

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ

Факультет

вычислительной техники

Кафедра

МО и ПЭВМ

Направление подготовки 09.04.02 Информационные системы и технологии
Магистерская программа Проектирование, разработка и эксплуатация
информационных систем

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему

**Оценка кредитоспособности физического лица на основе
интеллектуального анализа данных**

Студент

(подпись, дата)

Коркина Екатерина Игоревна

(ФИО полностью)

Руководитель

(подпись, дата)

Попова Наталия Александровна

(ФИО полностью)

Нормоконтролер

(подпись, дата)

Такташкин Денис Витальевич

(ФИО полностью)

Рецензент

(подпись, дата)

Афонин Александр Юрьевич

(ФИО полностью)

Работа допущена к защите (протокол заседания кафедры от _____ № _____)

Заведующий кафедрой

(подпись)

Макарычев П.П.

(фамилия, инициалы)

Работа защищена с отметкой _____ (протокол заседания ГЭК от _____ № _____)

Секретарь ГЭК

(подпись)

Попова Н.А.

(фамилия, инициалы)

Пенза, 2020

Список сокращений

ИАД – Интеллектуальный анализ данных

ЛДА – Линейный дискриминантный анализ

Содержание

Введение.....	5
1 Анализ методик и методов оценки кредитоспособности.....	8
1.1 Сущность кредитоспособности физических лиц.....	8
1.2 Оценка доступности кредита.....	8
1.3 Кредитоспособность клиентов.....	11
1.4 Модель кредитного скоринга.....	13
1.4.1 Критерии, используемые в моделях кредитного скоринга.....	15
1.5 Методы оценки кредитоспособности.....	16
1.6 Системы оценки кредитоспособности.....	20
Выводы.....	21
2 Поиск оптимальных решений для оценки кредитоспособности на основе методов ИАД.....	23
2.1 Обзор методов ИАД для решения задачи оценки кредитоспособ- ности физических лиц.....	23
2.2 Типы данных для анализа.....	25
2.3 Методы ИАД, используемые в кредитном скоринге.....	28
2.3.1 Линейный дискриминантный анализ.....	31
2.3.2 Логистическая регрессия.....	32
2.3.3 Деревья классификации и регрессии.....	34
2.3.4 Нейронные сети.....	36
2.4 Анализ программных средств для обработки и анализа данных.....	39
2.5 Выбор метода для оценки кредитоспособности.....	41
2.6 Этапы алгоритма оценки кредитоспособности.....	43
Выводы.....	45
3 Разработка алгоритма оценки кредитоспособности и проведе- ние экспериментов.....	47
3.1 Разработка алгоритма оценки кредитоспособности.....	47
3.2 Анкетирование потенциальных заемщиков.....	47
3.3 Описание исходного набора данных.....	51

3.4	Архитектура хранилища данных.....	54
3.5	Анализ данных в Deductor.....	55
3.5.1	Дерево решений.....	58
3.5.2	Нейронная сеть.....	60
3.5.3	Карты Кохонена.....	62
3.5.4	Логистическая регрессия.....	64
3.6	Работоспособность полученных моделей.....	67
3.7	Пример расчета кредитоспособности для потенциального заемщика...	69
	Выводы.....	71
	Заключение.....	73
	Список используемых источников.....	74
	Приложение А. Дипломы.....	80

Введение

В современном мире люди часто используют системы принятия решений, для решения разного рода задач. Актуальной на данный момент является задача оценки кредитоспособности физического лица.

Кредитный рейтинг (кредитоспособность) – это мнение определенной организации, выдающей кредиты относительно способности и готовности физического лица выполнять свои финансовые обязательства в полном объеме и в установленные сроки. Кредитный рейтинг также указывает на вероятность дефолта должника. По сути, кредитный рейтинг является мерой кредитоспособности частного лица (кредитный скоринг). Кредитор оценивает кредитоспособность, анализируя информацию, предоставленную заемщиком для получения кредита [1].

Современные системы не всегда позволяют объективно оценить финансовое состояние заемщика, основываясь только на количественном анализе данных. Так как методика количественной оценки позволяет установить уровень кредитоспособности физического лица, с помощью бальной системы, так образом рейтинг заёмщика определяется количеством баллов, набранных в процессе оформления заявки на кредит. Чем больше баллов, тем больше вероятность получения кредита.

Фирмы используют различные методы и процессы для оценки кредитного риска и доступности. Они могут быть автоматизированными, ручными или их комбинацией. Фирмы могут оценивать кредитный риск и доступность вместе или по отдельности. Все это может обеспечить благоприятный исход как для клиентов, так и для банков [2].

Потребительский кредит играет ключевую экономическую функцию и является в значительной степени выгодным видом деятельности. Большинство заемщиков погашают вовремя и без финансовых затруднений. Тем не менее, существуют особые риски в отношении кредитов и заемщиков. Для того, чтобы этого по возможности избежать, банкам необходимо учитывать эти

риски как в своей политике и процедурах, так и при принятии индивидуальных решений о кредитовании [3].

Большинство фирм имеют сильный коммерческий стимул для оценки кредитного риска, включая вероятность дефолта. В конечном итоге бизнес будет прибыльным только в том случае, если достаточное количество его клиентов погасит свои кредиты. Это также должно работать в интересах клиентов, получивших кредиты, с дальнейшей возможностью их погашения.

Цель диссертации состоит в проведении исследования различных методик оценки кредитоспособности физических лиц и разработке алгоритма для оценки кредитоспособности физических лиц с помощью методов интеллектуального анализа данных.

Задачи исследования. В соответствии с целью диссертации были выделены следующие задачи:

- проведение анализа банковских методик оценки кредитоспособности физических лиц;
- обзор методов интеллектуального анализа данных, используемых в кредитном скоринге;
- проведение исследования, с помощью методов интеллектуального анализа данных на базе реальных заемщиков банка;
- оценка эффективности полученных результатов исследования.

Объектом исследования являются методы интеллектуального анализа данных.

Предметом исследования являются алгоритмы методов интеллектуального анализа данных, используемые для оценки кредитоспособности физических лиц.

Научная новизна диссертационной работы заключается в следующем:

- 1) был разработан алгоритм, позволяющий сделать оценку кредитоспособности клиентов банка;

2) были предложены процедуры для решения задачи оценки кредитоспособности физических лиц с использованием Deductor.

Практическая значимость Данная методика оценки, на основе интеллектуального анализа данных позволит сделать более качественную и объективную оценку кредитоспособности физических лиц, что позволит банкам увеличить количество заемщиков, тем самым получить большую прибыль, а также расширить круг клиентов, которым представится возможность кредитования.

Результаты, выносимые на защиту:

- существующие методики оценки кредитоспособности потенциальных заемщиков;
- методы интеллектуального анализа данных, используемые в кредитном скоринге;
- алгоритм для оценки кредитоспособности физических лиц, полученный в результате исследования.

Публикации. По материалам диссертации были опубликованы две статьи.

Структура и объём работы. Диссертация состоит из введения, трех глав, заключения, списка использованных источников и приложений.

1 Анализ методик и методов оценки кредитоспособности

1.1 Сущность кредитоспособности физических лиц

Кредитование играет важную роль в обществе и является в значительной степени выгодным видом деятельности, позволяющим заемщикам приобретать товары и услуги и распределять выплаты с течением времени.

Для того, чтобы выдать кредит физическому лицу, необходимо определить его кредитоспособность.

Кредитоспособность включает в себя кредитный риск и доступность. Кредитный риск – это риск для кредитора, что клиент не будет погашать кредит, в то время как доступность заключается в том, насколько трудно клиенту погасить кредит. Поэтому кредитный риск может рассматриваться как «ориентированный на кредитора», а доступность - «ориентированная на заемщика». Эти две характеристики, соединяют вместе, для того чтобы дать конечный ответ – одобрить/ не одобрить кредит [4,5].

Банкам необходимо учитывать эти риски как в своей политике и процедурах, так и при принятии индивидуальных решений относительно своих клиентов.

Необходимо сделать разумную оценку не только того, будет ли клиент платить, но также свести к минимуму риску его финансовые затруднения, а также личные, социальные и экономические последствия. Но не стоит забывать, что в конечном итоге бизнес будет прибыльным только в том случае, если достаточное количество его клиентов погасит свои кредиты.

Организация, выдающая кредит должна придерживаться подхода, который учитывает стоимость и риски кредита для отдельного клиента и вероятность его погашения, для того, чтобы не было финансовых потерь.

1.2 Оценка доступности кредита

Существует значительное совпадение между кредитным риском и доступностью. Фирмы могут использовать одну и ту же информацию для их оценки как часть единой комплексной оценки. Тем не менее, это не всегда

учитывает потенциальную возможность того, что кредит может оказать негативное влияние на финансовое положение клиента и его благосостояние. Например, клиент не должен существенно сокращать расходы на еду или другие важные расходы или не выполнять другие обязательства.

Поэтому необходимо учитывать, как показатели доступности кредитов, так и кредитоспособность клиентов. Рисунок 1 наглядно демонстрирует это.

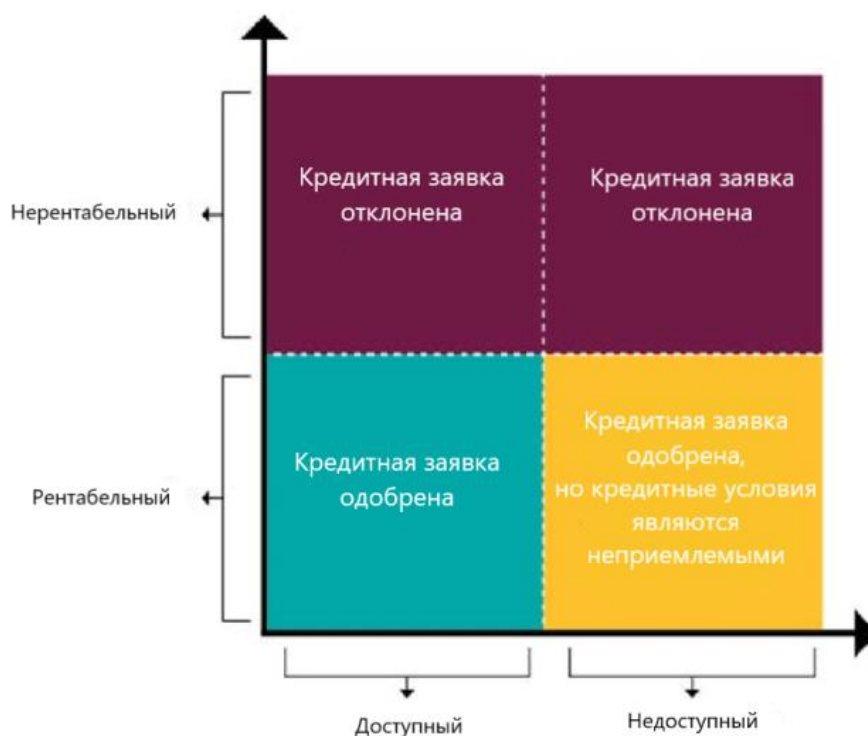


Рисунок 1 – Зависимость доступности кредита и кредитоспособности клиента

Поскольку банки заинтересованы получать только прибыль от кредитов, нерентабельным заемщикам однозначно будет отказано в выдаче кредитов, как это можно увидеть в верхних квадратах рисунка 1.

Отсутствие необходимости в том, чтобы фирмы оценивали доступность, могут побуждать фирмы предлагать недоступные кредиты тем заемщикам, которые, как они ожидают, будут прибыльными, даже если они подвергаются высокому риску финансовых затруднений. Это группа, указанная в правом нижнем квадрате рисунка 1. Степень такого несоответствия между банком и заемщиками на кредит будет зависеть от ситуации клиента и условий кредита.

