых методов. Критериями для ранжирования служили процент ошибок при классификации и ROC-кривая. Результаты исследования приведены в таблице 1.

Таблица 1 – Сравнение методов анализа скоринга

Метод	Средний процент ошибок
Нейронные сети	3.2
Опорные вектора	3.3
Логистическая регрессия	3.5
Линейный дискриминантный анализ	5.3
Линейные LS-SVM	5.5
Расширенное дерево Байеса	5.6
Наивный байесовский классификатор	7.8
Радикально базисные функции	9.1
k-ближайших соседей (k=100)	9.5
Линейный SVM	10.8
Квадратично дискриминантный анализ	11.9
Дерево принятия решений	14.1
k-ближайших соседей (k=10)	14.7

Из таблицы 1 видно, что нейронные сети, метод опорных векторов и логистическая регрессия являются наилучшими из представленных методов. Кроме того традиционные методы, такие как линейный и дискриминантный анализ показали себя конкурентоспособными. Отсюда следует, что, вероятно, большинство данных для кредитного скоринга лишь немного нелинейны. Вследствие чего линейные методы показали себя на уровне с нелинейными.

Не существует оптимальной скоринговой модели для любой ситуации. Выбор модели зависит от данных и цели, на которую направленно создание

модели. Кроме того, метод, оценивающий наилучшим образом, не обязательно будет лучшим в данной ситуации. [11]

Очень часто для реализации математического аппарата кредитного скоринга используют Байесовские сети и Байесовские классификаторы. Байесовские сети позволяют представлять многомерные распределения $p = (x, G)$ и $p = (x, B)$ в виде комбинации нескольких распределений более низкой размерности.

При этом в ходе построения модели могут быть учтены причинно-следственные связи (благодаря чему они широко распространены в таких областях как медицинская диагностика, поиск технических неисправностей и т. п.). Хотя в кредитном скоринге и других задачах классификации как правило рассматриваются переменные, которые не связаны друг с другом причинно, однако можно предположить существование скрытых переменных, обуславливающих ту или иную условную зависимость или независимость. Безусловным преимуществом байесовских сетей является возможность вывода по неполным данным. Если информация о потенциальном заёмщике не является полной, то алгоритм вывода по байесовской сети вычислит вероятность дефолта, основываясь лишь на доступных данных (что эквивалентно усреднению по переменным, чьи значения неизвестны). Байесовский подход позволяет также относительно легко производить интеграцию системы апостериорного скоринга с другими используемыми моделями, в частности, с системой априорного скоринга, чтобы учесть изменения в распределении дефолтов, связанные с изменением экономической ситуации.

Байесовские сети позволяют записать многомерное совместное распределение в виде комбинации нескольких распределений меньшей размерности. Байесовской сетью называется так же пара $N = (G, \theta)$, где G ориентированный ацикличный граф (ОАГ), а θ – набор условных распределений. Каждая вершина графа соответствует одной из переменных $X_1, ..., X$. Для каждой вершины задано условное распределение

$\theta_{X_i}\Pi_{X_i} = P(X_i, \Pi_{X_i})$ где Π_{X_i} – множество непосредственных предшественников X_i в графе G. Байесовская сеть N определяет следующее совместное распределение: $P(X_1,..., X_n) = \prod_{i=1}^{n} P(X_i, \Pi_i)$ [12].

Если известны реализовавшиеся значение одной или нескольких переменных, то известные алгоритмы вывода по $X_i = x_i (i \in I)$ байесовской сети позволяют оценить условные вероятности $P(X_1,..., X_n)$ остальных переменных.

В общем случае определение оптимальной структуры сети и вычисление маргинальных вероятностей по имеющимся наблюдениям являются NP-трудными задачами, поэтому используют приближённые алгоритмы обучения и вывода. Для построения классификатора, основанной на байесовской сети, необходимо вначале найти адекватную структуру графа G, а затем оценить условные распределения θ. Последняя задача имеет довольно простое решение (при условии отсутствия пропусков в данных), основную же трудность представляет именно поиск оптимальной структуры.

Когда байесовские сети применяются к задаче классификации, граф G условно разделяется на две части: вершина C, соответствующая классу клиента, и все остальные вершины. При этом выделяются следующие основные структуры классифицирующей сети.

2.3 Принципы построения логистической регрессии в DEDUCTOR

Deductor – это аналитическая платформа, основа для создания законченных прикладных решений в области анализа данных. Реализованные в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: от консолидации данных до построения моделей и визуализации полученных результатов.

До появления аналитических платформ анализ данных осуществлялся в основном в статистических пакетах. Их применение требовало от пользователя высокой квалификации.

Большинство алгоритмов, реализованных в статистических пакетах, не позволяло эффективно обрабатывать большие объемы информации. Для

автоматизации рутинных операций приходилось использовать встроенные языки программирования. [13]

В конце 80-х гг. произошел стремительный рост объемов информации, накапливаемой на машинных носителях, и увеличились потребности бизнеса в применении анализа данных. Ответом стало появление новых парадигм в анализе, таких как хранилища данных, машинное обучение, Data Mining, Knowledge Discovery in Databases.

Это позволило популяризировать анализ данных, поставить его на промышленную основу и решить огромное число бизнес-задач с большим экономическим эффектом. Финалом развития анализа данных стали специализированные программные системы – аналитические платформы, которые полностью автоматизировали все этапы анализа от консолидации данных до эксплуатации моделей и интерпретации результатов.

Первая версия Deductor была выпущена в 2000 г., и с тех пор идет непрерывное развитие данной платформы. Разработчик – компания BaseGroup Labs (Россия). Deductor – яркий представитель как настольной, так и корпоративной системы анализа данных последнего поколения.

В скоринговой карте наряду с классификацией объектов требуется ещё оценивать степень их принадлежности тому или иному классу или «степень уверенности» классификации [13].

Это позволяет делать логистическая регрессия – распространенный статистический инструмент для решения задач регрессии и классификации. Иными словами, с помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретной испытуемой выборки.

Логистическая регрессия – это разновидность множественной регрессии, общее назначение которой состоит в анализе линейной связи между несколькими независимыми переменными и зависимой переменной.

Когда предсказываемых классов два, то говорят о бинарной логистической регрессии. В традиционной множественной линейной регрессии существует следующая проблема: алгоритм не «знает», что переменная отклика

бинарна по своей природе. Это неизбежно приведет к модели с предсказываемыми значениями большими 1 и меньшими 0.

Но такие значения вообще не допустимы для первоначальной задачи. Таким образом, множественная регрессия просто игнорирует ограничения на диапазон значений для *y* [14].

Для решения проблемы задача регрессии может быть сформулирована иначе: вместо предсказания бинарной переменной мы предсказываем непрерывную переменную со значениями на отрезке [0,1] при любых значениях независимых переменных. Это достигается применением логит-преобразования вида: $P = \frac{1}{1+e^{y}}$, где P – вероятность того, что произойдет интересующее событие; e=2,71… – основание натуральных логарифмов; y – стандартное уравнение регрессии [14]:

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n . \quad (7)$$

Существует несколько способов нахождения коэффициентов логистической регрессии. На практике часто используют метод максимального правдоподобия. Он применяется в статистике для получения оценок параметров генеральной совокупности по данным выборки. Основу метода составляет функция правдоподобия (likehood function), выражающая плотность вероятности (вероятность) совместного появления результатов выборки.

Для поиска максимума, как правило, используется оптимизационный метод Ньютона, для которого здесь всегда выполняется условие сходимости.

Для облегчения вычислительных процедур максимизируют не саму функцию правдоподобия, а ее логарифм. В результатах обычно выводят численное значение либо на каждом шаге алгоритма, либо на последнем шаге.

Бинарная логистическая регрессия эквивалента построению рейтинговой или балльной модели, т.к. если признак f_i наблюдается в объектах, то к сумме баллов добавляется вес a_j. Классификация производится путём сравнения набранной суммы баллов с пороговым значением.

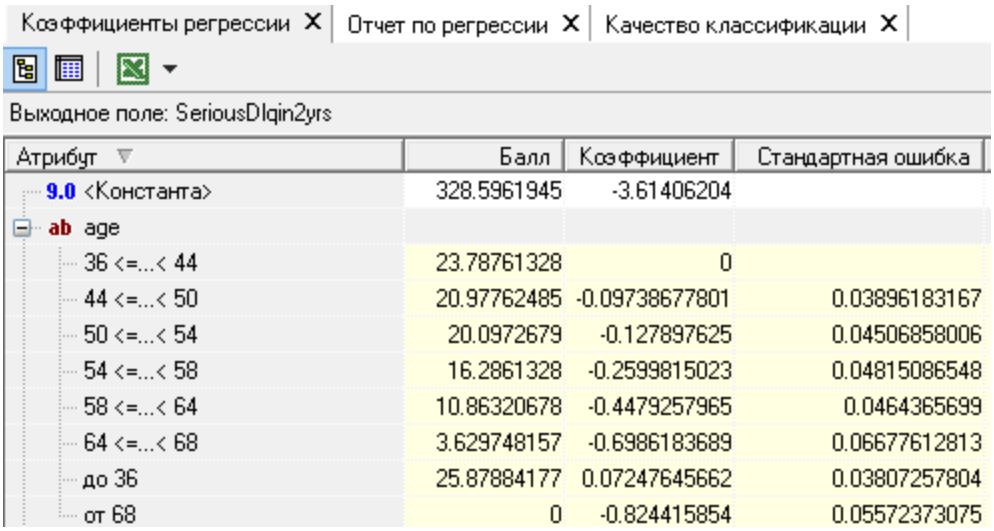
Для построения модели логистической регрессии готовится обучающая выборка так же, как и для нейросети. Но выходное поле может быть только дискретного типа и бинарное (т.е. количество уникальных значений по нему должно быть равно двум).

На этапе определения входов модели необходимо помнить, что естественное стремление учесть как можно больше потенциально полезной информации приводит к включению избыточных шумовых признаков. Экспериментально установлено, что для успешного обучения число примеров должно в несколько раз (примерно в 5) превосходить число входных признаков.

Но даже если все признаки информативны, количества обучающих примеров может просто не хватить для надёжного определения коэффициентов регрессии при всех признаках [15].

Когда данных мало, приходится искусственно упрощать структуру регрессионной модели, оставляя наиболее существенные признаки.

На рисунке 1, который представляет собой скриншот с оригинального исследования, представлена рассчитанная визуализация скоринговой карты с бальной системой. По признаку «возраст» осуществлена разбивка клиентов выборки по категориальным возрастным группам от 36 до 68 лет и начисляемые баллы по каждой из категорий. Для удобства демонстрации данные параметры представлены в таблице 2.


Коэффициенты регрессии | Отчет по регрессии | Качество классификации

Выходное поле: SeriousDlqin2yrs

Атрибут	Балл	Коэффициент	Стандартная ошибка
9.0 <Константа>	328.5961945	-3.61406204	
ab age			
36 <=...< 44	23.78761328	0	
44 <=...< 50	20.97762485	-0.09738677801	0.03896183167
50 <=...< 54	20.0972679	-0.127897625	0.04506858006
54 <=...< 58	16.2861328	-0.2599815023	0.04815086548
58 <=...< 64	10.86320678	-0.4479257965	0.0464365699
64 <=...< 68	3.629748157	-0.6986183689	0.06677612813
до 36	25.87884177	0.07247645662	0.03807257804
от 68	0	-0.824415854	0.05572373075

Рисунок 1 – Визуализация коэффициентов регрессии

Таблица 2 – Бальная характеристика категории «возраст»

Возрастная категория (года)	Скоринговый балл	Вес коэффициента	Стандартная ошибка
до 36	26	0,07	0,03
от 36 до 44	24	0	Пусто
от 44 до 50	21	-0,09	0,03
от 50 до 54	21	-0,01	0,04
от 54 до 58	16	-0,25	0,04
от 58 до 64	11	-0,44	0,04
от 64 до 68	4	-0,69	0,06
от 68	0	-0,82	0,05
Итого	123	-0,28	0,04

С помощью алгоритма можно подобрать оптимальные пороги диагностических показателей, оценить чувствительность и специфичность модели, рассчитать ложноположительные и ложноотрицательные результаты. Это позволяет сделать тесты более эффективными в сравнении с традиционными методиками. Алгоритм оценивает зависимость состояния банка от показателей финансовой устойчивости, ликвидности, рентабельности, деловой активности. Это позволяет предупреждать возможности возникновения кризисной ситуации, сохранить устойчивое финансовое состояние и повысить эффективность банковской деятельности. [15]

Для оценки качества или «адекватности» логистической регрессии в качестве модели, используют ROC-анализ или ROC-кривую. ROC-кривая (Receiver Operator Characteristic) – кривая, которая наиболее часто используется для представления результатов бинарной классификации в кредитном скоринге. Название пришло из систем обработки сигналов. Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами. ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. При этом предполагается, что у классификатора имеется некоторый параметр, варьируя который, мы будем получать то или иное

разбиение на два класса [7]. Этот параметр часто называют порогом, или точкой отсечения (cut-off value). В зависимости от него будут получаться различные величины ошибок I и II рода. В логистической регрессии порог отсечения изменяется от 0 до 1 – это и есть расчетное значение уравнения регрессии. Будем называть его рейтингом. Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности (confusion matrix), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам – таблица 3.