Идеальным является совпадение показателей доступности и кредитоспособности, для того чтобы банк одобрил кредит заемщику, без ухудшения финансового положения последнего. Это показано на рисунке 1 в нижнем левом квадрате.

Таким образом, необходимо проводить комплексную оценку для того, чтобы улучшить уровень оценки кредитоспособности, а также расширить доступ к кредитным продуктам банка большему числу заемщиков. Тем самым необходимо минимизировать риски как по отношению к банкам, так и заемщикам [6].

Потребительский кредит выполняет важную экономическую функцию и в значительной степени является выгодным видом деятельности. Это позволяет заемщикам приобретать товары и услуги и распределять выплаты по времени, управлять денежными средствами и планировать будущее. Он также имеет более широкие экономические преимущества и может способствовать росту всей экономики.

Однако некоторые заемщики могут испытывать финансовые и другие трудности с получением кредита. Например, если у клиента недостаточный доход или слишком юный возраст, а также это может произойти, если клиент не сможет полностью или частично погасить кредит, или имеет плохую кредитную историю.

В дальнейшем у заемщика могут возникнуть финансовые трудности, которые могут произойти по ряду причин. В некоторых случаях это может быть связано с изменением обстоятельств или другими вопросами, которые кредитор или заемщик не могли разумно предвидеть. В других это может быть частично или полностью из-за безответственного кредитования. В частности, фирмы могут не проявлять достаточной осторожности при принятии решений о кредитовании или могут не относиться к клиентам справедливо.

Именно для этого необходимо провести предварительно качественную и объективную оценку заемщика.

Как показывает практика, что люди, испытывающие финансовые затруднения, как правило, моложе, с более низкими доходами, с меньшей вероятностью трудоустроиться. Они также более склонны к долгосрочным кредитным выплатам. Это указывает на то, что оценки доступности могут оказать относительно большее влияние на снижение рисков для такого типа клиентов [7].

Одной из целей регулирования потребительского кредитования является минимизация предсказуемых финансовых затруднений путем поощрения кредитования и обеспечения ответственного заимствования. Это также должно способствовать устойчивому функционированию кредитных рынков. На высоком уровне рынок потребительского кредитования можно считать устойчивым, если он является преимущественным для большинства заемщиков и, если банки могут эффективно покрывать убытки от дефолтов и оставаться прибыльными. Выполняя комплексную оценку, необходимо придерживаться нескольких целей, а именно: обеспечить грамотную оценку кредитоспособности заемщика, с выдачей кредита на условиях, которые не вызовут финансовые трудности, а также поддержание банковской системы.

Поскольку большинство клиентов берут кредиты без затруднений, необходимо уделить особое внимание тем ситуациям, в которых интересы клиентов и банков не соответствуют друг другу, или там, где существуют определенные риски, связанные с конкретными бизнес-моделями. То есть, в спорных вопросах, относительно выдачи кредита, необходимо прибегать к дополнительным проверкам и оценкам возможностей заемщика.

1.3 Кредитоспособность клиентов

Кредитоспособность – это один из ключевых элементов обеспечения стабильно функционирующего рынка потребительских кредитов. Эффективная оценка кредитоспособности помогает избежать вреда, отказывая в кредитовании тем, кто наиболее подвержен финансовым проблемам. А также правильно применяемая политика также должна означать, что потребители

могут получить кредит при соответствующих обстоятельствах. Для этого нужно минимизировать риск предсказуемых финансовых проблем и в то же время избежать чрезмерного ограничения доступа к кредитам [8,9].

В идеальном мире клиентам будет предоставлен кредит только в том случае, если они смогут позволить себе его полное погашение - без неоправданных трудностей, при выполнении других разумных обязательств и без какого-либо ущерба для их финансового состояния.

На практике это никогда не может быть гарантировано. Оценка доступности не является точной наукой. Кредит может быть доступным для одного заемщика, но недоступным для другого в аналогичной ситуации. Не все это можно предсказать заранее, и затраты на это могут быть непропорциональными или могут неоправданно ограничивать доступность кредита для тех, кто может его себе позволить.

Много лет назад большинство финансовых учреждений полагалось почти исключительно на субъективный анализ или так называемые банковские экспертные системы для оценки кредитного риска по корпоративным кредитам. По сути, банкиры использовали информацию о различных характеристиках заемщика, чтобы прийти к субъективному суждению относительно того, предоставлять ли кредит или нет. Из-за того, что банковские организации, как правило, чрезмерно пессимистично настроены, финансовые институты перешли к системам, которые более объективны. Среди них: однофакторные системы кредитного скоринга, основанные на учете, где лицо, принимающее решения в финансовом учреждении, сравнивает различные ключевые учетные коэффициенты потенциальных заемщиков с отраслевыми или групповыми нормами, и многовариантные модели, в которых ключевые переменные учета объединяются и взвешиваются для получения либо оценки кредитного риска или вероятности дефолта. Существует как минимум четыре методологических подхода к разработке многомерных систем оценки кредитоспособности: линейная

модель вероятности, модель логита, модель пробита и модель дискриминантного анализа.

Кредитный скоринг можно формально определить как статистический (или количественный) метод, который используется для прогнозирования вероятности того, что заемщик не выполнит своих обязательств [10].

Кредитные оценки помогают уменьшить дискриминацию, потому что модели кредитного скоринга обеспечивают объективный анализ кредитоспособности потребителя. Это позволяет банкам сосредоточиться только на информации, касающейся кредитного риска, и избежать личной субъективности кредитного аналитика. Они также помогают повысить скорость и последовательность процесса подачи заявки на кредит и позволяют автоматизировать процесс кредитования. Таким образом, это значительно снижает необходимость вмешательства человека в оценку кредита и стоимость предоставления кредита. С помощью кредитных баллов финансовые учреждения могут количественно оценить риски, связанные с предоставлением кредита конкретному заявителю в более короткие сроки. Время, сэкономленное на обработке кредитов, можно использовать для решения более сложных вопросов.

1.4 Модель кредитного скоринга

Розничное кредитование играет важную роль в обществе и является в значительной степени выгодным видом деятельности, позволяющим заемщикам приобретать товары и услуги и распределять выплаты с течением времени.

Банки используют различные методы и процессы для оценки кредитного риска и доступности. Они заинтересованы в том, чтобы проводить разумную оценку не только того, будет ли клиент платить, но также и его способности платить по приемлемым ценам, и это не окажет существенного влияния на его финансовое положение. Это сводит к минимуму риск финансовых затруднений для клиентов и его личные, социальные и экономические

последствия. В то же время необходимо избегать чрезмерных предписаний, поскольку это может иметь вредные непреднамеренные последствия, в том числе для стоимости и доступности кредита. Банкам необходимо придерживаться пропорционального подхода, который учитывает стоимость и риски кредита для отдельного клиента и вероятность того, что они могут пострадать в результате кредита.

Большинство заемщиков погашают вовремя и без финансовых затруднений. Тем не менее, существуют особые риски в отношении кредитов. Банкам следует учитывать эти риски как в своей политике и процедурах, так и при принятии индивидуальных решений о кредитовании. Большинство имеют сильный коммерческий стимул для оценки кредитного риска, включая вероятность дефолта. В конечном итоге бизнес будет прибыльным только в том случае, если достаточное количество его клиентов погасит свои кредиты. В целом это также должно работать в интересах клиентов, предотвращая их от долга, который они не смогут погасить. Поэтому банки используют как кредитный риск, так и доступность для единой комплексной оценки. Однако не всегда возможно в достаточной степени учитывать потенциальный риск как для заемщика, так и для банка. Именно для этого необходимо прибегать к различным методам анализа кредитоспособности.

Прежде чем одобрять кредит физическим лицам, следует изучить их финансовое положение. На основе финансового положения своих заявителей, запрашивающих кредит, банки присваивают кредитный рейтинг (баллы), и опираясь на кредитные баллы банк решает, предлагать ли кредит этим заявителям, а также определяет кредитные лимиты.

«Кредитный рейтинг» является независимой переменной. Кредитный рейтинг – это число, которое обозначает кредитоспособность заявителей. Чем выше кредитный рейтинг, тем выше кредитоспособность заявителя, а чем ниже кредитный рейтинг, тем ниже кредитоспособность заявителя.

1.4.1 Критерии, используемые в моделях кредитного скоринга

Чтобы ответить на вопрос о том, каковы основные детерминанты дефолта, необходимо разработать спецификацию модели, содержащую правильные критерии. Критерии можно разделить на четыре основные категории, как показано в таблице 1.

Таблица 1 – Критерии, используемые для оценки кредитоспособности

Демографические критерии	Финансовые критерии	Критерии занятости	Потребительские критерии
1. Возраст заемщика	1. Общая сумма активов заемщика	1. Тип занятости	1. Проверка наличия кредитной истории
2. Пол заемщика	2. Общий доход заемщика	2. Общий стаж на текущей работе	2. Наличие текущих кредитов
3. Семейное положение заемщика	3. Общий доход семьи	3. Общий стаж на предыдущем месте работы	3. Наличие непогашенных займов
4. Количество иждивенцев	4. Ежемесячные расходы		4. Срок кредита
5. Наличие собственной недвижимости			
6. Адрес проживания			

Первая категория содержит демографические показатели. Эти переменные, как правило, не имеют наибольшего значения, но они полезны при выявлении различных региональных, гендерных и других соответствующих различий. Например, часто обнаруживается, что пожилые люди менее рискованны, чем молодые люди. В целом, риск дефолта уменьшается с возрастом, а также ниже для лиц, состоящих в браке. Заемщики с недвижимостью также представляют менее рискованную категорию из-за дома в качестве залога. Подобные связи могут помочь лучше различать хороших и плохих кандидатов.

Вторая категория содержит данные о финансовом положении. При рассмотрении заявления на получение кредита, банку необходимо знать, какие другие ресурсы имеются у заемщика, каковы его доходы и расходы, и, следовательно, исходя из этих элементов информации, каков реалистичный или потенциальный максимально возможный ежемесячный платеж. Важность этих переменных очевидна.

Источник дохода и статус занятости относятся к третьей категории критериев. Как правило, люди, работающие не по найму, часто получают более низкий балл при оценке заявок на получение ссуды, чем работающие на организации люди. Это связано с тем, что стабильность занятости может служить признаком стабильности платежей. Характер и продолжительность занятости также являются решающими факторами: частая смена рабочих мест предполагает низкий балл.

Критерии четвертой категории являются необходимой информацией, которую можно использовать для кредитного скоринга. Этот тип информации значительно снижает проблему асимметричной информации между банком и клиентом. Если у клиента есть какая-либо история с банком, то этот банк может легко проверить, например, историю кредитов. Банк знает, был ли у клиента уже кредит, был ли этот кредит успешно погашен, или возникли проблемы. Банки часто делятся этим типом информации, поскольку предыдущий дефолт / просрочка является серьезным фактором, определяющим будущие проблемы с погашением долгов.

1.5 Методы оценки кредитоспособности

Современные методы оценки заемщика позволяют установить уровень кредитоспособности физического лица, с помощью бальной системы, таким образом рейтинг заёмщика определяется количеством баллов, набранных в процессе оформления заявки на кредит. Чем больше баллов, тем больше вероятность получения кредита [11].

Как правило, каждый банк имеет свою систему оценки кредитоспособности заемщика. Именно с ее помощью банк может определить платёжеспособность клиента, чтобы избежать риска невозврата средств.

Одними из ключевых составляющих кредитоспособности являются:

- текущая платёжеспособность;
- дисциплинированную оплату платежей;
- потенциальную возможность погашения своего долга.

Рассматривая физическое лицо, главными факторами для оценки его кредитоспособности является: уровень заработной платы, наличие движимого и недвижимого имущества, социальный статус, страховой стаж, уровень образования, семейное положение и др.

Основной метод кредитной оценки – анкетирование. Клиенту задается ряд вопросов, за ответ на каждый из которых он получает определенное количество баллов. После опроса, баллы суммируются и на основании полученных результатов, выдаётся решение о выдаче или не выдаче потенциальному заемщику кредита.

Данные, которые рассматриваются при анкетировании:

- личные (паспортные данные, семейное положение, пол, возраст, наличие детей, их возраст и количество);
- финансовые (доход, место работы, должность, период работы на предыдущем месте, наличие недвижимости, автомобиля);
- дополнительные (сведения о дополнительном доходе, возможность предоставления поручителя, и др.).

Современные системы не всегда позволяют объективно оценить финансовое состояние заемщика, основываясь только на количественном анализе данных.

Для более качественной оценки кредитоспособности физического лица следует использовать метод анализа, который включает в себя как количественные, так и качественные критерии оценки.

Рассмотрим более подробно существующие методы оценки кредитоспособности заемщиков. На рисунке 2 представлены категории методов, используемых в кредитном скоринге.

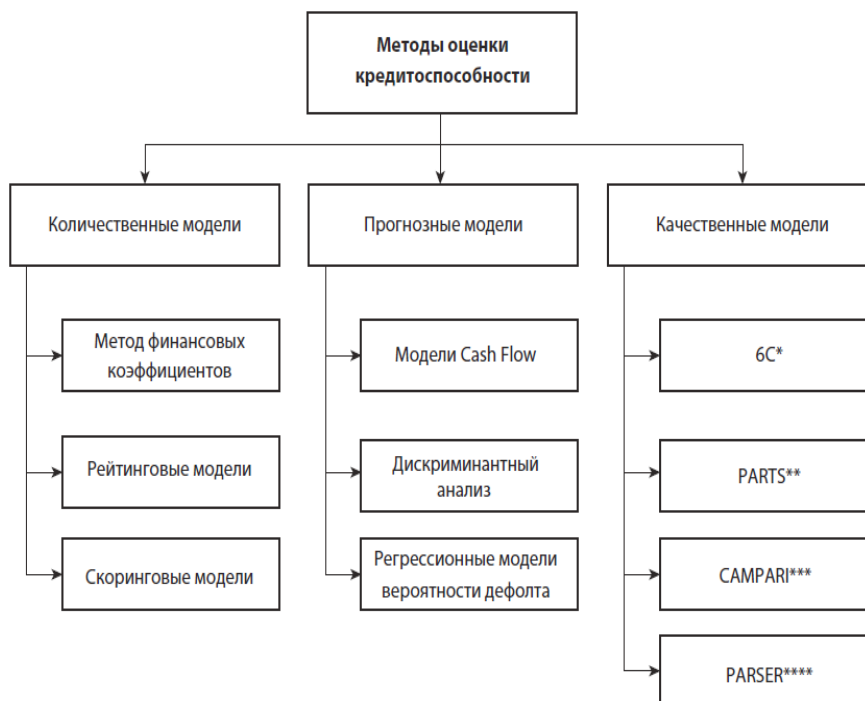


Рисунок 2 – Методы оценки кредитоспособности

В количественных моделях используется определенный перечень финансовых и нефинансовых показателей, которые с их помощью позволяют присвоить заемщику определенную рейтинговую оценку. Что касается прогнозных моделей, то они используют статистику прошлых лет и направлены на моделирование динамики финансового состояния компании и оценку вероятности банкротства (дефолта) клиента. Качественные модели опираются только на определенную систему различных качественных показателей.

Проведем сравнительный анализ данных методов и определим их преимущества и недостатки. Результаты сравнения можно увидеть на рисунке 3.

Наименование метода	Достоинства	Недостатки
Метод коэффициентов	Позволяет комплексно оценить финансовое положение заемщика.	Не учитывает качественные показатели, статистику прошлых лет. Не автоматизированный метод, который требует постоянной интерпретации значений отдельных показателей.
Рейтинговые модели	Позволяют автоматизировать оценку методом коэффициентов путем вычисления интегрального показателя. Отличаются удобством и простотой использования.	Учитывают только финансовые показатели, не используют статистику прошлых лет. Требует перестройки для различных типов компаний.
Скоринговые модели	Позволяют получить оценку кредитоспособности в бальном эквиваленте и отнести клиента к группе плохих или хороших заемщиков. Просты и удобны в использовании, помогают оценивать нефинансовые (качественные) показатели. При оценке весовых коэффициентов статистическими методами позволяют учесть данные по уже выданным кредитам.	Не универсальны, требуют перестройки под определенные типы компаний. Требуют большого объема данных для исследования весовых коэффициентов.
Модели Cash Flow	Позволяют оценить будущие денежные потоки и сопоставить с долговой нагрузкой заемщика.	Не учитывают качественные показатели заемщика, могут давать несовместимые с реальностью результаты.
Модели дискриминантного анализа	Позволяют определить вероятность дефолта заемщика на основе статистики за прошлые годы.	Результаты сильно зависят от обучающей выборки, при исследовании модели на других данных могут выдать результаты не соответствующие реальности. Требуют большого объема данных для исследования весовых коэффициентов.
Модели регрессионного анализа	При использовании модели на корректных данных можно получить результаты высокой степени достоверности.	Результаты сильно зависят от обучающей выборки, чувствительность к пропущенным значениям. Требуют большого объема данных для исследования весовых коэффициентов.
Модели качественного анализа	Позволяют провести комплексный анализ заемщика.	Отсутствие математических методов приводит к субъективной оценке отдельных показателей и к ошибкам. Нет четких регламентов оценки кредитоспособности.

Рисунок 3 – Достоинства и недостатки методов оценки кредитоспособности

Рассмотрев некоторые из действующих методик для оценки кредитоспособности клиентов, можно сделать вывод, что все они имеют свои плюсы и минусы. Поэтому необходимо разработать методику оценки, которая позволит проводить объективную оценку, на возможных наборах данных, учитывая, как качественные, так и количественные показатели, выдавая при этом результаты соответствующие реальности.

Ввиду того, что все вышеописанные методы имеют свои ограничения и недостатки и не являются достаточно эффективными, нами предлагается новый метод оценки кредитоспособности заёмщика, который заключается в использовании средств DataMining.

1.6 Системы оценки кредитоспособности

Скоринговая система банка – это механизм отбора потенциальных заемщиков путем экспертной проверки факторов, влияющих на платежеспособность и риск невозврата полученных займы денежных средств. В настоящее время существует много систем, которые позволяют оценить кредитоспособность заемщика. Наиболее известными являются - SAS Credit Scoring for Banking и FICO Scores.

Система Credit Scoring for Banking для оценки кредитных рисков включает набор методик и инструментов, которые позволяют предсказать поведенческую модель клиента, определить вероятность невыплаты кредита, ухода заемщика в дефолт.

Инструментарий SAS Credit Scoring for Banking позволяет создавать модели кредитного скоринга для потребительских кредитов, кредитных карт, овердрафтов, автомобильных, ипотечных и других кредитных продуктов. Скоринг направлен на решение различных задач от оценки вероятности дефолта клиента до определения стратегии работы коллекторского подразделения и создания рейтинговой системы.

SAS Credit Scoring анализирует данные потенциального заемщика и выдает готовые рекомендации по выдаче кредита с учетом возможных рисков. Данная система оценки кредитных рисков обеспечивает сбор, преобразование, стандартизацию и очистку всех релевантных данных для создания полного профиля клиента с использованием внутренних данных компании, а также с возможностью привлечения внешней информации (например, данные налоговых служб).

Система FICO это система кредитной оценки, созданной Fair Isaac Corporation. Кредиторы используют оценки FICO заемщиков вместе с другими данными кредитных отчетов заемщиков для оценки кредитного риска и определения необходимости предоставления кредита. Баллы FICO учитывают различные факторы в пяти областях для определения кредитоспособности: история платежей, текущий уровень задолженности, типы используемых

кредитов, продолжительность кредитной истории и новые кредитные счета. На базе данной системы создаются потребительские кредитные рейтинги, которые финансовые учреждения используют при принятии решения о предоставлении кредита или выдаче кредита.

FICO является системой количественной оценки кредитоспособности и оценивает рейтинг заемщиков в баллах. Баллы варьируются от 300 до 850. В целом, баллы выше 650 указывают на хорошую кредитоспособность. Напротив, лицам с оценками ниже 620 часто трудно получить финансирование по выгодным ставкам. Чтобы определить кредитоспособность, кредиторы принимают во внимание балл заемщика, но также учитывают и другие детали, такие как доход, продолжительность работы заемщика и тип запрашиваемого кредита.

FICO помогает кредиторам принимать точные, надежные и быстрые решения по кредитному риску на протяжении всего жизненного цикла клиента. Оценка кредитного риска определяет для потребителей то, насколько вероятно, что они выполнят свои кредитные обязательства в соответствии с договоренностью.

Выводы

Таким образом кредитование населения по-прежнему составляет основу доходов коммерческих банков и других кредитных учреждений. С технической точки зрения процесс кредитования в целом представляет собой относительно простую серию действий с участием двух основных сторон. Эти действия варьируются от первоначальной заявки на кредит до успешного или неудачного погашения кредита. Хотя потребительское кредитование относится к числу наиболее прибыльных инвестиций в портфели активов кредиторов, увеличение объемов займов также приводит к увеличению числа просроченных займов, то есть займов, которые либо вообще не погашаются, либо заемщик имеет проблемы с выплатой долгов. Таким образом, основная проблема любого кредитора заключается в разграничении между «хорошими»

и «плохими» заемщиками до предоставления кредита. Такая дифференциация возможна с помощью метода кредитного скоринга. Поэтому необходимо провести анализ методов кредитного скоринга и разработать наиболее эффективную методику оценки кредитоспособности физических лиц.

В последние годы российский рынок розничного кредитования развивается впечатляющими темпами. Объем выданных кредитов ежегодно продолжает расти, растет также количество работающих на этом рынке банков и предлагаемых ими кредитных продуктов. В условиях острой конкуренции важнейшим фактором успешной работы банка является возможность максимально точно оценить кредитоспособность заемщиков. Банк, способный наиболее точно «отбраковывать» нежелательных клиентов, сможет предложить рынку более привлекательные кредитные продукты с меньшими кредитными ставками и, следовательно, получить определенные конкурентные преимущества. В таком случае на помощь приходит узкоспециализированная эффективная система принятия решений – кредитный скоринг.

Актуальность задачи по совершенствованию существующих методик кредитоспособности подтвердил последний экономический кризис, который выявил неэффективность многих методик, применяемых коммерческими банками. Это выразилось, в частности, в том, что многие из заемщиков, которым был присвоен высокий рейтинг, тем не менее, оказались проблемными. Это заставляет задуматься о грамотной интеграции оценки кредитоспособности в кредитный процесс банка, в том числе и с точки зрения автоматизации, поскольку на сегодняшний день информационные технологии являются неотъемлемой частью бизнес-процесса, поэтому к информационным технологиям банковской сферы в текущих реалиях выставляются повышенные требования.

2 Поиск оптимальных решений для оценки кредитоспособности на основе методов ИАД

2.1 Обзор методов ИАД для решения задачи оценки кредитоспособности физических лиц

На данный момент в современном мире существует огромное количество программных продуктов, которые значительно упрощают работу предприятий. Каждая организация стремится обзавестись новыми технологиями, для того чтобы значительно увеличить производительность с помощью комплекса средств автоматизации, который позволит находить решение на основе автоматизации технических процессов и будет реализовывать выполнение определенных задач.