Таблица 3 – Обозначение результатов выхода данных по шкале отрицательности

Модель	Положительно	Отрицательно
Положительно	TP	FP
Отрицательно	FN	TP

Далее представлена расшифровка сокращений:

TP (True Positives) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);

TN (True Negatives) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);

FN (False Negatives) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый «ложный пропуск» – когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи. В случае кредитного скоринга важен ответ на вопрос вернёт ли клиент деньги или нет в кредитную организацию (банк). При анализе

чаще оперируют не абсолютными показателями, а относительными – долями (rates), выраженными в процентах.

Доля истинно положительных примеров (True Positives Rate)[16]:

$$TPR = \frac{TP}{TP + FN} \times 100\%. \quad (8)$$

Доля ложно положительных примеров (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} \times 100\%. \quad (9)$$

Для оценки «адекватности» логистической регрессии в скоринге также используют два параметра: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

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

$$S_e = TPR = \frac{TP}{TP + FN} \times 100\%. \quad (10)$$

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

$$S_p = \frac{TN}{TN + FP} \times 100\%. \quad (11)$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

Если рассуждать в терминах скоринга – задачи определения «дефолтности» клиента, где модель классификации заёмщиков на дефолтных и недефолтнымх называется диагностическим тестом, то получится следующее:

Чувствительный диагностический тест проявляется в скоринге – максимальном предотвращении пропуска дефолтных клиентов [16];

Специфичный диагностический тест диагностирует только доподлинно дефолтных. Это важно в случае, когда, например, выявление потенциального

«мошенника», чьи действия могут быть связаны с серьезными последствиями для банка и «гипервыявление» дефолтников не желательно.

Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом *dx* (например, 0.01) рассчитываются значения чувствительности S_e и специфичности S_p. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.

Строится график зависимости: по оси *Y* откладывается чувствительность S_e, по оси *X* – 100%–S_p (сто процентов минус специфичность), или, что то же самое, *FPR* – доля ложно положительных случаев.

Канонический алгоритм построения ROC-кривой имеет вид:

Входы: *L* – множество примеров; *f*[*i*] – рейтинг, полученный моделью, или вероятность того, что *i*-й пример имеет положительный исход; *min* и *max* – минимальное и максимальное значения, возвращаемые *f*; *dx* – шаг; *P* и *N* – количество положительных и отрицательных примеров соответственно. [16]

1. *t=min*
2. *повторять*
3. *FP=TP=0*
4. *для всех примеров* i *принадлежит L {*
5. *если f[i] >= t тогда // этот пример находится за порогом*
6. *если i положительный пример тогда*
7. *{ TP=TP+1 }*
8. *иначе // это отрицательный пример*
9. *{ FP=FP+1 }*
10. *}*
11. *Se=TP/P*100*
12. *100_m_Sp=FP/N // расчет (100 минус Sp*
13. *Добавить точку (100_m_Sp, Se) в ROC кривую*
14. *t=t+dx*
15. *пока (t>max)*

В ходе данного исследования было построено несколько логистических моделей, одна из них представлена на рисунке 2.

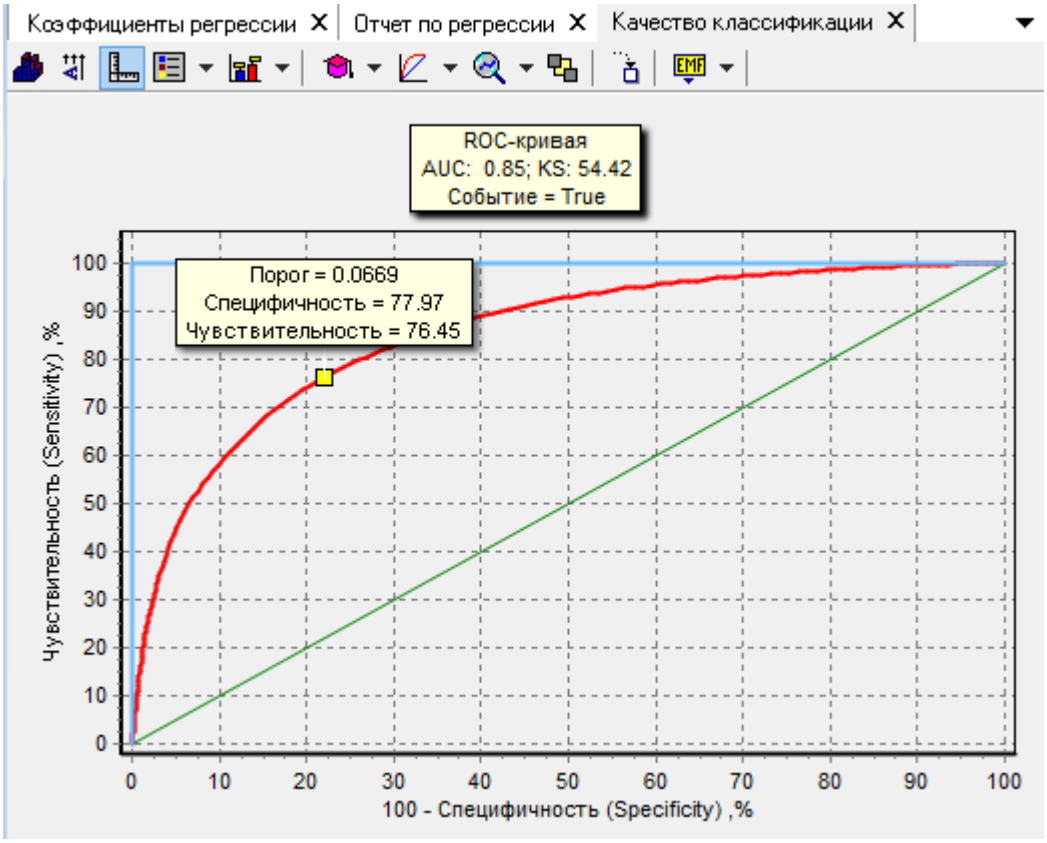


Рисунок 2 – Модель логистической регрессии и ROC-кривой

Для идеального классификатора график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100% или 1.0 (идеальная чувствительность), а доля ложно положительных примеров равна нулю. Поэтому чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагональной прямой, тем менее эффективна модель. Диагональная линия соответствует «бесполезному» классификатору, т.е. полной неразличимости двух классов.

3 МАТЕМАТИЧЕСКАЯ АПРОБАЦИЯ СКОРИНГОВЫХ ДАННЫХ

В данной главе происходит построение и исследование логистической регрессии и с помощью неё строится скоринговая модель. Построение будем вести исходя из статистики по потребительским кредитам 150 000 клиентов банка.

Постановка задачи.

Предположим, имеется некий банк, занимающийся кредитованием частных лиц. В банк за получением кредита обращаются клиенты. Решение о выдаче кредита банк выносит на основе информации о клиенте.

Информацию о клиенте банк получает из разных источников: от самого клиента, от кредитного бюро и из других источников. В данном исследовании будет рассматриваться информация, предоставляемая самим клиентом. Банк получает её через заполненную заемщиком анкету. В анкете заемщик указывает следующие данные: пол, возраст, семейное положение, наличие детей, ежемесячный доход, наличие недвижимости и прочее.

На основании этих данных разобьём клиентов на группы, в которых они схожи по определённым признакам. Для каждого клиента найдём рейтинг – эмпирическая вероятность того, что клиент вернёт кредит при условии, что он принадлежит данной группе. Найдя распределение рейтингов, мы тем самым построим скоринговую модель. [17]

Для применения метода необходимо, чтобы данные удовлетворяли следующим условиям:

1) независимость – клиенты не имеют сговора по выплате кредита;

2) однородность – данные взяты из одной генеральной совокупности;

3) разновероятность – клиенты равновероятно распределяются по группам.

3.1 Реализация задачи методом Байеса

Введем вероятностное пространство (Ω, F, P) Обозначим в этом пространстве $\omega \in \Omega$ – клиент банка.

Каждый клиент банка имеет набор характеристик согласно заполненной анкете. Например: в браке или нет, уровень дохода, разбитый по категориям, наличие машины и прочие характеристики. Согласно этим характеристикам введём разбиение пространства Ω на множества B:

1) $\bigcup_{i=1}^{n} B_i = \Omega$;

2) $P(B_i) > 0, \forall_i$.

Далее для упрощения задачи применим эмпирический Байесовский подход, который основывается на формуле Байеса[17]:

$$P(\frac{A}{B}) = \frac{P(\frac{B}{A})P(A)}{P(B)} \tag{12}$$

где $P(\frac{A}{B})$ – эмпирическая вероятность события A при условии B, которую называют апостериорной вероятностью;

$P(\frac{B}{A})$ – эмпирическая вероятность события B при условии A;

$P(A)$ – априорная эмпирическая вероятность события A;

$P(B)$ – эмпирическая вероятность события В.

Данная формула позволяет переоценить вероятность события A, учитывая тот факт, что произошло событие B.

Из определения условной вероятности можем записать [18]:

$$P(X_i, Y_i) = P(X_i \frac{X_i}{Y_i}) P(Y_i). \tag{13}$$

Далее заметим, что:

$$P(Y_I) = \sum_{i=1}^{2} P(X_{ij}, Y_j). \tag{14}$$

Выразив из (13) $P(X_i \frac{X_i}{Y_i})$ и подставив в (14) формулу выражение для $P(Y_i)$ получим:

$$P(X_i \frac{X_i}{Y_i}) = \frac{P(X_i, Y_i)}{\sum_{i=1}^{2} P(X_{ij}, Y_j)}. \quad (15)$$

Выраженная величина является рейтингом клиента из *j*-го множества. Исходя из её значений, осуществляется классификация на «хороших» и «плохих».

Для проверки гипотезы о независимости в данном исследовании используется ранг критерия Спирмена. $X_1, X_2, ..., X_n$ Статистикой данного критерия является коэффициент ранговой корреляции, определяемый следующим образом.

Даны два ряда наблюдений: $X_1, X_2, ..., X_n$ и $Y_1, Y_2, ..., Y_n$. На основании этих наблюдений построим пары рангов $(R_i, S_i), i = 1$. Под рангом R_i понимаем номер места, занимаемого наблюдением X_i в вариационном ряду $X_1 \le X_2, ..., X_n$. Аналогично $X_1 \le X_2, ..., X_n$ понимаем ранг S_i. Затем, переставляем пары рангов в порядке возрастания первой компоненты. Получившийся ряд обозначим $(1, T_1), (2, T_2), ..., (n, T_n)$ [19].

Коэффициент корреляции находится по формуле:

$$\rho = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (T_i - i)^2. \quad (16)$$

Критическая область критерия:

$$\tau_{1\alpha} = \{|\rho| > t_\alpha(n)\} \quad (17)$$

Для нахождения области критерия воспользуемся тем, что закон распределения $\sqrt{n}\rho$ стремится к $N(0,1)$, при больших n. Отсюда:

$$t_\alpha(n) = \frac{t_{\alpha/2}}{\sqrt{n}}, \Phi(-t_{\alpha/2}) = \alpha / 2, \quad (18)$$

где $\Phi(x)$ – функция распределения стандартного Гауссова закона.

При уровне значимости $\alpha = 0.05$, $t_{0.05}$=1.959964. Граница критической зоны $t_{0.05}(n) = 0.062354$. Найденный коэффициент корреляции $\rho = -0.00942$.

Таким образом, статистика критерия не попадает в его критическую область, и мы можем принять гипотезу о независимости при уровне значимости 0.05. Формулируется гипотеза следующим образом. Даны две выборки $X = (X_1,...,X_n)$ и $Y = (Y_1,...,Y_m)$ из распределений $T(\eta)$ и $T(\theta)$ соответственно, с функциями распределений $F_1(x)$ и $F_2(x)$. Тогда гипотеза об однородности $H_0 : F_1(x) = F_2(x)$ [19].

Для проверки данной гипотезы воспользуемся критерием Смирнова.