В то время как некоторые отказываются от идеи доверять сложным компьютерным программным решениям для принятия решений за них, большинству удобнее использовать компьютерную статистику для понимания ключевых тенденций. К ним относятся такие аналитические данные, как статистика продаж, процентные ставки по гарантиям и тенденции движения денежных средств, которые являются важными показателями, помогающими руководителям определить состояние их бизнеса и обуславливающими необходимость корректирующих действий.

Сложность состоит в том, что этот уровень информации не может определить, какая из нескольких возможностей максимизирует отдачу при достижении желаемого результата. Он также не может предвидеть внешние изменения, которые могут повлиять на рентабельность, что является важным фактором, поскольку большинство компаний работают в неопределенной среде, регулируемой настроениями потребителей, правовыми нормами и интенсивной конкуренцией. Кроме того, компании подвержены внешним воздействиям, таким как политическая неопределенность, глобальные погодные явления и торговые споры.

Эти факторы иногда объединяются, чтобы создать идеальные условия, в которых принятие решений затрудняется отсутствием предсказуемости, а также неспособностью обрабатывать данные достаточно быстро для поддержки решений. Вот почему актуальны методики, основанные на интеллектуальном анализе данных, которые могут быстро анализировать данные, определять шаблоны и оценивать несколько альтернатив, оказываются неоценимыми для лидеров бизнеса.

ИАД предназначен для помощи в принятии решений путем доступа к большим объемам информации, генерируемой из данных, участвующих в организационных бизнес-процессах [12].

ИАД использует сводную информацию, исключения, шаблоны и тенденции с использованием аналитических моделей. Лица, принимающие решения, собирают полезную информацию из необработанных данных, документов, личных знаний и / или бизнес-моделей для выявления и решения проблем и принятия решений.

Выбор наиболее подходящего метода ИАД зависит от сложности и в определенной степени, от структуры данных. Без сомнения, наибольшая выгода заключается в выборе алгоритмов ИАД, основанных на предсказательной аналитике, которая моделирует бизнес и дает возможность определять наиболее выгодное решение на основе определенных критериев, таких как доход и прибыльность. Хотя такое решение влечет за собой большие инвестиции в ресурсы, однако оно с большей вероятностью может превзойти ожидания и повысить рентабельность инвестиций. Кроме того, модель копирует бизнес, поэтому эти методы с большей вероятностью предложат выполнимые и рациональные решения.

Наиболее эффективными примерами ИАД являются те, которые определяют лучшее решение на основе определенных критериев. Такие системы устраняют субъективность и предвзятость в процессе принятия решений. Кроме того, они могут оценить множество альтернативных сценариев и определить лучшие.

Обычный подход начинается с разработки математической модели бизнеса. Далее необходимо посмотреть, как он принимает решения, и использовать программное обеспечение для оптимизации, чтобы определить результаты различных сценариев. Этот метод основан на предписывающей аналитике и является чрезвычайно мощным. Хотя некоторые полагают, что моделировать следует только процесс принятия решений, разработка полной модели организации повышает универсальность и повышает точность с точки зрения финансовых результатов.

Существует два подхода к оптимизации: правила и модели оптимизации. Подход, основанный на правилах (эвристический) работает хорошо, когда возможные результаты могут быть в значительной степени предопределены, например, при оценке страхового риска. С другой стороны, модели оптимизации являются более адаптируемыми, могут обрабатывать более сложные проблемы и иметь дело с множеством ограничений и компромиссов.

Таким образом, ИАД позволяет анализировать данные для решения определенных задач. Основной задачей ИАД является предоставление возможности эффективного анализа данных. Насколько анализ будет качественным зависит от некоторых факторов – квалификации экспертов, скорости обработки, объема данных, а также от формы представления данных. Рассмотрим один из этих факторов – представление данных для дальнейшего анализа. Существует три типа данных, которые можно использовать в СППР: структурированные данные, слабоструктурированные данные и неструктурированные данные.

2.2 Типы данных для анализа

Структурированные данные – это данные, отражающие отдельные факты предметной области и упорядоченные определенным образом с целью применения дальнейшего эффективного анализа. Они организованы в виде отформатированного хранилища, которое обычно является базой данных. Как правило представлены в SQL, таблицей со строками и столбцами. Они имеют

реляционные ключи и могут быть легко отображены в предварительно разработанные поля. Сегодня эти данные в приорите при разработке и являются наиболее простым способом управления информацией. Большинство алгоритмов машинного обучения, статистического и интеллектуального анализа данных работают только со структурированными данными.

Неструктурированные данные – это данные, которые не организованы заранее определенным образом или не имеют заранее определенной модели данных, поэтому они не подходят для основной реляционной базы данных.

Слабоструктурированные данные – это данные, который не находятся в реляционной базе данных, но имеют некоторые организационные свойства, облегчающие анализ. Вы можете сохранить их в базе данных отношений, однако, это может быть очень сложно для некоторых видов слабоструктурированных данных. Эти данные не соответствуют модели данных, но имеют некоторую структуру, им не хватает фиксированной или жесткой схемы [13].

Неструктурированные данные не организованы в формате, который облегчает доступ и обработку. Структурированные данные в основном противоположны неструктурированным: они отформатированы и организованы в структуру данных, так что к элементам можно обращаться и получать доступ к ним различными способами, чтобы лучше использовать информацию. Слабоструктурированные данные находятся где-то между двумя крайностями. Они не организованы сложным образом, из-за чего доступ к ним становится более сложным и анализ провести труднее, но добавляется такая информация, как тегирование метаданных, которая позволяет обращаться к элементам данных.

Слабоструктурированные данные составляют подавляющее большинство всех данных. Просто посмотрите на огромное количество видеофайлов, аудиофайлов и сообщений в социальных сетях, добавляемых каждую минуту, и вы поймете, почему возник термин «большие данные».

Такие данные представляют больше половины всех данных. И в дальнейшем будут расти только после того, как машинное обучение, искусственный интеллект и интернет вещей возрастут на рынке. Это приведет к огромному количеству систем, которые будут заполняться данными каждую секунду. Например, ожидается, что в ближайшие пять лет количество таких систем достигнет десятков миллиардов. Они будут генерировать много слабоструктурированных данных, которые необходимо будет обрабатывать. Слабоструктурированные данные являются более сложными и трудными для работы, даже не столько потому что они содержатся в больших объемах, а сколько из-за сложности структурирования этих данных. Однако на данный момент существуют одни из новейших технологий, которые работают с такими данными, такие как Hadoop, NoSQL или MongoDB. Данные программы ослабляют обычные требования к модели данных и позволяют хранить данные в гораздо более неструктурированном формате, чем, например, сбор данных в наборе данных SAS или реляционной базе данных Oracle. Но количество данных растет и будет продолжать расти, и сейчас потоки слабоструктурированных данных уже демонстрируют спутниковые снимки, цифровую микроскопию, исследования гидролокатора, каналы Twitter, публикации в Facebook на YouTube и так далее.

Почтовые сообщения являются хорошим примером слабоструктурированных данных. Хотя фактический контент не структурирован, он содержит структурированные данные, такие как имя и адрес электронной почты отправителя и получателя, время отправки и т.д. Другим примером является цифровая фотография. Само изображение не структурировано, но, если фотография была сделана, например, на смартфоне, на нем были бы отметки даты и времени, географическая пометка и идентификатор устройства. После сохранения фотографии также могут быть присвоены теги, которые будут обеспечивать структуру, например «собака» или «домашнее животное».

Такие данные используются в разных областях, таких как бизнес и менеджмент (прайсинг, рабочая сила, продукты, стратегия и т.п.), инжиниринг (дизайн продукта, контроль качества), финансы (кредитование и займы), медицина (лекарства, виды лечения, диагностика) и многое другое.

С увеличением количества систем, в которых присутствуют слабоструктурированные данные, появилась необходимость в развитии и улучшении моделей извлечения информации, путем поиска новых мер подобия и новых подходов в работе со слабоструктурированными данными. Это позволит принимать во внимание содержание и структуру слабоструктурированных данных для более удобного анализа и повышения точности результатов.

Вычислительная техника создавалась прежде всего для обработки данных. В настоящее время современные вычислительные системы и компьютерные сети позволяют накапливать большие массивы данных для решения задач обработки и анализа. К сожалению, сама по себе машинная форма представления данных содержит информацию, необходимую человеку, в скрытом виде, и для ее извлечения нужно использовать специальные методы анализа данных.

2.3 Методы ИАД, используемые в кредитном скоринге

Процесс кредитного скоринга очень важен для банков, поскольку им необходимо отделить «хороших заемщиков» от «плохих заемщиков» с точки зрения их кредитоспособности. Кредитный скоринг в его автоматизированной форме часто является единственным способом оценки кредитоспособности, поскольку у банков нет достаточных ресурсов для индивидуальной обработки каждого небольшого риска. Методы, обычно используемые для кредитного скоринга, основаны на статистических методах.

На данный момент существует большое количество методов анализа данных, которые работают на большом массиве данных. В зависимости от задач, которые необходимо решить, можно определить необходимый вид

анализа данных. Наиболее популярным для больших данных является интеллектуальный анализ данных. Интеллектуальный анализ данных (ИАД) – это исследование и анализ больших данных для выявления значимых закономерностей и правил [14].

Эти методы помогают предсказывать будущее и затем принимать соответствующие решения. Они также помогают анализировать рыночные тенденции и увеличивать доходы компании.

Важной концепцией, относящейся к кредитному скорингу, является концепция понесенных расходов, связанных с вероятностью погашения кредитов. Для простоты предположим, что совокупность ссуд состоит из двух групп или классов G и B , которые обозначают ссуды, которые (после предоставления) станут хорошими и плохими в будущем, соответственно. Хорошие кредиты погашаются полностью и в срок. Плохие кредиты подвержены различным степеням дефолта.

Обычно размеры классов очень разные, так что для вероятности того, что случайно выбранный кандидат (клиент) принадлежит группе G , обозначаемой как p_G , мы имеем $p_G > p_B$. Пусть x – вектор независимых переменных (называемый вектором измерения) используется в процессе принятия решения о том, принадлежит ли заявитель группе G или B . Пусть вероятность того, что заявитель с вектором измерения x принадлежит группе G , равна $p(G|x)$, а вероятность B равна $p(B|x)$. Пусть вероятность $p(x|G)$ указывает, что у хорошего кандидата есть вектор измерения x . Точно так же для плохого заявителя вероятность равна $p(x|B)$. Задача состоит в том, чтобы оценить вероятности $p(.|X)$ из набора данных о кандидатах, которые оказываются хорошими или плохими, и найти правило для того, как разделить пространство X всех векторов измерений на две группы A_G и A_B на основе этих вероятностей, чтобы что в A_G будут векторы измерений претендентов, которые окажутся хорошими, и наоборот. Однако, как правило, невозможно найти идеальную классификацию, так как может случиться так, что один и тот же вектор задается двумя заявителями, где один хороший, а другой плохой.

Поэтому необходимо найти правило, которое минимизирует расходы агентства, предоставляющего кредит, связанные с неправильной классификацией заявителей. Давайте обозначим C_G как расходы, связанные с ошибочной классификацией хорошего заявителя, и C_B как затраты, связанные с классификацией плохого заявителя как хорошего. Обычно $C_B > C_G$, потому что затраты, понесенные из-за неправильной классификации плохого клиента, являются финансово более разрушительными, чем затраты, связанные с первым типом ошибки. Если заявители x отнесены к классу G , ожидаемые затраты составляют $C_B p(B|x)$, а ожидаемые потери для всей выборки – $C_B \sum_{x \in A_G} p(B|x)p(x) + C_G \sum_{x \in A_B} p(G|x)p(x)$, где $p(x)$ - вероятность того, что вектор измерения равен x . Это сводится к минимуму, когда в группу G назначаются такие заявители, у которых есть группа векторов измерений:

$$A_G = \{x | C_B p(B|x) \leq C_G p(G|x)\},$$

что равнозначно:

$$A_G = \left\{x | p(G|x) \geq \frac{C_B}{C_B + C_G}\right\}.$$

Без потери общности мы можем нормировать затраты на неправильную классификацию $C_B + C_G = 1$. В этом случае правило классификации состоит в том, чтобы назначить кандидата с x для класса G , если $p(G|x) > C_B$, а в противном случае - для класса B .

Разумеется, важной задачей является определение стоимости ошибок кредитования для розничных кредитов и, при этом, более точное определение оптимального подхода к балльной оценке кредита. Именно банк должен выбрать оптимальный компромисс между прибыльностью и риском. Слишком строгая кредитная политика может обеспечить минимальные затраты с точки зрения просроченных ссуд, но альтернативные затраты на отклоненные ссуды могут превышать потенциальные затраты по сомнительным долгам, и, следовательно, прибыль не максимизируется. И наоборот, слишком либеральная политика может привести к большим убыткам от безнадежных долгов.

2.3.1 Линейный дискриминантный анализ

Целью линейного дискриминантного анализа (ЛДА) является классификация неоднородного множества на однородные подмножества и дальнейшее принятие решений по этим подмножествам. Можно предположить, что для каждого заёмщика имеется определенное количество доступных объясняющих переменных. Идея состоит в том, чтобы искать такую линейную комбинацию переменных, которая отделяет большинство подмножеств друг от друга. В простом случае цель заключается в том, чтобы из двух подмножеств найти линейную комбинацию объясняющих переменных, которая оставляет максимальное расстояние между средними значениями двух подмножеств.

В общем случае мы рассматриваем распределения $p(x|G)$ и $p(x|B)$, которые являются многомерными нормальными распределениями с общей дисперсией. Тогда уравнение (1) сводится к:

$$A_G = \{x | \sum w_i x_i > c\},$$

как следует из теории эконометрики. Здесь x_i – объясняющие переменные, w_i – связанные коэффициенты (веса) в линейной комбинации объясняющих переменных. Если взять $s(x) = \sum w_i x_i$, то можно различить в соответствии с этой оценкой и, таким образом, свести проблему только к одному измерению.

Дискриминантный анализ был введен Фишером (1936), который искал лучший способ разделения двух групп, используя линейную комбинацию переменных. Эйзенбайс (1977) подверг критике этот метод, заявив, что правило оптимально только для небольшого класса распределений. Однако Хэнд и Хенли (1997) утверждают, что «если переменные следуют за многомерным эллипсоидальным распределением (частным случаем которого является нормальное распределение), то линейное дискриминантное правило является оптимальным» [15, 16, 17].

Также было отмечено, что существует предвзятость выбора из-за того, что образец обучения для системы начисления баллов сделан из кандидатов, которым был предоставлен кредит. Это означает, что результаты смещены

применительно ко всему множеству. Эйзенбайс (1977) также видел проблемы в определении плохих и хороших групп в случае, когда между ними нет четкой границы и в предположении, что ковариационные матрицы двух распределений равны. В этом случае целесообразно использовать квадратичный дискриминантный анализ вместо линейного случая. Проблемы также возникают, когда кто-то хочет проверить значимость отдельных переменных, поскольку нет предположения о нормальности и поэтому невозможно выполнить статистический вывод [18].

Альтман (1968), который первым применил дискриминантный анализ, построил так называемый z -счет, который представляет собой линейную комбинацию нескольких объясняющих переменных для случая корпоративной проблемы предоставления кредита. Он обнаружил, что модель очень точна в правильном прогнозировании банкротства [19].

Преимущества метода ЛДА в том, что он прост и с его помощью можно очень легко оценить кредитоспособность, и он действительно работает очень хорошо; он часто используется банками для целей кредитного скоринга. Недостатком является то, что ЛДА требует нормально распределенных данных, но данные о кредитных операциях часто являются слабоструктурированными и не всегда есть возможность их классифицировать.

2.3.2 Логистическая регрессия

Распределение данных кредитной информации обычно ненормально, и этот факт теоретически может создавать проблемы при проведении ЛДА. Один из способов преодоления проблем с ненормальностью данных - это использовать расширение ЛДА, а именно с помощью модели, которая допускает некоторое параметрическое распределение. В этом случае подходящим расширением является обобщенная линейная модель, известная как логистическая модель. Учитывая вектор характеристик приложения x , вероятность дефолта p связана с вектором x соотношением:

$$\log\left(\frac{p}{1-p}\right) = w_0 + \sum w_i \log x_i.$$

Одним из преимуществ перед линейным дискриминантным анализом является использование метода максимального правдоподобия для оценки параметров w_i . Другое преимущество состоит в том, что можно обеспечить вероятность нахождения в каком-то конкретном классе. Логистический метод также может работать с группированными данными; решение состоит в том, чтобы взять фиктивные переменные для каждой группы данных. Несколько исследований показали, что эта модель превосходит дискриминантный анализ.

В последнее время данный анализ стал одним из основных подходов классификации кредитного скоринга в практике банков. Полученные коэффициенты имеют те же значения, что и в исследованиях, в которых использовалось правило принятия решений ЛДА. Тем не менее, они получены при гораздо более слабых предположениях. Фактические результаты классификации одинаковы для обоих методов анализа, и поскольку оба они чувствительны к высокой корреляции между объясняющими переменными, следует убедиться, что в обучающем наборе не осталось таких переменных. Другим недостатком этого метода является чувствительность к пропущенным значениям (все критерии с пропущенными значениями должны быть удалены). Аналогично, как и при линейном дискриминантном анализе, данный подход ограничен параметрической формой модели.

Лоуренс и Аршади (1995) использовали эту модель для анализа управления проблемными кредитами и определяющих факторов выбора решений с использованием ряда переменных заемщика и банка. В области ипотечного кредитования Кэмпбелл и Дитрих (1983) использовали немного другую модель, чтобы показать, что рост ипотеки, отношение кредита к стоимости, процентные ставки и уровень безработицы играют важную роль в объяснении предоплаты по ипотечным кредитам, просрочек и дефолтов. Недавно Charitou, Neophytou и Charalambous (2004) обнаружили, что

логистический метод превосходит другие методы прогнозирования значений по умолчанию [20, 21].

2.3.3 Деревья классификации и регрессии

Деревья классификации и регрессии – это непараметрический метод, разработанный Брейманом, Фридманом, Ольшеном и Стоуном (1984). Это гибкая и мощная техника; однако она используется в банковской практике главным образом только в качестве вспомогательного инструмента для сопровождения методов параметрической оценки, описанных ранее. Данный метод служит, например, в процессе выбора регрессоров или характеристик с наивысшей объяснительной силой. Он использует двоичные деревья и классифицирует набор данных в конечное число классов. Первоначально был разработан как инструмент для работы с бинарными ответами и как таковой он подходит для использования в кредитном скоринге, где ответы по умолчанию и не по умолчанию содержатся в данных.

Как и в случае с рассмотренными ранее методологиями, мы предполагаем наличие обучающего набора векторов измерений $X_T = \{x^j\}$ вместе с информацией о том, является ли отдельный j – ошибочным или нет (следовательно, y^j кодируется 1 или 0 соответственно). Дерево данного метода состоит из нескольких уровней узлов: первый уровень состоит из корневого узла; последний состоит из листовых узлов. Поскольку это двоичное дерево, каждый узел (кроме листьев) связан с двумя узлами на следующем уровне. Корневой узел содержит весь обучающий набор; другие узлы содержат подмножества этого набора. В каждом узле это подмножество делится на две непересекающиеся группы на основе одной конкретной характеристики x_i из вектора измерения. Если x_i – ординал, то разделение является результатом факта, связанного с конкретным человеком, относительно того, является ли $x_i > c$, для некоторой константы c . Если предыдущее утверждение верно, индивидуум j классифицируется в правый

узел; если нет, то индивидуум классифицируется на левый узел. Аналогичное правило применяется, если x_i является категориальной переменной.

X_i выбирается среди всех возможных критериев, а постоянная c выбирается так, чтобы результирующие подвыборки были настолько тождественными, насколько это возможно. Другими словами: x_i выбрана для минимизации различий результирующих подвыборок (разнообразие в этом контексте будет определено в настоящее время). Процесс классификации представляет собой рекурсивную процедуру, которая начинается в корневом узле, и в каждом последующем узле (за исключением листьев) выбираются один отдельный критерий и правило разделения (или константа c). Сначала наилучшее разделение находится для каждого критерия, затем среди этих критериев выбирается тот, у которого наилучшее разбиение.

Результатом этой процедуры является то, что подвыборки являются более однородными, чем первоначальная выборка. Процедура заканчивается тогда, когда содержит только индивидуумов с одинаковым y^j , или невозможно дальнейшее уменьшение разнородности. Для иллюстрации пусть $p(\cdot|t)$ будет пропорцией групп G и B , присутствующих в узле t . В качестве примера функций разделения можно взять индексную функцию Джини, которая определяется как $d(t) = p(G|t)p(B|t)$. Значение константы c , которое способствует разделению между узлами, определяется так, чтобы минимизировать взвешенную диверсификацию в дочерних узлах (то есть узлах, для которых узел t является родительским). Формально цель состоит в том, чтобы выбрать значение c , которое минимизирует $p_L d(t_L) + p_R d(t_R)$, где p_L и p_R – пропорции индивидуумов, идущих в узлы t_L и t_R соответственно. В итоге полученное дерево обычно очень большое, но существуют алгоритмы для того, чтобы преобразовать его в более простое конечное дерево. Преимущества метода деревьев решений в кредитном скоринге состоят в том, что он очень интуитивно понятен, прост для объяснения руководству и способен справляться с отсутствием наблюдения. Основным недостатком является вычислительная нагрузка в случае больших наборов данных,

поскольку в каждом узле необходимо изучить каждый критерий. Очень часто результирующее дерево достаточно велико, поэтому процесс обучения модели становится слишком трудоемким. В некоторых эмпирических исследованиях также отмечается, что часто деревья нестабильны, поскольку небольшие изменения в обучающем наборе могут значительно изменить структуру всего дерева. Существенной проблемой является также тот факт, что данный метод оптимизирует только локально по одной переменной за раз и таким образом, это не может минимизировать общие затраты на неправильную классификацию.

Первыми, кто применил метод деревьев решений в области кредитного скоринга, были Фридман, Альтман и Као (1985), которые обнаружили, что он превосходит линейный дискриминантный анализ. Относительно розничного кредитования является исследование Девани (1994), которое использовало методы логистического анализа и деревьев решений для отбора финансовых критериев, которые смогут максимально точно предсказать дефолт и обнаружили, что оба метода существенно различаются при выборе критериев: те, которые были выбраны методом деревьев решений, не были столь важны в соответствии с логистическим анализом. Недавним дополнением к литературе, посвященной методу деревьев решений, является работа Фельдмана и Гросса (2005). Авторы используют этот метод для данных по ипотечному кредиту и обсуждают плюсы и минусы метода деревьев решений по сравнению с традиционными методами [22, 23].

2.3.4 Нейронные сети

Нейронная сеть – это математическое представление, вдохновленное человеческим мозгом и его способностью адаптироваться на основе притока новой информации. Математически нейронные сети – это инструмент нелинейной оптимизации. Существует множество различных типов нейронных сетей, которые были описаны в литературе.