Статистикой данного критерия является:

$$D_{n,m} = \sup_{-\infty<x<\infty} \left|\hat{F}_{1n}(x) - \hat{F}_{2m}(x)\right|, \tag{19}$$

где $\hat{F}_{1n}, \hat{F}_{2m}$ – эмпирические функции распределения, построенные по выборкам *X* и *Y*. Критическая область задаётся в виде $\tau_{1\alpha} = \{D_{n,m} > t_\alpha(n,m)\}$. При больших *n* и *m* границу критической области $t_\alpha(n,m)$ можно принять равной $\sqrt{\frac{1}{n}+\frac{1}{m}}\lambda_\alpha$, где $K(\lambda_\alpha) = 1-\alpha$ – функция распределения Колмогорова:

$$K(t) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2t^2}. \tag{20}$$

Статистика критерия не попадает в критическую область, и мы можем принять гипотезу об однородности при уровне значимости 0.05. Сформулируем гипотезу. Дана выборка $X = (X_1,...,X_n)$ из распределения $T(\xi)$ с функцией распределения $F_\xi(x)$, которая неизвестна. Необходимо проверить, что $F_\xi(x) = F(x)$ – функция распределения равномерного распределения на отрезке [0;0.05].[20]

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{x}{x-0.05}, & 0 \le x \le 0.05 \\ 1, & x \ge 0.05 \end{cases} \tag{21}$$

Для этого воспользуемся критерием согласия Пирсона.

Статистикой критерия является [12]:

$$X^2 = n\sum_{i=1}^{k}\frac{(n_i / n - P_i)^2}{P_i}, \quad (22)$$

где n_i / n – частота попадания наблюдений в *i*-й отрезок, P_i – вероятность попадания в *i*-й отрезок. Если проверяемая гипотеза верна, при больших n статистика подчиняется распределению хи-квадрат с *k*-1 степенью свободы.

Значение статистики не превышает критического уровня и гипотеза о равномерном распределении клиентов по группам принимается при уровне значимости 0.05. Итак, данные удовлетворяют всем гипотезам, приведённым выше, и мы можем приступить к нахождению рейтингов.

Имеющиеся данные содержат множество различных характеристик клиентов. Для построения будем использовать 4 из них, наиболее значимых. Для обработки данных возьмём основное количество клиентов из выборки (150 000 элементов) и разделим его на 60 групп.

Выбранные характеристики: возраст и пол заёмщика, наличие детей, выплаты по кредиту в % от суммарного дохода заёмщика. Характеристика возраст принимает 3 значения – от 18 до 29, от 30 до 45, 46 и более …; пол заёмщика два значения – мужской и женский; наличие детей два значения – есть дети, и нет детей; выплаты принимают 5 значений – (<5%),(от 6 до 10%), (от 11 до 16%), (от 17 до 22%),(от 23 до 55%).

Каждое конкретное значение характеристики назовём свойством заёмщика. Разобьём всех наших клиентов на множества, опираясь на наличие конкретного свойства у данного клиента. К примеру, B_1^1 – множество клиентов у которых нет детей.[21]

Приведём эти обозначения:

$\vec{B}_1$ – дети, B_1^1= Нет детей, B_1^2= Есть дети;

$\vec{B}_2$ – возраст, B_2^1=Возр1(18-29), B_2^2=Возр2(30-45), B_2^3=Возр3(46-…);

$\vec{B}_3$ – пол, B_3^1=Женский, B_3^2=Мужской;

$\vec{B}_4$ – выплаты по кредиту в % от суммарного дохода заемщика, B_4^1=Вып1(<5), B_4^2=Вып2(6-10), B_4^3=Вып3(11-16), B_4^4=Вып4(17-22), B_4^5=Вып5(23-55).

Образуем новые множества, как комбинацию всех возможных свойств клиента – $B_i = B_1^J \bigcap B_2^k \bigcap B_3^j \bigcap B_4^m$ по всевозможным *j*, *k*, *l*,*m* к примеру, множество B_i состоит из женщин в возрасте от 18 до 29 лет без детей, выплачивающих <5% от своего суммарного дохода. Количество таких множеств равно 60. Рассматриваемые данные кодировки представлены в таблице 4 [21].

Таблица 4 – Кодировка множеств

	Возр 1	Возр 2	Возр 3	Жен-ский	Муж-ской	Нет детей	Есть де-ти	Вып 1	Вып 2	Вып 3	Вып 4	Вып 5
1	+			+		+		+				
2	+			+		+			+			
3	+			+		+				+		
4	+			+		+					+	
5	+			+		+						+
6	+				+	+		+				
7	+				+	+			+			
8	+				+	+				+		
9	+				+	+					+	
10	+				+	+						+
11		+		+		+		+				
12		+		+		+			+			
13		+		+		+				+		
14		+		+		+					+	
15		+		+		+						+
16		+			+	+		+				
17		+			+	+			+			
18		+			+	+				+		
19		+			+	+					+	
20		+			+	+						+
21			+	+		+		+				
22			+	+		+			+			
23			+	+		+				+		
24			+	+		+					+	

Продолжение 4

25			+	+		+						+
26			+		+	+		+				
27			+		+	+			+			
28			+		+	+				+		
29			+		+	+					+	
30			+		+	+						+
31	+			+			+	+				
32	+			+			+		+			
33	+			+			+			+		
34	+			+			+				+	
35	+			+			+					+
36	+				+		+	+				
37	+				+		+		+			
38	+				+		+			+		
39	+				+		+				+	
40	+				+		+					+
41		+		+			+	+				
42		+		+			+		+			
43		+		+			+			+		
44		+		+			+				+	
45		+		+			+					+
46		+			+		+	+				
47		+			+		+		+			
48		+			+		+			+		
49		+			+		+				+	
50		+			+		+					+
51			+	+			+	+				
52			+	+			+		+			
53			+	+			+			+		
54			+	+			+				+	
55			+	+			+					+
56			+		+		+	+				
57			+		+		+		+			
58			+		+		+			+		
59			+		+		+				+	
60			+		+		+					+

Построим совместное эмпирическое распределение двух дискретных случайных величин – $X_{i,j}$ и Y_j, где $X_{i,j}$={0,1}, Y_j={ B_i, i=1:60}. Строить его будем как отношение количества клиентов удовлетворяющих паре значений случайных величин (*X*, *Y*) к общему количеству клиентов. Зафиксируем количество

клиентов соответствующих каждой возможной паре (*X*, *Y*) и представим результат в таблице 5.

Таблица 5 – Количество клиентов вернувших и не вернувших кредит в каждой группе. 0 – вернули кредит, 1 – не вернули кредит.

X Y	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	20	45	44	21	15	12	23	31	15	4	18	28	35	14	12	19	34	45	21	11
1	12	22	20	10	10	3	17	17	8	3	12	23	13	5	6	5	25	15	4	3
X Y	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
0	18	36	47	19	5	20	49	52	30	28	10	21	20	16	8	11	27	27	8	6
1	6	11	11	6	3	5	10	16	4	10	5	17	13	3	4	6	17	10	4	5
X Y	41	42	43	44	45	46	47	48	4	50	51	52	53	54	55	56	57	58	59	60
0	49	65	65	28	14	49	69	58	2	20	5	12	7	3	2	2	6	6	4	1
1	12	32	15	9	2	14	27	18	9	3	2	3	3	2	2	1	2	4	0	1

Построим совместное эмпирическое распределение вероятностей. Для этого разделим количество клиентов вернувших и не вернувших кредит в каждой из групп на общее количество клиентов. Данные распределения отражены в таблице 6.

Таблица 6 – Эмпирическое распределение вероятностей

Номер	0	1	Номер	0	1
1	0,010116	0,00607	31	0,005058	0,002529
2	0,022762	0,011128	32	0,010622	0,008599
3	0,022256	0,010116	33	0,010116	0,006576
4	0,010622	0,005058	34	0,008093	0,001517
5	0,007587	0,005058	35	0,004047	0,002023
6	0,00607	0,001517	36	0,005564	0,003035
7	0,011634	0,008599	37	0,013657	0,008599
8	0,01568	0,008599	38	0,013657	0,005058
9	0,007587	0,004047	39	0,004047	0,002023
10	0,002023	0,001517	40	0,003035	0,002529
11	0,009105	0,00607	41	0,024785	0,00607
12	0,014163	0,011634	42	0,032878	0,016186
13	0,017704	0,006576	43	0,032878	0,007587
14	0,007081	0,002529	44	0,014163	0,004552
15	0,00607	0,003035	45	0,007081	0,001012
16	0,009611	0,002529	46	0,024785	0,007081
17	0,017198	0,012645	47	0,034901	0,013657
18	0,022762	0,007587	48	0,029337	0,009105
19	0,010622	0,002023	49	0,013657	0,004552
20	0,005564	0,001517	50	0,010116	0,001517
21	0,009105	0,003035	51	0,002529	0,001012
22	0,018209	0,005564	52	0,00607	0,001517

Продолжение таблицы 6

23	0,023773	0,005564	53	0,003541	0,001517
24	0,009611	0,003035	54	0,001517	0,001012
25	0,002529	0,001517	55	0,001012	0,001012
26	0,010116	0,002529	56	0,001012	0,000506
27	0,024785	0,005058	57	0,003035	0,001012
28	0,026302	0,008093	58	0,003035	0,002023
29	0,015175	0,002023	59	0,002023	0
30	0,014163	0,005058	60	0,000506	0,000506

Найдём *P* – эмпирическое распределение вероятностей попадания в каждую из групп. Для этого разделим количество клиентов в каждой группе на общее количество клиентов. Общее количество клиентов – 150 000. Данные представлены в таблице 7.

Таблица 7 – Эмпирическое распределение вероятностей попадания в каждую группу

№	P	№	P	№	P	№	P	№	P
1	0,016186	13	0,024279	25	0,004047	37	0,022256	49	0,018209
2	0,03389	14	0,009611	26	0,012645	38	0,018715	50	0,011634
3	0,032372	15	0,009105	27	0,029843	39	0,00607	51	0,003541
4	0,01568	16	0,01214	28	0,034396	40	0,005564	52	0,007587
5	0,012645	17	0,029843	29	0,017198	41	0,030855	53	0,005058
6	0,007587	18	0,030349	30	0,019221	42	0,049064	54	0,002529
7	0,020233	19	0,012645	31	0,007587	43	0,040465	55	0,002023
8	0,024279	20	0,007081	32	0,019221	44	0,018715	56	0,001517
9	0,011634	21	0,01214	33	0,016692	45	0,008093	57	0,004047
10	0,003541	22	0,023773	34	0,009611	46	0,031866	58	0,005058
11	0,015175	23	0,029337	35	0,00607	47	0,048558	59	0,002023
12	0,025797	24	0,012645	36	0,008599	48	0,038442	60	0,001012

Найдём рейтинги клиентов как эмпирическую вероятность того, что клиент вернёт кредит при условии, что он принадлежит какой-либо группе. Данные отображены в таблице 8.

Таблица 8 – Рейтинги клиентов

№	Рейтинг	№	Рейтинг	№	Рейтинг
1	0,625	21	0,75	41	0,803279
2	0,671642	22	0,765957	42	0,670103
3	0,6875	23	0,810345	43	0,8125
4	0,677419	24	0,76	44	0,756757
5	0,6	25	0,625	45	0,875

Продолжение таблицы 8

6	0,8	26	0,8	46	0,777778
7	0,575	27	0,830508	47	0,71875
8	0,645833	28	0,764706	48	0,763158
9	0,652174	29	0,882353	49	0,75
10	0,571429	30	0,736842	50	0,869565
11	0,6	31	0,666667	51	0,714286
12	0,54902	32	0,552632	52	0,8
13	0,729167	33	0,606061	53	0,7
14	0,736842	34	0,842105	54	0,6
15	0,666667	35	0,666667	55	0,5
16	0,791667	36	0,647059	56	0,666667
17	0,576271	37	0,613636	57	0,75
18	0,75	38	0,72973	58	0,6
19	0,84	39	0,666667	59	1
20	0,785714	40	0,545455	60	0,5

На основе полученных рейтингов можем вынести решение о выдаче кредита. Выделим 3 варианта:

Рейтинг клиента лежит в полуинтервале (0.7, 1] – 94 500 клиентов , значит клиент считается надёжным. Ему можно выдать кредит.

Рейтинг клиента лежит в полуинтервале (0.5, 0.7] – 46 500, значит клиент считается «среднерисковым». Если пересмотреть условия кредитования, то клиенту можно будет выдать кредит.

Рейтинг клиента меньше 0,5 клиент считается рисковым. Ему не стоит выдавать кредит – 10 000 клиентов.

Исходя из этого, получаем классификацию: клиенты из группы 60 рисковые; клиенты из групп 1-5, 7-12, 15, 17, 25, 31, 32, 33, 35-37, 39, 40, 42, 54-56, 58 – «среднерисковые»; клиенты из групп 6, 13, 14, 16, 18-24, 26-30, 34, 38, 41, 43-53, 57, 59 – надежные.

3.2 Построение модели логистической регрессии в DEDUCTOR

Логистическая регрессия – это математическая и статистическая модель, которая используется для определения вероятности для исследования, события. В рамках данной научной работы изучается вопрос вероятности возврата/невозврата кредитной ссуды, которую возьмёт потенциальный клиент

в будущем. Эксперимент базируется на выборке состоящей из 150 000 клиентов банковского учреждения. Каждый клиент обладает определённым признаком, который включает в себя: пол, месячный доход (доллары), возраст, количество членов семьи находящихся на иждивении, оборот по кредитным картам, выходил ли данный клиент в просрочку от 30 и более дней и т.д. Представленные параметры послужат входными данными, необходимыми чтобы построить данную модель в аналитической платформе DEDUCTOR.

В рассматриваемой выборке используются 150 000 клиентов, среди которых 20 000 имею «отрицательную» кредитную историю – это означает, что данные заёмщики уже «декретировали» себя в отношении обязательств по кредиту и уже находятся в группе риска. Среди методов построения скоринговых моделей наиболее распространёнными являются способы такие, как «прореживание выборки» и взвешивание наблюдений. Составляющей операции «прореживания» является отбрасывания части «положительных» заёмщиков, а для реализации построения скоринговой карты (в данном случае логистической регрессии) используется последующий второй и третий «положительный» клиент. Данный алгоритм приводит к увеличению доли «недобросовестных» клиентов в выборке.

«Взвешивание» выборки даёт возможным лучше проанализировать имеющуюся выборку, дав в итоге заданные пропорции «положительных» и «отрицательных» заёмщиков. Суть метода взвешивания заключается в том, что оно увеличивает репрезентативность выборки, позволяет максимально приблизить долю «отрицательных» и «положительных» заёмщиков к первоначальной совокупности [22].

Для построения скоринговой модели на основе логистической регрессии в DEDUCTOR необходимо пройти несколько важных этапов, каждый из которых заключается в обработке входных данных и их анализе не пригодность в рамках исследования. В данном исследовании использовались такие виды алгоритма, как: сэмплинг, категоризация, оценка качества модели и расчёт получаемых весов и бальных критериев.

В качества источника информации для исследования выступает набор данных, включающий себя параметры 150 000 обезличенных клиентов немецкой финансовой организации. Данная выборка была поделена на два класса: 1 – клиент дефолтный и кредит не вернёт; 0 – клиент является благонадёжным, и можно не сомневаться в его добросовестности.

Признаки клиента состоят из 12 параметров, среди них такие как пол, возраст, наличие и количество иждивенцев в семье, месячный доход, кредитная история и дни просрочки прошлых кредитов, если они были.

Первым этапом при загрузки данных в аналитическую платформу DEDUCTOR следует исследование выборки по параметру «Качество данных». Данное исследование является предпосылкой для дальнейшего изучения выборки на применимость логистической регрессии, как математической модели, для построения скоринговой карты. Алгоритм «Качество данных» в аналитической платформе DEDUCTOR позволяет выявить данные, а рамках исследования – это параметры заёмщиков, по степени необходимости их изменения и «исправления».

В исследовании параметры были разделены по категориям типов данных: вещественный, логический, целый. Логическая категория представлена в единственном виде, так как представляет собой основополагающую цель данного исследования, выявить дефолтных клиентов из выборки. [23]

Данный параметр – предполагаемая дефолтность, в рамках предсказательной бинарной функции, или свойства логистической регрессии.

В таблице 9 представлена расшифровка используемых в исследовании параметров клиентов. Данная расшифрока позволяет проанализировать статистические показатели выборки на пригодность к построению.

Таблица 9 – распределение параметров заёмщиков по типам и видам данных

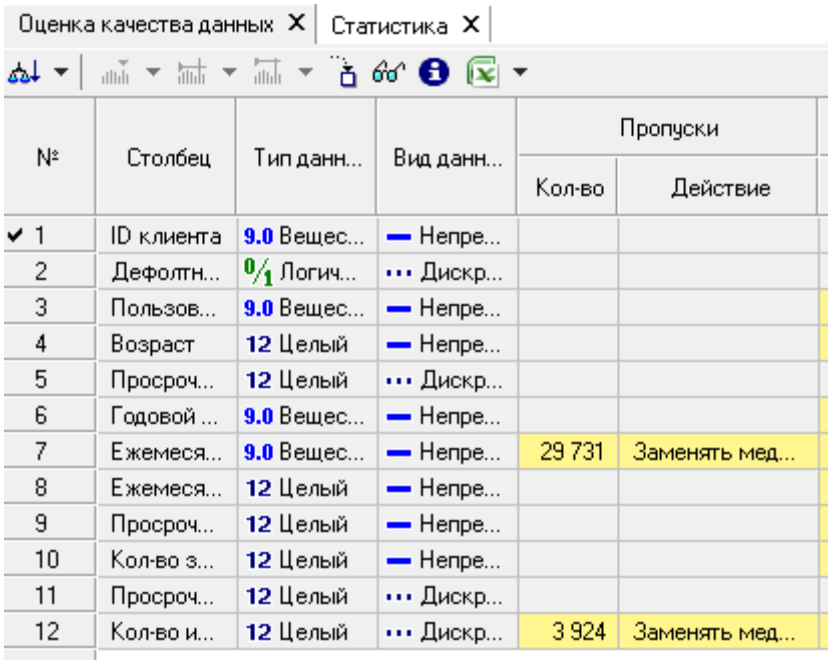
Параметр	Расшифровка	Тип данных	Вид данных
ID клиента	Порядковый номер клиента в выборке	Вещественный	Непрерывный
Дефолтность	Предполагаемая дефолтность клиента	Логический	Дискретный

Продолжение таблицы 9

Кредитные карты	Пользовался или пользуется ли клиент кредитными картами	Вещественный	Непрерывный
Возраст	Количество полных лет клиента	Целый	Дискретный
Просрочка 30-59 дней	Выходил ли клиент в просрочку от 30 до 59 дней	Целый	Непрерывный
Годовой доход	Суммарный ежемесячный доход клиента за 12 месяцев (в долларах)	Вещественный	Непрерывный
Ежемесячный доход	Величина дохода клиента в месяц	Вещественный	Непрерывный
Просрочка 90+	Выходил ли клиента, или находится на просрочке более 90 дней	Целый	Непрерывный
Количество закрытых кредитов	Общее количество закрытых кредитов у клиентов за счёт его средств	Целый	Непрерывный
Просрочка 60-89 дней	Находился ли клиент в просрочке от 60 до 89 дней	Целый	Дискретный

В подобных исследованиях, как построение скоринговой карты, а именно по банковской тематике, важным шагом является определение, какие данные подойдут для изучения. Какие параметры клиента могут при взаимодействии с другими данными, смогут дать в своей совокупности ответ на вопрос исследования. Использование специальных коэффициентов позволяет решить эту задачу. Например, отношение платежа по кредиту к доходу позволяет уже заранее предсказать вероятность дефолта у данного клиента. По практике, если данный коэффициент превышает отметку в 40%, то это указывает на вероятный дефолт заёмщика, а значит риск финансовых потерь для банка [23].

На рисунках 4 и 5 представлена визуализация метода «Качество данных».



Оценка качества данных | Статистика

№	Столбец	Тип данн...	Вид данн...	Пропуски	
				Кол-во	Действие
✔ 1	ID клиента	9.0 Вещес...	— Непре...		
2	Дефолтн...	0/1 Логич...	··· Дискр...		
3	Пользов...	9.0 Вещес...	— Непре...		
4	Возраст	12 Целый	— Непре...		
5	Просроч...	12 Целый	··· Дискр...		
6	Годовой ...	9.0 Вещес...	— Непре...		
7	Ежемеся...	9.0 Вещес...	— Непре...	29 731	Заменять мед...
8	Ежемеся...	12 Целый	— Непре...		
9	Просроч...	12 Целый	— Непре...		
10	Кол-во з...	12 Целый	— Непре...		
11	Просроч...	12 Целый	··· Дискр...		
12	Кол-во и...	12 Целый	··· Дискр...	3 924	Заменять мед...

Рисунок 4 – Визуализация метода «Качество данных»

На рисунке 4 представлено общее количество параметров, которые используются в данном исследовании. Продемонстрированы типы и виды данных, к которым было необходимо их отнести. Графа «Пропуски» отображает общее количество «поражённых» или «дефолтных» данных, которые обнаружила система, так же предлагается способ исправления данных проблем.

Выбросы		Экстремальные		Кол-во уникаль...	Качество данных	Резюме
Кол-во	Действие	Кол-во	Действие			
					1.0000	Пригоден
		10 026	Заменят...	2	0.3540	Предобр...
31	Ограничи...	160	Ограничи...		0.0021	Предобр...
46	Ограничи...				0.7949	Предобр...
		269	Заменят...	16	0.2193	Предобр...
421	Ограничи...	178	Ограничи...		0.0020	Предобр...
201	Ограничи...	120	Ограничи...		0.0010	Предобр...
1 752	Ограничи...	146	Ограничи...		0.6236	Предобр...
9	Ограничи...	269	Ограничи...		0.0078	Предобр...
1 009	Ограничи...	473	Ограничи...		0.0907	Предобр...
		269	Заменят...	13	0.0951	Предобр...
209	Заменят...	36	Заменят...	13	0.4603	Предобр...

Рисунок 5 – Продолжение алгоритма «Качество данных»

Благодаря визуализатору «Статистика», данные которой представлены в таблице 10 – можно наглядно увидеть детальное рассмотрение каждого параметра, и его расшифровку по математическим составляющим. В исследовании используются данные 150 000 клиентов.

Таблица 10 – Статистические данные параметров клиентов

Параметр	Минимальное значение	Максимальное значение	Среднее	Стандартное отклонение	Сумма
ID клиента	1	150 000	75 000,5	43301,41	1,125007
Дефолтность	–	–	0,06684	0,249	10026
Кредитные карты	0	50 708	6,0484	249,775	907265,7
Возраст	0	60	30,295	14,771	7844
Просрочка 30-59 дней	0	98	4,19	4,19	63155
Годовой доход	0	329 664	363,155	2139,709	54473256
Ежемесячный доход	0	30 8750	6670,02	14384,67	802208
Просрочка более 90 дней		98	0,2403	5,145	126794
Количество закрытых кредитов	0	54	1,01	4,169	39896
Просрочка 60-89 дней	0	98	0,24	4,1551	152736
Количество иждивенцев	0	20	0,75	1,115	110612

Показанная разбивка сатистических данных позволяет избеджать лишней погрешности при расчётах и увидеть параметры, которые нуждаются в «переработке». Например если величины одного параметра не соотностся друг с другом в математическом соотношении, это сигнализирует о том, что необходимо заново настроить выборку в исходном файле, который изначально был загружен в аналитическую платформу.

Визуализатор статистических показателей выборки представлен на рисунке 6.