Конструкция нейронной сети, называемая многослойным персептроном, особенно подходит для классификации и широко используется на практике. Сеть состоит из одного входного слоя, одного или нескольких скрытых слоев и одного выходного слоя, каждый из которых состоит из нескольких нейронов. Каждый нейрон обрабатывает свои входные данные и генерирует одно выходное значение, которое передается нейронам в следующем слое. Каждый нейрон во входном слое (индексированный $i = 1, \dots, n$) доставляет значение одного предиктора (или критерия) из вектора x . При рассмотрении дискриминации по умолчанию или не по умолчанию, один выходной нейрон является удовлетворительным. На каждом уровне распространение сигнала осуществляется следующим образом. Сначала для каждого нейрона вычисляется взвешенная сумма входных данных: выходное значение каждого нейрона на исходящем сетевом уровне умножается на соответствующий вес соединения с этим нейроном. Передаточная функция $g(x)$ затем применяется к этой взвешенной сумме, чтобы определить выходное значение нейрона. Итак, каждый нейрон в скрытом слое (индексированный $j = 1, \dots, q$) производит так называемую активацию:

$$a_j = g(\sum_i w_{ij} x_i).$$

Нейроны в выходном слое (индексированные $k = 1, \dots, m$) ведут себя аналогично нейронам скрытого слоя, создавая выходные данные сети:

$$y_k = f(\sum_j w'_{jk} a_j) = f[\sum_j w'_{jk} g(\sum_i w_{ij} x_i)],$$

где w_{ij} и w'_{jk} – веса.

Есть два этапа оптимизации. Во-первых, веса должны быть инициализированы, а во-вторых, реализована нелинейная схема оптимизации. На первом этапе веса обычно инициализируются небольшим случайным числом. Второй этап – это обучение нейронной сети. Наиболее популярным алгоритмом для обучения многослойных персептронов является алгоритм обратного распространения. Как следует из названия, ошибка, вычисленная из выходного слоя, распространяется по сети обратно, а веса изменяются в соответствии с их взаимодействием с функцией ошибки. По сути, обратное

распространение выполняет поиск локального градиента и, следовательно, его реализацию; хотя это не требует вычислительных затрат, оно не гарантирует достижения глобального минимума. Для каждого отдельного человека весовые коэффициенты изменяются таким образом, чтобы ошибка, вычисленная из выходного слоя, была минимизирована.

Нейронные сети были описаны в 1960-х годах, но их первое использование в литературе, касающейся кредитного скоринга, появилось только в начале 1990-х годов. Альтман (1994) использовал линейный дискриминантный анализ и нейронную сеть для диагностики корпоративных финансовых проблем для 1000 итальянских фирм и пришел к выводу, что нейронные сети не являются явно доминирующим математическим методом по сравнению с традиционными статистическими методами, такими как дискриминантный анализ, а также он не уступает нейронной сети в точности принятия решений. С другой стороны, с другими результатами Тем и Кианг (1992) изучили применение модели нейронной сети для прогнозирования банкротства Техаса на период 1985–1987 гг.; они сравнили точность прогноза нейронной сети с точностью линейного дискриминантного анализа, логистической регрессии, метода k-ближайших соседей и модели дерева решений. Их результаты показывают, что нейронная сеть все же является наиболее точной, после чего следует линейный дискриминантный анализ, логистическая регрессия, деревья решений и метод k-ближайших соседей [24].

Непосредственно касаясь розничного кредитного скоринга, Дисей, Крук и Оверстрит (1996) исследуют нейронную многослойную персептронную сеть, сочетание нейронной сети эксперта, линейного дискриминантного анализа и логистической регрессии для оценки кредитных заемщиков в индустрии кредитных союзов. Их результаты показывают, что настроенные нейронные сети предлагают очень многообещающий путь, если мерой производительности является процент правильно классифицированных плохих кредитов. Однако, если показателем эффективности является процент

правильно классифицированных хороших и плохих кредитов, модели логистической регрессии сопоставимы с подходом нейронных сетей [25].

Уэст (2000) исследовал точность оценки кредитоспособности пяти различных моделей нейронных сетей. Модели оценки кредитоспособности нейронной сети были протестированы с использованием 10-кратной перекрестной проверки с двумя реальными наборами данных (данные по кредитам Германии и Австралии). Результаты сравнивались с более традиционными методами, рассмотренными в этой статье. Их исследования показали, что логистическая регрессия является хорошей альтернативой нейронным моделям. Логистическая регрессия была немного более точной, чем модели нейронной сети для среднего случая, который включает в себя несколько итераций обучения нейронной сети низкого уровня [26].

Основным недостатком нейронных сетей является отсутствие у них возможности объяснения. Несмотря на то, что они могут достичь высокой степени точности прогноза, причины, почему и как было принято решение, недоступны. Например, в случае отказа в кредите невозможно определить, какой из критериев был именно ключевым, чтобы отклонить заявку. Следовательно, очень сложно объяснить результаты решения менеджерам.

Таким образом, были выявлены наиболее распространенные методы, используемые в процессе кредитного скоринга заявителей на розничные кредиты. Обзор концентрируется на наиболее актуальных методах, которые соответствуют их применению в практике банков.

2.4 Анализ программных средств для обработки и анализа данных

Проанализировав методы оценки кредитоспособности, далее необходимо выбрать программные средства для того, чтобы в дальнейшем провести анализ.

В последнее время получили широкое распространение программные средства или информационные системы, предназначенные для автоматизации

работ статистической обработки данных, которые позволяют собирать, хранить и обрабатывать большие объемы данных. Достоинством таких систем является адаптация информационной базы и функций системы к условиям функционирования организации. Однако, в силу сложности реализации математических методов, такие системы, как правило, включают лишь ограниченный набор аналитических методов.

В настоящий момент наиболее популярными статистические пакетами, которые могут быть легко подключены информационной системе обработки информации являются STATISTICA, SPSS и Deductor. Рассмотрим их подробнее.

Пакет прикладных программ STATISTICA – универсальная система анализа данных, разработанная компанией StatSoft, построенная по модульному принципу, каждый модуль выполняет определенный набор функций и может быть использован и автономно. К основным возможностям пакета относятся:

- реализация математических методов;
- графическая интерпретация результатов;
- возможность импорта и экспорта в другие программы.

STATISTICA может применяться в разнообразных сферах деятельности, а именно в банковской сфере, в исследованиях торговой и маркетинговой деятельности, в медицине и даже на производстве.

Пакет прикладных программ SPSS (Statistical Package for Social Science) – статистический пакет, разработанный компанией SPSS Inc, предназначенный для работы в операционной системе MS Windows. К его основным возможностям относятся:

- реализация методов математической и статистической обработки;
- возможность объединения нескольких баз данных;
- интерпретация результатов с помощью нестандартных отчетов, для более качественной оценки;

- возможность обмена данными с другими приложениями MS Windows.

Пакет прикладных программ Deductor – статистический пакет, разработанный фирмой Base Group Labs, который включает в себя многомерное хранилище данных Deductor Warehouse, аналитическое приложение Deductor Studio и Deductor Viewer.

Основными функциями данного пакета являются:

- реализация функций импорта, экспорта данных;
- возможность обработки данных и дальнейшего анализа;
- отображение результатов анализа в удобном для пользователя виде.

Реализованные в Deductor обработчики покрывают основную потребность в анализе данных и создании законченных аналитических решений на базе Data Mining. Data Mining позволяет извлекать из данных знания и превратить в конкурентные преимущества: качественно прогнозировать, точнее, выявлять целевые аудитории, предсказывать развитие событий, управлять рисками и прочее.

2.5 Выбор метода для оценки кредитоспособности

Ответ на вопрос о том, какой метод выбрать, не прост и зависит главным образом от предпочтений банка, доступности данных, его характеристик и т. д. Различные методы оценки можно сравнить по их результатам. Часто нет лучшего метода для разных наборов данных. Тем не менее, метод интеллектуального анализа данных является достаточно распространённым методом на практике, в основном из-за (почти) отсутствия каких-либо допущений в отношении переменных, за исключением отсутствующих значений и мультиколлинеарности среди переменных. В отличие от него, непараметрические методы могут иметь дело с отсутствующими значениями и мультиколлинеарностью (или корреляциями) среди переменных, но часто требуют вычислительных затрат. Однако правила, которые строятся на основе

некоторых из этих методов, может быть трудно объяснить, как руководителю, так и клиенту. Например, несмотря на то, что нейронные сети обладают потенциалом для получения превосходных результатов, неспособность объяснить, почему может представлять серьезный недостаток для применения этого метода на практике.

В любом случае, по мере увеличения объемов розничных кредитов, улучшение качества и методологии в процессах кредитного скоринга становятся обязательными для коммерческих банков. Это особенно важно в периоды резкого роста оптимизма в отношении будущих доходов. Если реальная производительность падает ниже этих ожиданий, серьезность понесенных убытков может быть очень высокой для коммерческих банков.

Были рассмотрены методы, с помощью которых можно оценить кредитоспособность заемщика. Необходимо на практике выяснить какой же из них будет удовлетворять всем требованиям, а именно проводить объективную оценку, на возможных наборах данных, учитывая, как качественные, так и количественные показатели, выдавая при этом результаты соответствующие реальности.

Для того, чтобы протестировать данные методы ИАД обратимся к такой программе как Deductor. Deductor – аналитическая платформа, которая позволяет создать эффективную систему поддержки принятия бизнес-решений. Она предназначена для выполнения сбора информации и структурирования ее в виде данных и последующей углубленной аналитики полученной бизнес-информации с возможностью визуализации, и создания необходимых отчетов.

Методики анализа Data Mining применяются для решения различных задач, выполняемых аналитической платформой Deductor, также их можно использовать для анализа кредитоспособности физических лиц.

Методы анализа данных являются лишь частью общей картины управления данными, которая также включает в себя архитектуру и моделирование данных, инструменты сбора данных, методы сбора данных,

типы визуализации данных, безопасность данных, метрики, отображение данных и интеграция, бизнес-аналитика и др.

Какой тип анализа данных использовать? Ни один метод или метод анализа данных не может быть определен как лучший метод для анализа данных. Все они имеют свою роль, значение, преимущества и недостатки.

Выбор методов зависит от конкретной проблемы и набора данных. Таким образом, выбор правильных методов анализа данных может стать решающим моментом для оценки кредитоспособности физического лица.

2.6 Этапы алгоритма оценки кредитоспособности

Алгоритм оценки состоит из нескольких этапов.

1. Анкетирование клиента.
2. Сбор данных.
3. Очистка данных.
4. Преобразование данных.
5. Проведение анализа данных.
6. Получение результата.

На рисунке 4 показана схема этапов алгоритма оценки кредитоспособности.

Первый этап – сбор данных с помощью анкетирования. На этом этапе проводится подробный опрос, для того, чтобы узнать необходимые личные данные физического лица для анализа, а именно персональные, контактные, демографические, сведения о занятости, финансовая информация, а также условия по кредитованию.

Когда все данные получены, можно перейти ко второму этапу – очистке данных. Полученные от клиента данные вносятся в базу для того, чтобы привести их к единой структуре удобной для хранения и последующей обработки. Затем нужно сделать проверку данных после очистки, если остались неочищенные данные, необходимо очистить их вручную, иначе преступаем к следующему этапу.