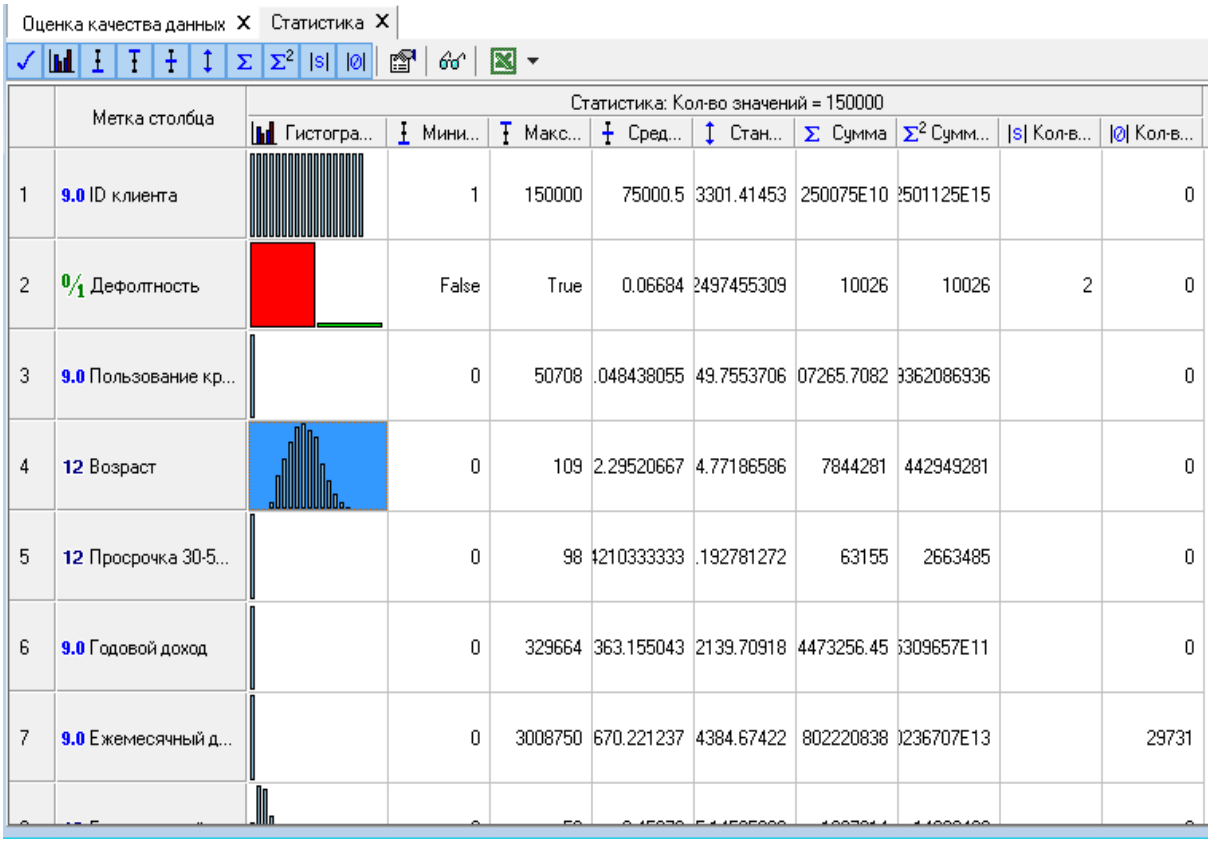


Рисунок 6 – Визуализатор статистических показателей

Платформа DEDUCTOR при оценки качества данных и их обработки, для возможности построения модели и проведения математически анализов, позволяет более наглядно представить каждый параметр выборки – разбив их на процентные соотношения.. Из всех параметров представленном на скришоте, «красным» цветом отмечена «Дефолтность».

Данный параметр является «выходным», именно он будет демонстрировать предсказаетльную фукцию логистической регрессии, в качестве ответа а поставленный вопрос. Несорразмерность гистограммы на даном скриншоте у параметра «Дефолтность», помогает уже на начальном этапе увиедть примерное процентное соотношение вероятных «дефолтных» и «недефолтных» клиентов. Изменеие гистограммы зависит от тестовой и эксперемтной выборки, которая используется в исследоваии, наряду с качеством данных. Данная процедура представлена на рисунке 7.

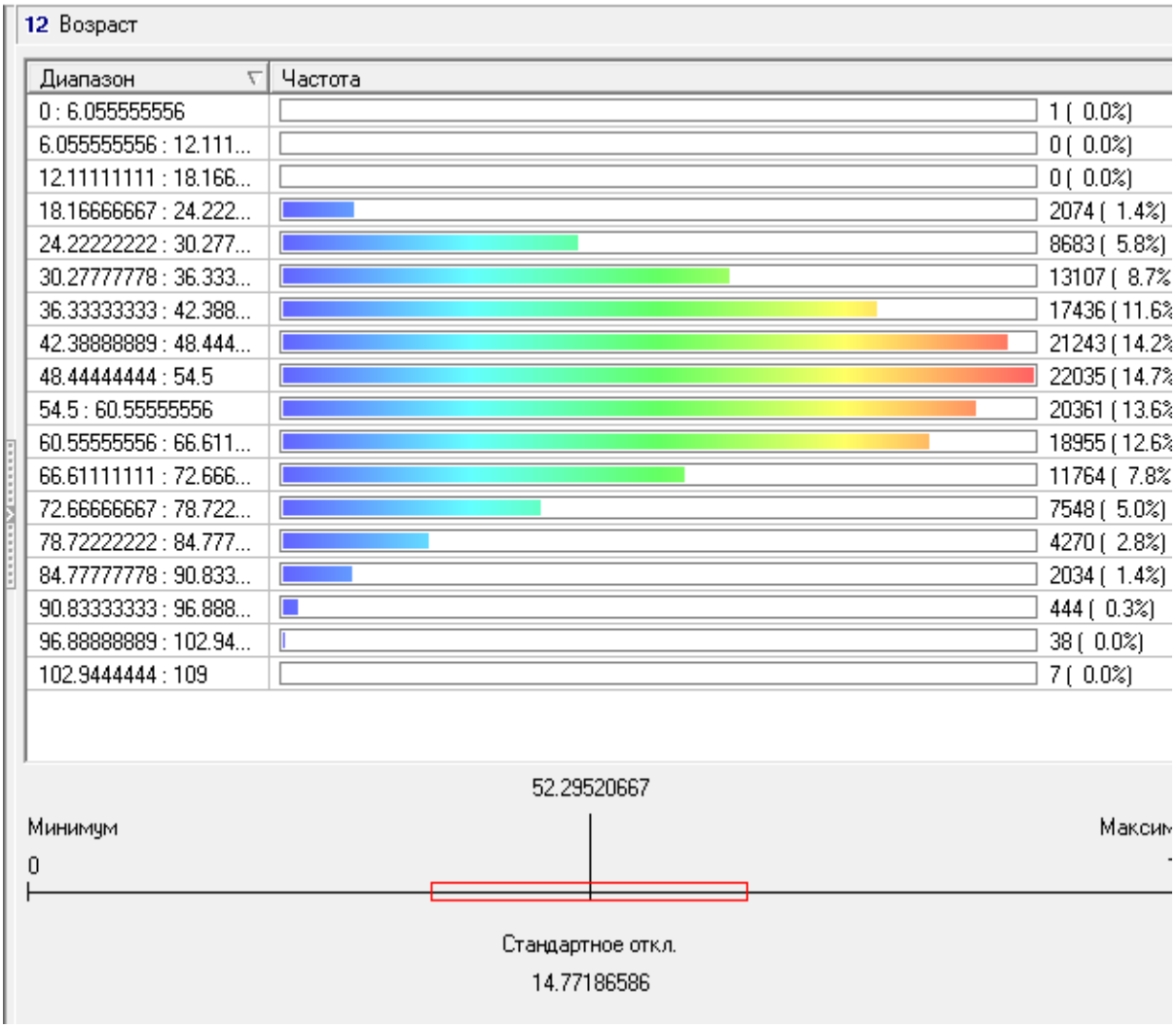


Рисунок 7 – Процентное соотношение по параметру «Возраст»

Как можно увидеть из представленного выше скриншота, категория возраст разбита на 20 подкатегорий, с интервалом 6-7 единиц. Среди клиентов финансового учреждения, которыми представлена данная выборка, наибольший процент по количеству имеют лица в возрасте от 42 до 54 лет (14,2 % и 14,7% соответственно).

Наличие в данной разбивке категорий возраста выше 60 лет объясняется особенность кредитной политики Германии, где социальные льготы пенсионерам позволяют оформлять небольшие ссуды по гарантию родственников. Более детальная разбивка параметра «Возраст» представлена в таблице 11.

Таблица 11 – Категориальная разбивка параметра возраст

Категория возраста	Процентное соотношение
18-24	1,4%
24-30	5,8%
30-36	8,7%
36-42	11,6%
42-48	14,2%
48-54	14,7%
54-60	13,6%
60-66	12,6%
66-72	7,8%

Согласно таблице 11 можно сделать вывод, что у данного кредитного учреждения минимальный процентный порог у категории 1,4% – данные лица среди реальных заёмщиков наименее часто встречаются в выборке.

Коэффициент «Отношение Шансов» – OR для категории возраст в исследуемой выборке можно представить в ввиде диапазона от 36 до 54 лет, ввиду наибольшего дольного процента у данных категорий. Можно сделать вывод если клиент находится в диапазоне данных категорий, то вероятность его попадания в группу «Благонадёжных» клиентов увеличится в 7,14 раз.

Управление статистическими показателями позволяет более детально рассмотреть и другие параметры. Например, категория «Количество закрытых кредитов» тоже является важным нюансом для определения клиента по степени дефолтности [24].

Данный признак заёмщика позволяет просмотреть его кредитную историю. Выявить «хороших» клиентов и разделить заёмщиков по категориям «новый клиент» и «старый клиент, что также влияет оценку «благонадёжности» клиента. Детальная расшифровка данного параметра, представлена на рисунке 8.

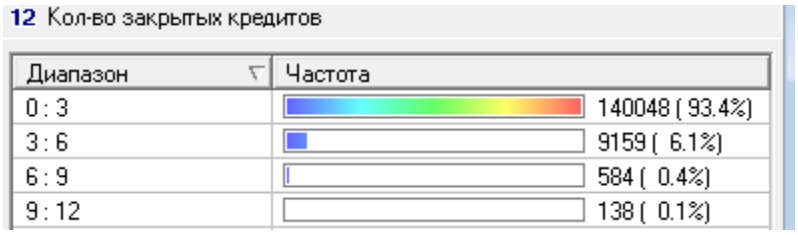


Рисунок 8 – Визуализация параметра «Количество закрытых кредитов»

Представленная выше статистика по категории «Количество закрытых кредитов» демонстрирует факт того, что наибольшее число заёмщиков из выборки, а именно 93,4% – 140 048 физических лиц либо являются новыми клиентами данного банка, либо имели от до кредитов и успешно их закрывали. Около 10 000 клиентов имеют в данном банке большее количество закрытых кредитов, тем самым относя себя к категории самых желательных клиентов банковского учреждения. Но весовая доля данных заёмщиков в исследуемой выборке очень не велика, что не влияет на общую статистику.

Процентная разбивка показателя «Количество иждивенцев» представлена на рисунке 9.

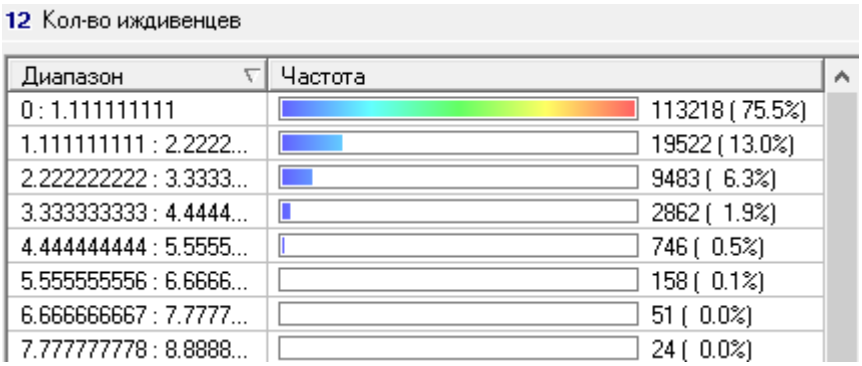


Рисунок 9 – Визуализация статистических показателей параметра «Количество иждивенцев»

Согласно данному скриншоту с оригинального исследования, можно проследить, что наибольший процент в выборке имеют клиенты, у которых либо нет лиц на обеспечении, либо до одного физического лица. Эта категория представлена 75,5% от выборки и в количестве 113 218 человек.

Для оценки качества выборки исследуются показатели на степень их влияния в данном исследовании на поставленную задачу. А именно на

вероятную дефолтность клиентов, которую возможно будет проследить благодаря предсказательной функции логистической регрессии. Данный вид анализа и вычислений в платформе DEDUCTOR проводится с помощью корреляционного анализа – анализа на степень соотношения влияния параметров выборки друг на друга. Результат данного анализа представлен на рисунке 10.

Входные поля		Корреляция с выходными полями
№	Поле	Дефолтность
1	Пользование кредитными картами	0.112
2	Возраст	-0.115
3	Просрочка 30-59 дней	0.271
4	Годовой доход	-0.021
5	Ежемесячный доход	-0.032
6	Просрочка более 90 дней	0.353
7	Кол-во закрытых кредитов	-0.020
8	Просрочка 60-89 дней	0.266

Рисунок 10 – «Матрица корреляции»

Согласно «Матрице корреляции» можно сделать вывод о том, что наибольшую весовую долю в выборке имеют такие показатели как: факт нахождения в просрочке от месяца до 59 дней, факт нахождения в просрочке более 90 дней (в данном исследовании рассматривается факт нахождения в просрочке у закрытых кредитов в прошлом, либо просрочка в данный момент, в момент открытого действующего кредита), факт нахождения в просрочке от двух месяцев до 89 дней. Данный результат можно объяснить особенностью функционирования банковской системы и её взаимодействием с потенциальным и действительным клиентом. Наибольшую степень риска для банка несут заёмщики, которые выходят на просрочку с момента первого платежа, что и представлено в «Матрице корреляции». Тем самым они являются «дефолтными» клиентами, попадаю автоматически в данную категорию в проводимом оригинальном исследовании [25].

Основополагающим этапом построения включает в себя работу с мастером обработки «конечные классы». Исследования данного алгоритма необходимо для переработки выборки и повышения степени качества получаемой на выходе логистической регрессии.

Использование данного метода целесообразно в случаях:

Необходимость сохранения информации;

Исправление выбросов;

Заполнением пропусков;

Уменьшения размерности параметров

В качестве ограничителей для конечных классов выступают:

минимальное число наблюдений в конечном классе;

минимальное число конечных классов

Алгоритм уменьшения уникальных параметров признака выполняется в данной последовательности:

Создаётся начальная выборка оригинальных параметров для исследования их методом «начальные классы»;

В рамках исследования можно задать любые другие величины данных параметров, необходимых для реализации метода «конечные классы». Визуализация данного алгоритма оригинального исследования представлена на рисунке 11.

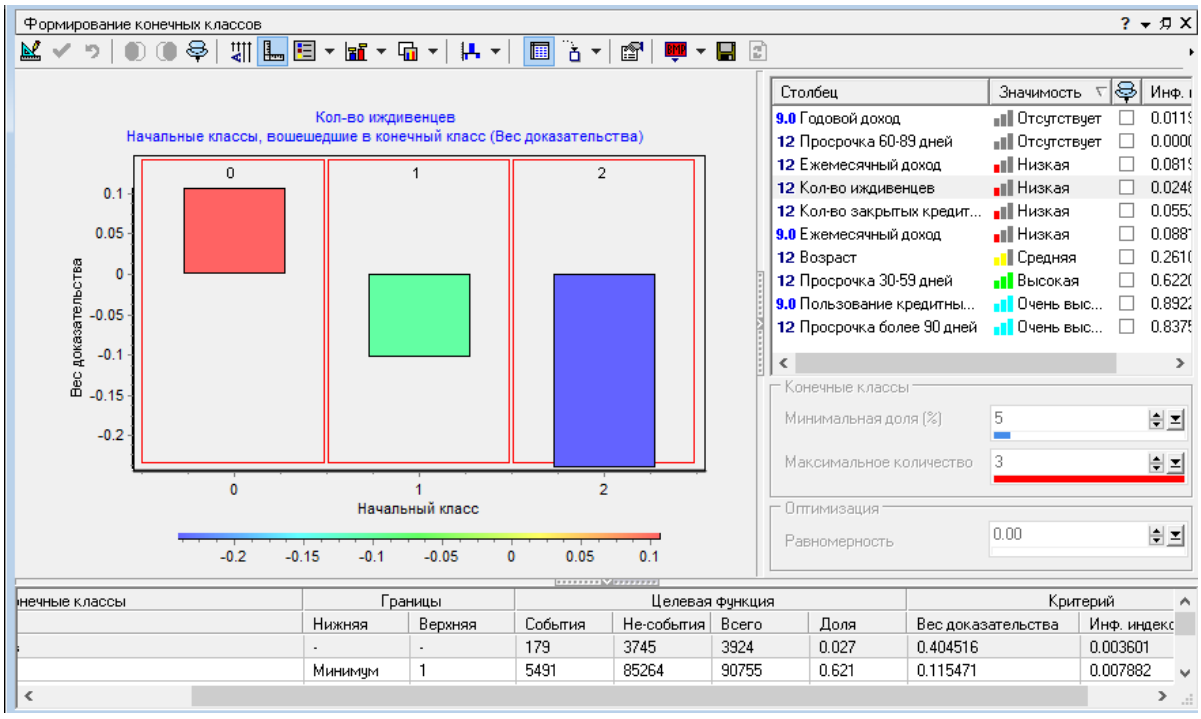


Рисунок 11 – Реализация метода «Конечные классы» в DEDUCTOR

На рисунке 11 можно так же увидеть вес параметров исследования, которые разбиты по группам значимости, исходя из анализа выборки. Согласно данном анализу можно выделить пять категорий веса:

Значимость (вес) отсутствует – её можно не использовать в исследовании;

низкая значимость для данного исследования;

средняя значимость;

Высокая значимость;

очень высокая значимость.

Исходя из аннограмы, представленной выше, можно заметить что, согласно анализу «Конечные классы» наивысшую степень веса («очень высокую значимость») имеют параметры: пользование кредитным картами, и наличие в просрочке 90 и более дней по раннее взятым кредитам. Следует отметить, что выход клиента на просрочку 90 и более дней с момента первого платежа автоматически приравнивает его категории «Дефолтный клиент». Поэтому выходной анализ при использовании данного алгоритма можно считать актуальным и обоснованным. Также высокую степень значимости для данного исследования выступает признак – просрочка 30-59 дней с момента первого платеже. Данная информация по клиенту, обладание им такого признака, так же позволяет отнести его в категорию «Дефолтный клиент».

Для реализации метода «Конечные классы» используется Wo анализ – методология доказательств. Данный алгоритм включает в себя математический метод оценки факторов на актуальность выдвигаемого исследования.

Следующим этапом обработки данных выступает удаление из исследования тех параметров заёмщика, которые не имеют весовой значимости. А значитт эти показатели не влияют на выход клиента в просрочку-дефолт. На рисуке 12 представлен иснтрумент выполнения данного алгоритма.

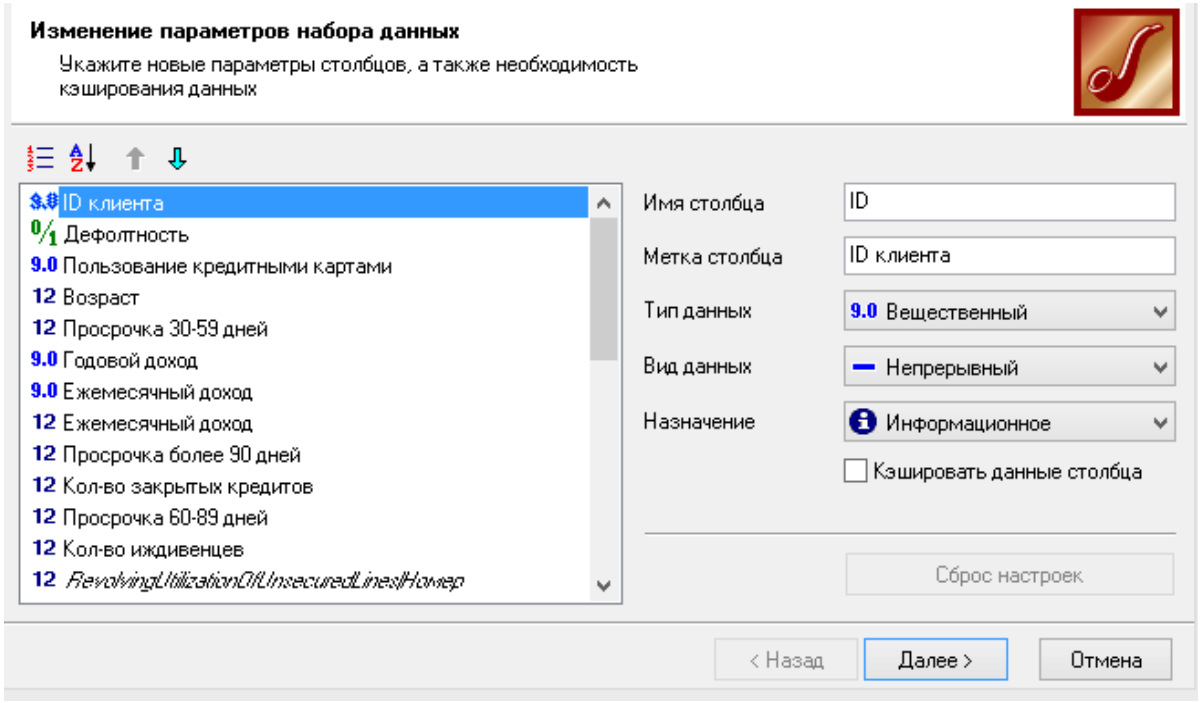
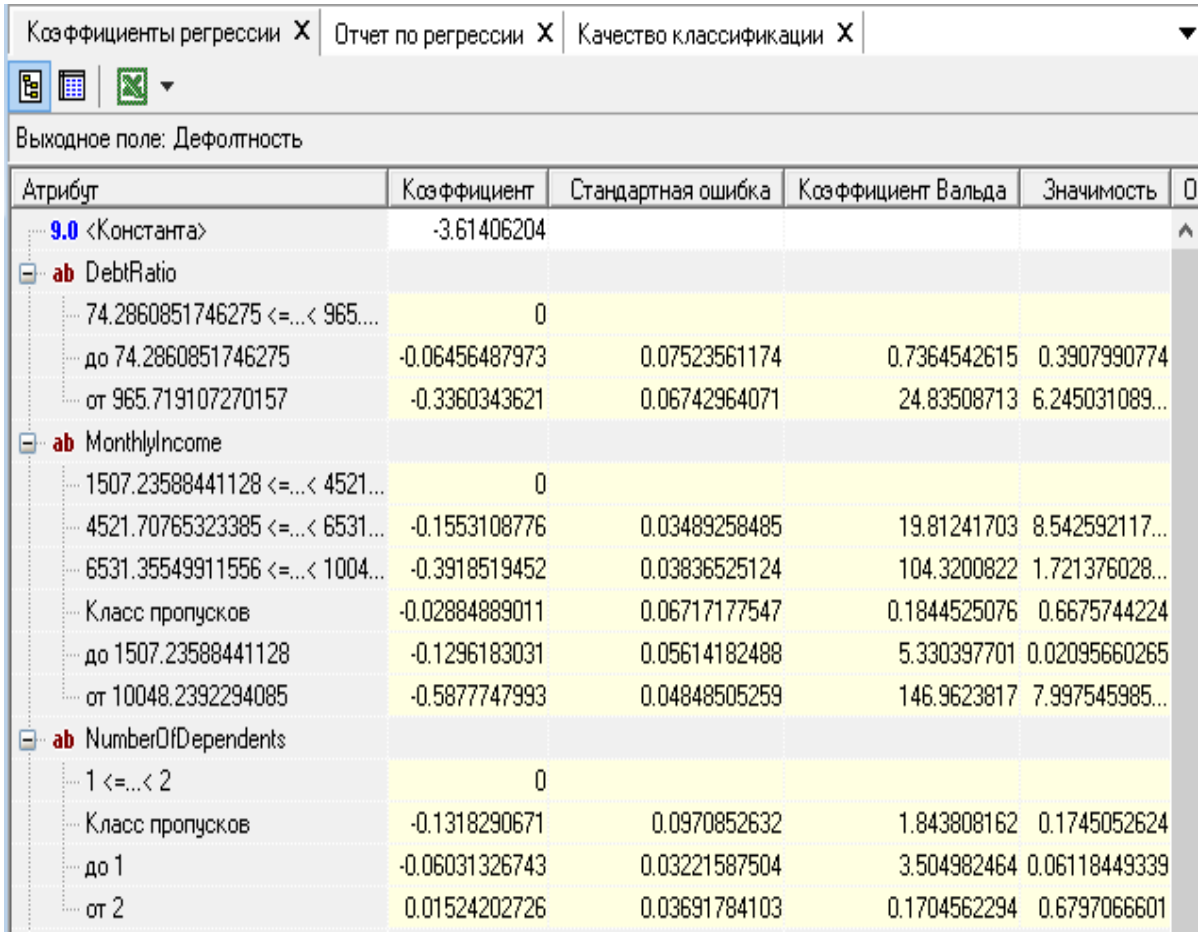


Рисунок 12 – Панель изменения набора данных исследования

В качестве методики выполенения данного алгоритма выступает нахождение столбцов с малой весовой долей и изменние у них колонки «Назначение» на «Информационное» значение. Данные параметры после проведения этой операции перестанут влиять на модель, полученную на платформе.[26]

Следующим шагом агоритмизациит предложено настроить обучающие и тестовые множества, в рамках данного исследования было решено сделать 20 % тестовым. В последствии праграмма предлагает изменить параметры алгоритма логистической регрессии . По умолчанию порог классификации равен 0,5 , что является оптимальным и для рассматриваемой модели.