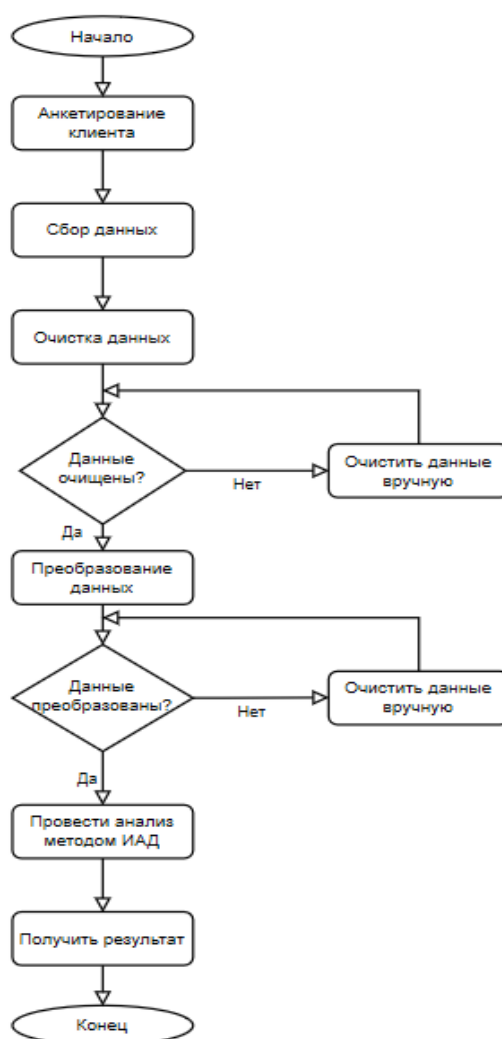


Рисунок 4 – Схема алгоритма

Далее идет этап преобразования данных. Данные, полученные в результате очистки следует привести к виду, с помощью которого можно провести анализ данных. Есть эталонная модель, в которой собраны ключевые показатели для оценки кредитоспособности, каждый показатель имеет свою категорию, которая оценивается конкретным баллом. Поэтому мы преобразовываем данные по этой модели для дальнейшего анализа. Затем опять делаем проверку на корректность данных. Если данные не проходят проверку, делаем их корректировку вручную, если проходят переходим к следующему этапу.

Предпоследним этапом идет анализ очищенных и преобразованных данных в Deductor с использованием методов интеллектуального анализа данных. На основании проведенных экспериментов, оптимальным методом для решения задачи оценки кредитоспособности физических лиц, наиболее высокую точность результатов, оказался метод Деревьев решений.

И после того, как данные были загружены и проведен анализ по всем показателям, в конечном итоге выдается результат – выдать или не выдать кредит клиенту.

Используя данный алгоритм, банки могут охватить более подробную информацию о заемщике: личные данные, кредитную историю, финансовое положение и много другое, что позволяет изучить большее количество показателей, что дает возможность банку выработать индивидуальный подход к любому потенциальному заемщику, в рамках которого будут учтены необходимые показатели.

Для определения кредитного рейтинга целесообразно использовать методы интеллектуального анализа данных, так как они позволяют классифицировать очередного заемщика по степени принадлежности к одному или другому классу – хороший или плохой, что позволит банку принять правильное решение относительно предоставления кредита и тем самым снизить кредитный риск.

Выводы

Одним из основных методов управления кредитным риском в банке является предупреждение, то есть устранение предпосылок возникновения кредитного риска в будущем. Важную роль в этом играет оценка кредитоспособности заемщика и установление его кредитного рейтинга. То есть, чтобы повысить эффективность кредитного процесса, серьезное внимание следует уделить разработке методов анализа кредитоспособности и платежеспособности заемщиков, которые позволяют давать обоснованные рекомендации по предоставлению кредитов, тем самым повышая их качество.

Наш современный информационный век ведет к динамичному и чрезвычайно высокому росту мира интеллектуального анализа данных. Нет сомнений, что для этого требуются адекватные и эффективные методы, методы и инструменты анализа данных различных типов, которые могут удовлетворить постоянно растущие потребности в бизнес-исследованиях.

Оценить кредитоспособность отдельных лиц гораздо сложнее, потому что зачастую для этого недостаточно информации. Этот недостаток можно компенсировать с помощью технологии Data Mining. Так как наиболее перспективным направлением в построении скоринговых систем для оценки кредитоспособности заемщиков, является технология Data Mining, которая представляет собой процесс «обнаружения ранее неизвестных, нетривиальных, практически полезных и доступных интерпретаций знаний в необработанных данных, необходимых для принятия решений в различных областях человеческой деятельности».

3 Разработка алгоритма оценки кредитоспособности и проведение экспериментов

3.1 Разработка алгоритма оценки кредитоспособности

С непрерывным развитием финансовых и банковских учреждений, кредиты играют все более важную роль в экономике. Экономическая глобализация и появление новых каналов, через которые предоставляются услуги, такие как Интернет, позволяют клиентам искать и выбирать кредиторов без региональных ограничений. Из-за этой тенденции кредиторы должны быть готовы и способны расширить свой бизнес на других рынках. Поэтому финансовые учреждения сталкиваются с жесткой глобальной конкуренцией, что побуждает их уделять больше внимания технологиям, с помощью которых анализируются заявки на кредит.

Поэтому необходимо найти оптимальный алгоритм для прогнозирования кредитного риска в условиях потребительского рынка. Для этого будут исследованы возможности скоринговых моделей в сочетании с использованием интеллектуального анализа данных.

Цель этого исследования – провести анализ кредитного скоринга для физических лиц, но не с использованием классических статистических методов, а с использованием алгоритмов интеллектуального анализа данных, для того чтобы проверить эффективности набора алгоритмов интеллектуального анализа данных.

3.2 Анкетирование потенциальных заемщиков

Первым этапом является анкетирование потенциальных заемщиков. Рассмотрим подробнее какие данные будут запрашиваться у клиентов.

Во-первых, персональные данные, а именно ФИО, дата рождения и паспортные данные. На рисунке 5 представлена часть анкеты, в которой заполняются персональные данные физического лица.

ПЕРСОНАЛЬНЫЕ ДАННЫЕ

ФИО

Дата рождения

Паспорт гражданина РФ

Рисунок 5 – «Персональные данные»

Далее необходимо заполнить контактную информацию о клиенте, к ней относятся номера телефонов, адреса регистрации и проживания, срок проживания, а также основание проживания. На рисунке 6 можно увидеть поля, которые будут отражать контактные данные заемщика.

КОНТАКТНЫЕ ДАННЫЕ

Телефон моб. Телефон дом.

Адрес по месту регистрации

☐ совпадает/ ☐ не совпадает с адресом регистрации

Адрес по месту жительства

Срок проживания

Основание для проживания ☐ Собственность ☐ Аренда ☐ Служебная

Рисунок 6 – «Контактные данные»

Следующее, что банку необходимо знать о клиенте, это его демографические данные (пол, семейное положение, наличие детей, данные об образовании). На рисунке 7 можно увидеть часть анкеты, в которой заполняются демографические данные физического лица.

ДЕМОГРАФИЧЕСКИЕ ДАННЫЕ

Пол ☐ муж ☐ жен

Семейное положение ☐ холост/не замужем ☐ женат/замужем ☐ в разводе
☐ вдова/вдовец ☐ гражданский брак

Количество детей Из них несовершеннолетних

Образование ☐ среднее ☐ среднее специальное
☐ незаконченное высшее ☐ высшее
☐ 2 и более высших/ учебная степень

Рисунок 7 – «Демографические данные»

Потом заполняются сведения о занятости заемщика, то есть все данные о текущей и предыдущих работах (название и данные организации, должность, общий стаж работы, стаж на последнем месте работы). На рисунке 8 можно увидеть поля, которые будут отражать сведения занятости клиента.

СВЕДЕНИЯ О ЗАНЯТОСТИ

Полное наименование организации

Наименование должности

Тип занятости ☐ постоянная занятость ☐ не работаю
☐ частичная занятость

Род занятий ☐ наемный сотрудник
☐ ИП/ собственный бизнес

Общий трудовой стаж лет мес

Стаж на последнем месте работы лет мес

Юридический адрес работодателя регион
 район город
 улица дом офис

Рабочий телефон

Рисунок 8 – «Сведения о занятости»

Далее необходимо оценить финансовые данные заемщика, а именно сведения о основных и дополнительных доходах, а также наличие собственности. На рисунке 9 можно увидеть часть анкеты, в которой заполняются сведения о финансовом положении физического лица.

The image shows a light blue rectangular box with the title 'ФИНАНСОВАЯ ИНФОРМАЦИЯ' in bold black text at the top center. Below the title, there are five rows of input fields. The first three rows are for income: 'Основной доход', 'Пенсии и прочие соц.выплаты', and 'Прочие доходы', each followed by a white rectangular input box. The fourth row is for 'Автомобиль', followed by two radio button options: 'есть' and 'нет'. The fifth row is for 'Недвижимость', followed by a white rectangular input box.

Рисунок 9 – «Финансовая информация»

И в последней части анкеты будут рассматривать условия кредита, в которые входят сумма, срок и цель кредита. На рисунке 10 можно увидеть поля, которые необходимо заполнить для определения условия кредита.

The image shows a light blue rectangular box with the title 'ЗАПРАШИВАЕМЫЕ УСЛОВИЯ' in bold black text at the top center. Below the title, there are three rows of input fields. The first two rows are for 'Сумма кредита' and 'Срок кредита (мес)', each followed by a white rectangular input box. The third row is for 'Цель кредита', followed by a list of six radio button options: 'ремонт', 'покупка автомобиля/недвижимость', 'развитие бизнеса', 'покупка бытовой техники/мебели', 'погашение кредита', and 'прочее'.

Рисунок 10 – «Финансовая информация»

После того, как все поля анкеты были заполнены, у нас появляются сведения о клиенте. В анкете есть вопросы, ответы на которых представлены как в числовом, так и в тестовом виде. Поэтому для дальнейшего анализа необходимо преобразовать данные и привести их к единой структуре.

3.3 Описание исходного набора данных

После анкетирования полученные данные были собраны в таблицу Excel. На рисунке 11 фрагмент базы, полученной в результате анкетирования клиентов.

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Заемщик	Возраст	Семейное положение	Дети	Образование	Место проживания	ТС	Должность	Место работы	Стаж последний	Общий стаж	Доход	Доп. доход	Кред
2	1	25	не женат	0	высшее	собственная квартира	нет	продавец	"Евросеть"	6 месяцев	3 года	25000	нет	
3	2	19	не женат	0	среднее общее	у родителей	нет	студент	-	0	0	65000	1400	
4	3	35	женат	1	высшее	собственная квартира	да	директор	ОАО "Аврора"	3 года	7 лет	80000	нет	
5	4	26	замужем	0	высшее	собственная квартира	нет	оператор	МДМ Банк	1 год	3 года	20000	нет	
6	5	20	не женат	0	среднее общее	у родителей	нет	студент	-	0	0	7000	нет	
7	6	55	разведен	1	днее специаль	у родителей	нет	менеджер	"Перекресток"	5 лет	15 лет	29000	10000	
8	7	60	замужем	2	высшее	собственная квартира	нет	пенсионер	-	7 лет	28 лет	17000	нет	
9	8	34	не замужем	1	днее специаль	съемная квартира	нет	дприинимат	Салон красоты	2 года	5 лет	40000	нет	
10	9	29	не женат	0	днее специаль	съемная квартира	да	продавец	Шины и диски	3 года	8 лет	27000	нет	
11	10	23	не замужем	0	высшее	у родителей	да	редактор	Компас	2 года	5 лет	30000	нет	
12	11	24	не замужем	0	высшее	съемная квартира	нет	бухгалтер	АВИА	1 месяц	1 год	45000	нет	
13	12	57	не замужем	1	среднее общее	собственная квартира	нет	продавец	PALLA	3 года	28 лет	15000	нет	
14	13	24	не замужем	0	высшее	у родителей	нет	маркетолог	Цемент+	2 года	3 года	33000	7000	
15	14	26	женат	0	днее специаль	съемная квартира	да	продавец	12 Студьев	1 год	4 года	46000	15000	
16	15	29	не женат	0	днее специаль	у родителей	нет	незработный	-	4 месяца	2 года	-	нет	
17	16	71	не замужем	2	среднее общее	собственная квартира	нет	пенсионер	-	5 лет	35 лет	18000	нет	
18	17	22	не женат	0	высшее	съемная квартира	нет	программист	Ростелеком	2 года	4 года	70000	нет	

Рисунок 11 – База потенциальных заемщиков

Данные, полученные на этапе анкетирования не структурированы, для того чтобы выполнить с ними анализ была разработана модель, которая позволила упорядочить данные и привести их к удобному для анализа виду. Были определены показатели, каждый из которых имеет свою категорию, в свою очередь категории присваивается определенный балл. Таким образом данные были преобразованы и представлены в виде удобном для дальнейшего анализа.