Заключительным шагом является построение самой модели логистической модели и рассматроение полученных коэффициентов. Визуализатор Коэффициенты регрессии наглядно показывает рассчитанные коэффициенты логистической регрессии, которые являются прототипом скоринговой карты. Данные представленны на рисунке 13.

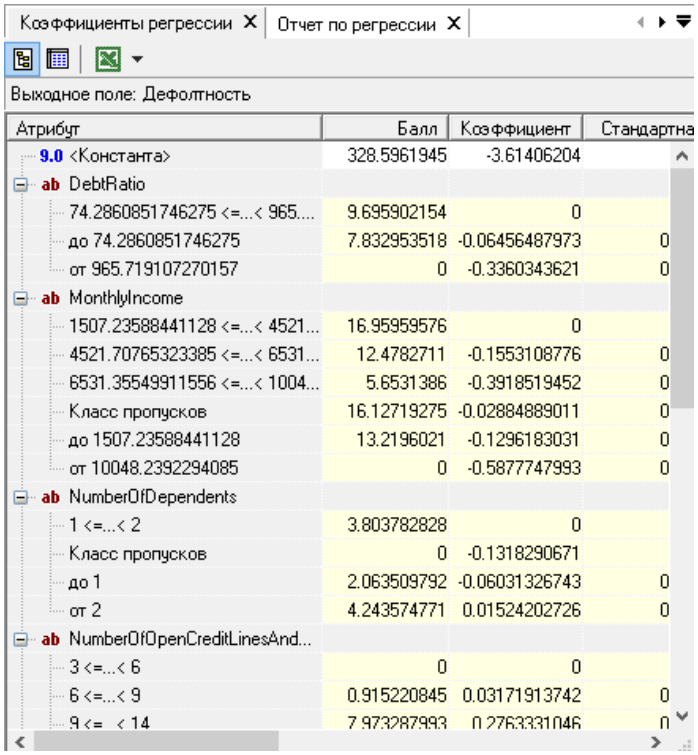


Атрибут	Коэффициент	Стандартная ошибка	Коэффициент Вальда	Значимость	О
9.0 <Константа>	-3.61406204				
ab DebtRatio					
74.2860851746275 <=...< 965....	0				
до 74.2860851746275	-0.06456487973	0.07523561174	0.7364542615	0.3907990774	
от 965.719107270157	-0.3360343621	0.06742964071	24.83508713	6.245031089...	
ab MonthlyIncome					
1507.23588441128 <=...< 4521...	0				
4521.70765323385 <=...< 6531...	-0.1553108776	0.03489258485	19.81241703	8.542592117...	
6531.35549911556 <=...< 1004...	-0.3918519452	0.03836525124	104.3200822	1.721376028...	
Класс пропусков	-0.02884889011	0.06717177547	0.1844525076	0.6675744224	
до 1507.23588441128	-0.1296183031	0.05614182488	5.330397701	0.02095660265	
от 10048.2392294085	-0.5877747993	0.04848505259	146.9623817	7.997545985...	
ab NumberOfDependents					
1 <=...< 2	0				
Класс пропусков	-0.1318290671	0.0970852632	1.843808162	0.1745052624	
до 1	-0.06031326743	0.03221587504	3.504982464	0.06118449339	
от 2	0.01524202726	0.03691784103	0.1704562294	0.6797066601	

Рисунок 13 – Визуализатор «Коэффициенты регрессии»

В данном визуализаторе присутсвует та часть исследования, которая характеризует скоринговую карту – бальная система распределения заёмщиков по категориальной шкале – от «плохих» к «хорошим». Следует отметить, что значимость показателя в системе определяется экономистом индивидуально для каждого заемщика в зависимости от политики данного коммерческого банка, особенностей клиента, ликвидности его баланса, положения на ссудном рынке. Например, высокая доля краткосрочных ресурсов, наличие просроченной задолженности по ссудам и неплатежей поставщикам повышают роль коэффициента ликвидности, который оценивает способность предприятия к оперативному высвобождению денежных средств. Втягивание ресурсов банка в кредитование постоянных запасов,заниженность размера собственных средств повышают рейтинг показателя обеспеченности собственными средствами.[26] Нарушение экономических границ кредита, закредитованность клиентов выдвигают на первое место при оценке кредитоспособности уровень

коэффициента покрытия. Данные бальные признаки продемонстрированы на рисунке 14.



Атрибут	Балл	Коэффициент	Стандартна
9.0 <Константа>	328.5961945	-3.61406204	
ab DebtRatio			
74.2860851746275 <=...< 965....	9.695902154	0	
до 74.2860851746275	7.832953518	-0.06456487973	0
от 965.719107270157	0	-0.3360343621	0
ab MonthlyIncome			
1507.23588441128 <=...< 4521...	16.95959576	0	
4521.70765323385 <=...< 6531...	12.4782711	-0.1553108776	0
6531.35549911556 <=...< 1004...	5.6531386	-0.3918519452	0
Класс пропусков	16.12719275	-0.02884889011	0
до 1507.23588441128	13.2196021	-0.1296183031	0
от 10048.2392294085	0	-0.5877747993	0
ab NumberOfDependents			
1 <=...< 2	3.803782828	0	
Класс пропусков	0	-0.1318290671	
до 1	2.063509792	-0.06031326743	0
от 2	4.243574771	0.01524202726	0
ab NumberOfOpenCreditLinesAnd...			
3 <=...< 6	0	0	
6 <=...< 9	0.915220845	0.03171913742	0
9 <= ... < 14	7.973287993	0.2763331046	0

Рисунок 14 – Бальная система классификации клиентов

Из данного представления следует, что максимальный балл, который может получить клиент, оценный по данной скоринговой карте, равен 328 баллам. Вариант отчёта по модели скоринговой карты «Отчёт по регрессии» позволяет увидеть статистику использования данных в построение модели. Согласно выводам по построенной логистической регрессии в данном исследовании использовалось 53% весомых коэффиентов наборов параметров клиента.[27] Статистика представлена двумя направлениями: «Финальная» - которая используется при построении модели, и «Нулевая» – те данные, которые были отброшены. «Отброшенные» данные являются теми параметрами, которые система посчиталал «мало-весомыми» для разрабатываемой скоринговой карты. Визуализатор данной статистики представлен на рисунке 15.

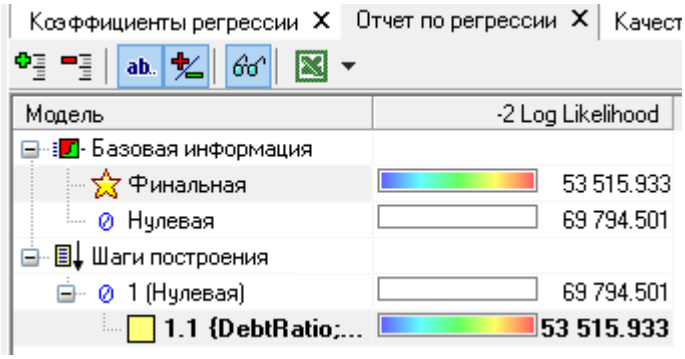


Рисунок 15 – «Базовая информация» построенной логистической регрессии

Логистическая регрессия на выходе рассчитывает значение рейтинга, которое можно трактовать как вероятность того, что событие наступит для конкретного испытуемого. Поэтому часто желательно указать, вероятность какого именно (из двух вариантов выходного поля) события будет оцениваться, чтобы оно кодировалось истиной. В данной кодировки используются математические коэффициенты, представленные на рисунке 16.

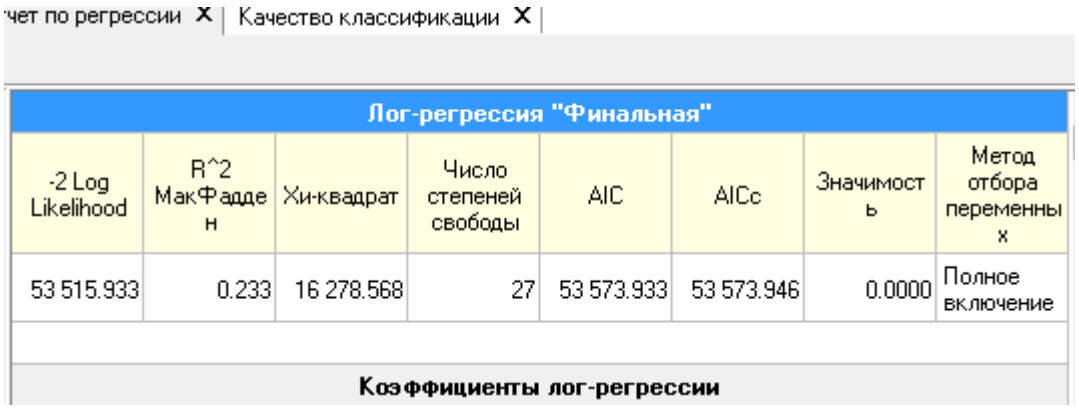


Рисунок 16 – Визуализатор схемы математических коэффициентов регрессии

Непосредественно после анализма и простотра коэффициетов «пред-модели» скоринговой карты, необходимо построить саму логистическую регрессию, которая представлена на рисунке 17. Логическая регрессия является разновидность множественной регрессии, и позволяет более эффективно анализировать выборку по бинарному классфификатору. В случае данного исследования вернёт ли потенциальный заёмщик кредитную ссуду или нет – предсказательная функция дефолтности.

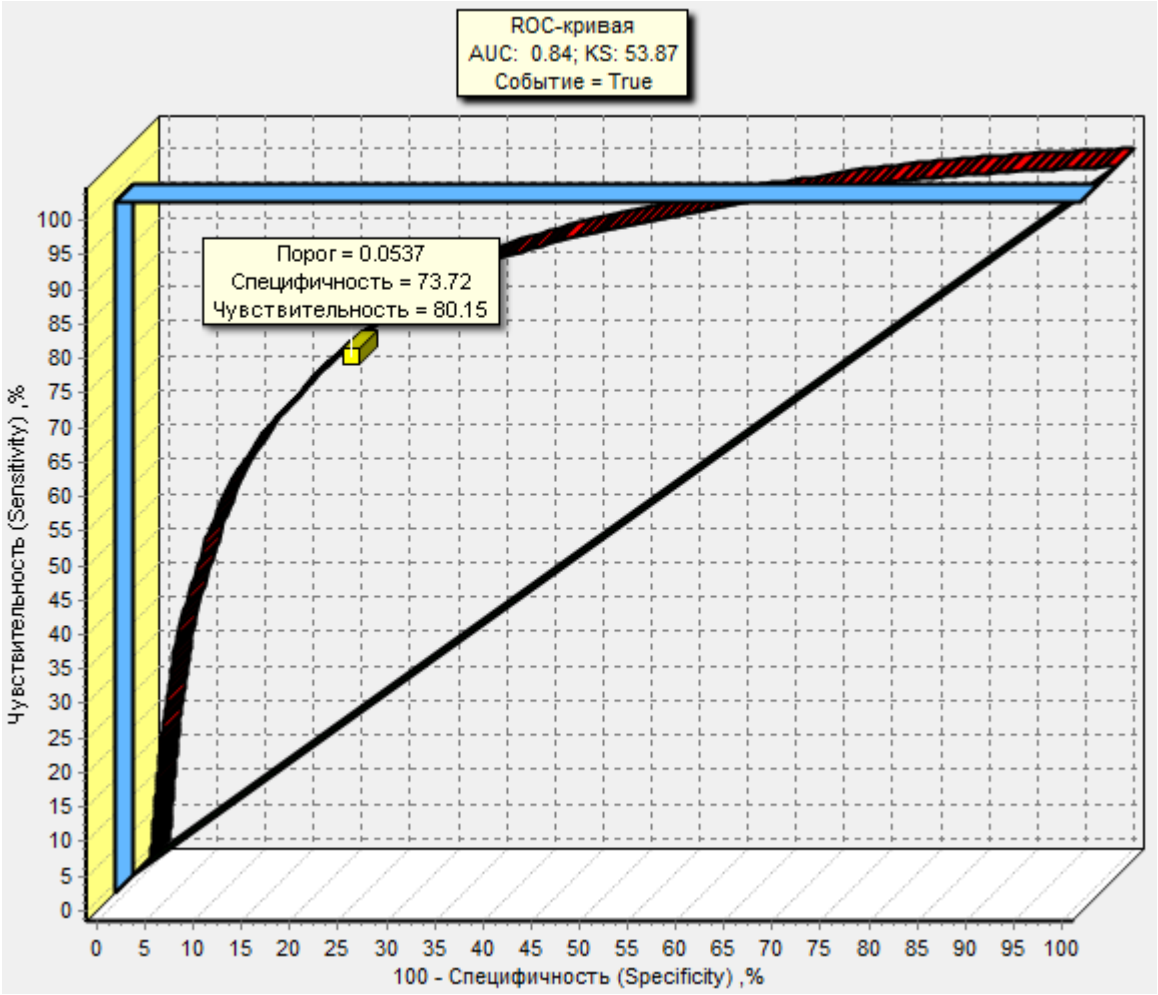


Рисунок 17 – График модели логистической регрессии

Представленная на рисунке 17 модель логистической регрессии обозначена красным цветом. Сопряженные линии с ней: нижняя линия – демонстриет самый худший вариант исхода событий, тоестьв рамках данного исследования – вернут ли клиенты кредит в банк. Линия, обозначенная голубым цветом, демонстрирует так называемую «идеальную» модель. Чем ближе график логистической регрессии к этой линии, тем более построенная нами модель является оптимальной для ответа на поставленный вопрос.[28]

Логистическая регрессия представленная в данном исследовании обладает рядом коэффициентов, которые могут отнести её к «идеальной» или «плохой» модели.

В данной модели коэффициент чувстивельности равен 80,15. Что говорит о том, что 80% вероятности правильности выставления баллов клиентам, или что 80% буду распознаны верно.

В рассматриваемой модели коэффициент специфичности имеет вес 73,72. Такая величина коэффициента говорит о том, что 73% дефолтных клиентов будут распозаны – что является очень хорошим предсказательной способностью модели.

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

Если рассуждать в терминах скоринга – задачи определения «благонадежности» клиента, где модель классификации заёмщиков на дефолтных и недефолтнымх называется диагностическим тестом, то получится следующее:

Чувствительный диагностический тест проявляется в скоринге – максимальном предотвращении пропуска дефолтных клиентов.[29]

В рамках вывода по данной логистической регрессии по выборке 150 000 наблюдений можно разделить рассматриваемых клиентов с долей «плохих» наблюдений 26% и «хороших» 74%. Точность предсказания «плохих» наблюдений составила всего 50%.

Было реализовано два альтернативных подхода. В первом были взяты все «плохие» наблюдения и случайно отобрано 50% «хороших» наблюдений.

Таким образом, соотношение «плохих-хороших» составило 42% и 58% соответственно. Объем откорректированной выборки составил 9 757 наблюдений. Это позволило увеличить процент верно предсказанных «плохих» наблюдений до 71%. При этом доля верного предсказания «хороших» наблюдений составила 82%.

Во втором подходе все данные были перевзвешенны, чтобы соотношение «плохих-хороших» в выборке было также 30% и 70% процентов. Число наблюдений в выборке сохранилось и составило 15000.

Это привело практически к идентичным результатам с точки зрения точности предсказания наблюдений по полученной модели. Данный расчёт представлен в таблице 12.

Таблица 12 – Сравнение точности классификации при различных стратегиях формирования обучающей выборки

Параметр	Исходная модель	Взяты все «плохие» и 50% хороших	Перезвешанные данные, чтобы увеличить долю «плохих»
Количество плохих	49 515	49 515	51 783
Доля плохих	26%	42%	42%
Количество хороших	100 485	99 757	81 694
Доля хороших	74%	68%	58%
Объём выборки	150 000	49 757	132 000
переменная 1	0,57	0,59	0,56
переменная 2	-0,08	-0,07	-0,10
переменная 3	-0,24	-0,24	-0,26
переменная 4	0,09	0,09	0,09
Константа	-0,79	-0,22	-0,95
Процент корректных «Нет»	50,3%	82,1%	79,9%
Процент корректных «ДА»	90,2%	70,5%	73,8%
Общий процент	81,4%	77,2%	77,3%

Интервал коэффициента AUC данной регрессии, по качеству модели, можно разбить на отрезки, представленные в таблице 13. Показатель AUC может изменяться от 0,5 («бесполезный» классификатор) до 1,0 (идеальная модель), в данной модели AUC=0,84. Вес данного коэффициента позволяет утверждать, что представленная модель является «адекватной» и будет максимально информативно отвечать на заданный вопрос исследвания. Коэффициент AUC имеет некоторую специфику, которая менятся в зависимости от его велечины. Если его вес максимален или приближен к значению – 1, то можно утверждать, что полученная модель не является адекватной и не отвечает на поставленный вопрос. Это исходит от того, что на

практике идеальных моделей не существует. Размерность коэффициента отображена в таблице 13.

Таблица 13 – Интервал коэффициента AUC

Интервал AUC	Качество модели
0,8-0,9	Отличное
0,7-0,8	Очень хорошее
0,6-0,7	Хорошее
0,5-0,6	Удовлетворительное

С большими допущениями можно считать, что чем больше показатель AUC, тем лучшей прогностической силой обладает модель. Однако следует знать, что [30]:

- показатель AUC предназначен скорее для сравнительного анализа нескольких моделей;

- AUC не содержит никакой информации о чувствительности и специфичности модели.

В результате анализа функциональных возможностей платформы Deductor и применении их в построении скоринговой модели можно сделать вывод, что грамотный выбор множества значений для входных переменных повышает качество и скорость построения модели. Отбор переменных для модели в ходе оценки обеспечивает наибольшую репрезентативность результатов, делает скоринговую модель устойчивой и также повышает скорость ее построения. Все этапы моделирования можно выполнить без использования специальных программных средств, но использование Deductor позволяет аналитику сконцентрироваться на интеллектуальной работе и сделать процесс анализа данных полуавтоматическим. Таким образом, возможно сократить время на создание и повысить качество скоринговой модели.

ЗАКЛЮЧЕНИЕ

В данной работе были описаны наиболее применяемые методы для построения скоринговой моделей. Приведено теоретическое и практическое обоснование эмпирического Байесовского метод, а также было осуществлено построение модели скоринговой карты на основе логистической регрессии. Данная модель, визуализированная в аналитической платформе DEDUCTOR 5.3, обладает параметрами, которые могут отнести её в категорию «идеальных» моделей. Однако следует учесть, что «идеальных» моделей на практике на существует. Если построенная на определённой выборке модель обладает вышеописанными качествами, можно утверждать о её неспособности дать актуальный ответ на поставленную задачу. Данное исследование было применено к реальным кредитным обезличенным данным банковского учреждения и данная модель послужит для развития и создания более эффективной модели скоринговой карты для ОАО «АТБ» – «Азиатский Тихоокеанский Банк». В результате были получены рейтинги клиентов и была выявлена основная категория клиентов в двух параметрах – дефолтные и не дефолтные заёмщики.

Был предложен и реализован новый метод анализа скоринговой карты, основанный на визуализации модели с помощью построения логистической регрессии и расчёта рейтингов клиентов с помощью эмпирического Байесовского метода. Описано его применение к тем же финансовым данным. Полученная комбинация методов позволит более эффективно производить оценку клиентов банка в рамках их «благонадёжности». Следующим этапом развития направления данного исследования может послужить использование Байесовского метода в выявлении рейтинговых показателей клиентов, осуществляемый перед построением логистической регрессии. И данные операции позволят в скоринговую карту внести новый параметр – рейтинг заёмщика.

Из этого можно сделать вывод о том, что представленные в данной работе методы анализа данных скоринговой карты может послужить новой ступенью для развития кредитного скоринга. При этом метод основанный на методе Байесовского выделяется тем, что математически обоснован и определены условия его применения.

В настоящее время, скоринг получил широкое распространение в России и показал себя эффективным финансовым инструментом. В качестве примера, можно назвать область розничного кредитования. В данной области отказались от использования экспертной оценки в пользу скоринговых систем.

Однако, не смотря на широкое распространение, скоринг слабо освещён в российской литературе, хотя ему посвящено множество зарубежных работ. Скоринг обладает большим потенциалом применения, но все ещё является для людей использующих его «черным ящиком». Следует продолжать изучать и совершенствовать скоринговые системы.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1 Сиддики, Н. Скоринговые карты для оценки кредитных рисков: / Н. Сиддики – М.: НИЯУ МИФИ, 2014. – 142 с.

2 Мэйлз, Э. Руководство по кредитному скорингу / Э.Мэйлз. – СПб.:СПбПУ, 2008. – 266 с.

3 Сергиенко, Д.С. Легенды и мифы скоринга / Д.С. Сергиенко. - М.: Мир, 2013. – 176 с.

4 Сорокин, А. С. Развитие современного скоринга/ А.С. Сорокин. – Новосибирск: Изд-во НГТУ, 2015. – 382 с.

5 Воронцов, К.В. Лекции по кредитному скорингу / К.В.Воронцов. – М.: МГУ, 2014. – 298 с.

6 Зволыхин, Е.С. Интеллектуальные модели анализа кредитной информации / Е.С. Зволыхин. – Л.: ЛГУ, 2010. – 324 с.

7 Кабаков, Р.И. Анализ и визуализация кредитных данных / Р.И. – М.: ДМК Пресс, 2014. – 462 с.

8. Шипунов, А. Б., Балдин, Е.М. Математическое моделирование/ А.Б. Щипунов, Е.М. Балдин. – СПб.: Профессия, 2011.– 360 с.

9 Алиферов, С.К. Основные понятия математического моделирования / С.К. Алиферов. – Новосибирск: Изд-во НГТУ, 2010. – 296 с.

10 Кудрявцев, А.А. Введение в математический анализ / А.А. Кудрявцев. – СПб.:СПбПУ, 2016. – 316 с.

11 Елисеева И.И. Общая теория математики в банковском деле / И.И. Елисеева. – М.: Финансы и статистика, 2013. – 656 с.

12 Ефимова М.Р. Практикум по математическому анализу в банковском деле / М.Р. Ефимова. – М.: Финансы и статистика, 2012. – 368 с.

13 Мелкумов Я.С.Принципы построения моделей в DEDUCTOR STUDIO. – М.: ИМПЭ-ПАБЛИШ, 2013. – 200 с.

14 Башина, О. Э. Общая теория использования математиеского моделирования в DEDUCTOR STUDIO / О.Э. Башина. – М.: Финансы и статистика, 2016. – 440 с.

15 Салин В.Н. Курс теории использования логистической регрессии в DEDUCTOR STUDIO / В.Н. Салин. – М.: Финансы и статистика, 2015. – 480 с.

16 Езекил, М. Методы анализа регресиий в DEDUCTOR STUDIO / М. Езекил. – М.: Статистика, 2011. – 393 с.

17 Шевчук, С.И. Байесовский метод в статистике: учебное пособие для вузов / С.И.Шевчук. – Новосибирск: Изд-во НГТУ, 2015. – 380 с.

18 Дрейнер,Н. Прикладной статистический анализ/ Н. Дрейнер – М.:Статистика, 2009. – 140 с.

19 Гладилин, А.В. Эконометрика / А.В. Гладилин. – М.:КНОРУС, 2012. – 368 с.

20 Дмитровский, В.В. Эконометрика / В.В. Дмитровский.– М.: Новый учебник, 2014. – 327 с.

21 Рязанов, И.Л.. Эконометрика.Введение в статистику / И.Л.Рязанов. – М.: КНОРУС, 2016. – 260 с.

22 Паклин, Н.Б. Оптимальное квантование для повышения качества бинарных классификаторов / Н.Б. Паклин, В.В. Афанасьев. – М.: ДМК Пресс, 2013. – 399 с.

23 Кацюба, О.А. Оценивание параметров многомерной линейной авторегрессии / О.А. Кацюба. – Москва-Ижевск, 2015. – 107 с.

24 Комарёк, В.И. Логистическая регрессия и её анализ / И.В. Комарёк. – СПб.:СПбПУ, 2016. – 246 с.

25 Томас, Х. Скоринговая карта. Построение и анализ / Х.Томас, И. Чарльз, Л. Рональд. – М.: ДМК Пресс, 2014. – 452 с.

26 Воунг, Л.С. Построение скоринговой карты/ Л.С. Воунг – М.: КНОРУС, 2015. – 210 с.

27 Лесовец, К.Б. Анализ, как способ повышения качества бинарных классификаторов / К.Б. Лесовец – М.: ДМК Пресс, 2014. – 219 с.

28 Гаррисон, О. Анализ скорингвой карты / О. Гаррисон. – Москва-Ижевск, 2016. – 207 с.

29 Димидин, К.Ф. Логистическая регрессия и её анализ /К.Ф.Димидин – СПб.:СПбПУ, 2014. – 346 с.

30 Вильямс, В. Способы анализа скоринговой карты/ В.Вильямс. – М.: ДМК Пресс, 2013. – 441 с.