Описание показателей и их категории с балльными эквивалентами представлены в таблице 2.

Таблица 2 – Описание показателей

Показатель	Описание	Категория	Балл
Кредит	Кредитоспособность (DV)	Хороший	1
		Плохой	0
Баланс_тек_счета	Баланс текущего счета	пустой счет	2
		0 <= ... < 50000	3
		... >= 50000 или текущий счет не менее 1 года	4
		нет действующего счета	1
Продолжительность	Продолжительность в месяцах		
кредит_история	Оплата предыдущих кредитов	нет предыдущих кредитов / вернул все предыдущие кредиты	2
		погасил предыдущие кредиты в этом банке	4
		нет проблем с текущими кредитами в этом банке	3
		плохая история с предыдущими кредитами	0
		проблемный текущий счет / есть еще кредиты в других банках	1
Цель_кредита	Цель кредита	новый автомобиль	1
		б/у автомобиль	2
		мебель и фурнитура	3
		гаджеты	4
		бытовая техника	5
		ремонт	6
		образование	7
		отпуск	8
		переподготовка	9
		бизнес	10
		другое	0
сбережения	Наличие сбережений или акций	< 10000	2
		10000,- <= ... < 50000	3
		50000,- <= ... < 100000	4
		>= 100000	5
		отсутствуют	1
стаж_работы	Общий стаж работы	безработный	1
		<= 1 year	2
		1 <= ... < 4лет	3
		4 <= ... < 7 лет	4
		>= 7 лет	5
Рассрочка	Рассрочка в% от доступного дохода	>= 35	1
		25 <= ... < 35	2
		20 <= ... < 25	3
		< 20	4

Продолжение таблицы 2

Показатель	Описание	Категория	Балл
сем_положение	Семейное положение / Пол	мужчина: разведен / живет отдельно	1
		женщина: разведена / живет отдельно / замужем	2
		мужчина: холост	2
		мужчина: женат / овдовел	3
		женщина: не замужем	4
поручитель	Наличие поручителей	нет	1
		созаявитель	2
		поручитель	3
проживание	Проживание в текущем месте жительства	< 1 года	1
		1 <= ... < 4 лет	2
		4 <= ... < 7 лет	3
		>= 7 лет	4
собственность	Наличие собственности	право собственности на дом или землю	4
		страхование жизни	3
		автомобиль/другое	2
		отсутствуют	1
возраст	Возраст		
Тек_кредит	Текущие кредиты	в других банках	1
		в магазине	2
		нет текущих кредитов	3
место_жительства	Место жительства	съемная квартира	2
		собственная квартира	3
		у родителей/родственников	1
количество_кредит	Количество предыдущих кредитов в этом банке (включая текущий)	один	1
		два или три	2
		четыре или пять	3
		шесть и более	4
род_деятельности	Род деятельности	безработный / без постоянной работы	1
		работающий на постоянной основе	2
		квалифицированный работник / государственный служащий	3
		предприниматель	4
иждивенцы	Наличие иждивенцев	от 0 до 2	2
		3 и более	1
телефон	Телефон	нет	1
		да	2
доп_доход	Дополнительный доход	да	1
		нет	2

На рисунке 12 можно увидеть фрагмент структурированной базы данных, полученный в результате преобразования данных по категориям показателей. Набор данных, который будет использоваться в анализе представляет собой таблицу заемщиков. Одна запись представляет собой данные одного клиента, представленные в виде переменных. Каждый заемщик обладает набором определенных переменных.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Кредит	баланс_тек_счета	продолжи_кредит	и	Цель_кред	Сумма_к	сбережен	стаж_раб	Рассрочк	сем_полс	поручите	проживае	собствен	возраст	Тек_кред	место_жи	количество	род_деят	индивен	телефон	доп_дохо
2	1	1	18	4	2	1049	1	2	4	2	1	4	2	21	3	1	1	3	1	1	1
3	1	1	9	4	0	2799	1	3	2	3	1	2	1	36	3	1	2	3	2	1	1
4	1	2	12	2	9	841	2	4	2	2	1	4	1	23	3	1	1	2	1	1	1
5	1	1	12	4	0	2122	1	3	3	3	1	2	1	39	3	1	2	2	2	1	2
6	1	1	12	4	0	2171	1	3	4	3	1	4	2	38	1	2	2	2	1	1	2
7	1	1	10	4	0	2241	1	2	1	3	1	3	1	48	3	1	2	2	2	1	2
8	1	1	8	4	0	3398	1	4	1	3	1	4	1	39	3	2	2	2	1	1	2
9	1	1	6	4	0	1361	1	2	2	3	1	4	1	40	3	2	1	2	2	1	2
10	1	4	18	4	3	1098	1	1	4	2	1	4	3	65	3	2	2	1	1	1	1
11	1	2	24	2	3	3758	3	1	1	2	1	4	4	23	3	1	1	1	1	1	1
12	1	1	11	4	0	3905	1	3	2	3	1	2	1	36	3	1	2	3	2	1	1
13	1	1	30	4	1	6187	2	4	1	4	1	4	3	24	3	1	2	3	1	1	1
14	1	1	6	4	3	1957	1	4	1	2	1	4	3	31	3	2	1	3	1	1	1
15	1	2	48	3	10	7582	2	1	2	3	1	4	4	31	3	2	1	4	1	2	1
16	1	1	18	2	3	1936	5	4	2	4	1	4	3	23	3	1	2	2	1	1	1
17	1	1	6	2	3	2647	3	3	2	3	1	3	1	44	3	1	1	3	2	1	1
18	1	1	11	4	0	3939	1	3	1	3	1	2	1	40	3	2	2	2	2	1	1
19	1	2	18	2	3	3213	3	2	1	4	1	3	1	25	3	1	1	3	1	1	1
20	1	2	36	4	3	2337	1	5	4	3	1	4	1	36	3	2	1	3	1	1	1
21	1	4	11	4	0	7228	1	3	1	3	1	4	2	39	3	2	2	2	1	1	1
22	1	1	6	4	0	3676	1	3	1	3	1	3	1	37	3	1	3	3	2	1	1
23	1	2	12	4	0	3124	1	2	1	3	1	3	1	49	1	2	2	2	2	1	1
24	0	2	36	2	5	2384	1	2	4	3	1	1	4	33	3	1	1	2	1	1	1
25	1	2	12	4	4	1424	1	4	4	3	1	3	2	26	3	2	1	3	1	1	1
26	1	1	6	4	0	4716	5	2	1	3	1	3	1	44	3	2	2	2	2	1	1
27	1	2	11	3	3	4771	1	4	2	3	1	4	2	51	3	2	1	3	1	1	1
28	1	1	12	2	2	652	1	5	4	2	1	4	2	24	3	1	1	3	1	1	1
29	1	2	9	4	3	1154	1	5	2	3	1	4	1	37	3	2	3	2	1	1	1

Рисунок 12 – Преобразованные данные

Правильно обработанная, а затем структурированная информация о заемщиках приносит пользу банкам. Таким образом банк может больше узнать о поведении и предпочтениях клиентов, а также более точно определить кредитный рейтинг заемщика.

Бальный рейтинг не всегда может показать объективный результат, поэтому необходимо провести анализ полученных данных с помощью средств Data Mining.

3.4 Архитектура хранилища данных

Далее рассмотрим многомерное хранилище данных, в котором собрана вся необходимая для анализа предметной области информация. Вся информация в хранилище содержится в структурах типа "звезда", где в центре расположены таблицы фактов, а "лучами" являются измерения.

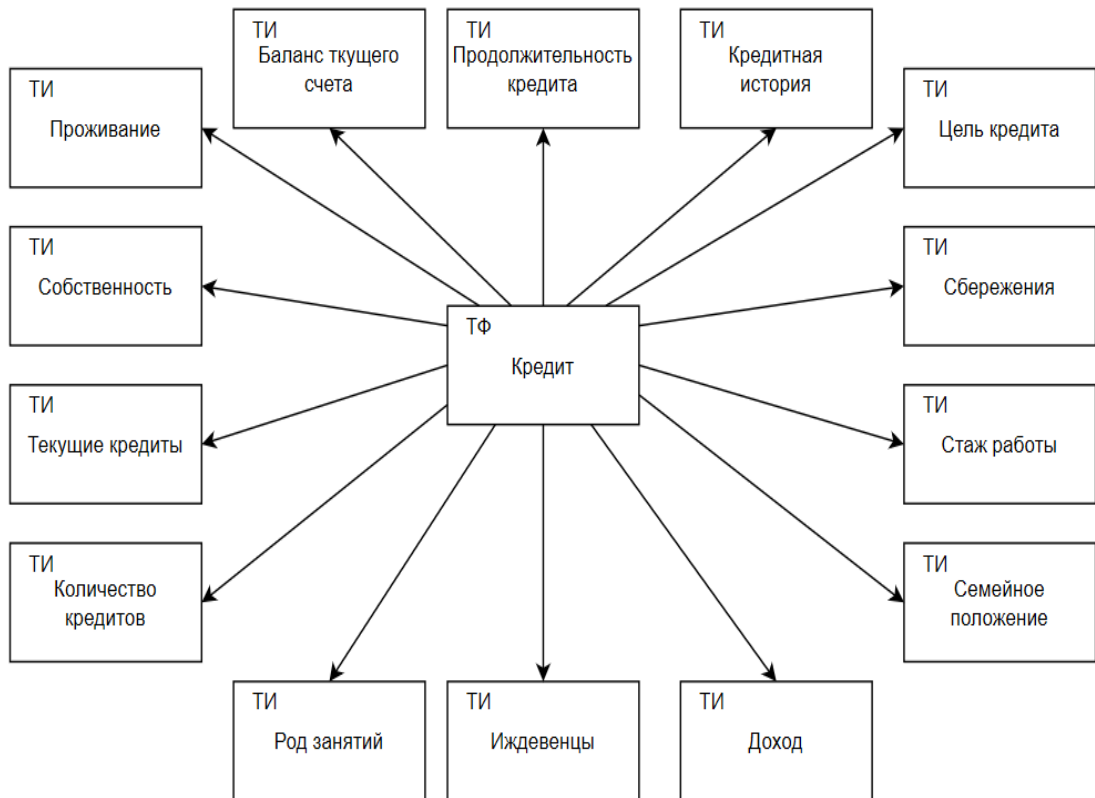


Рисунок 13 – Таблица фактов и измерений

На рисунке 13 изображена структура типа «Звезда». Данный вид архитектуры хранилища наиболее адекватен для решения задач анализа данных с помощью Data Mining.

3.5 Анализ данных в Deductor

Следующий этап, это непосредственно этап анализа данных, с помощью Deductor. Для начала необходимо импортировать базу в Deductor, с помощью Мастера импорта. На рисунке 14 изображено начало процесса импорта данных.

Далее необходимо проверить правильно ли структурированы данные из базы для дальнейшего переноса и работы с ней.

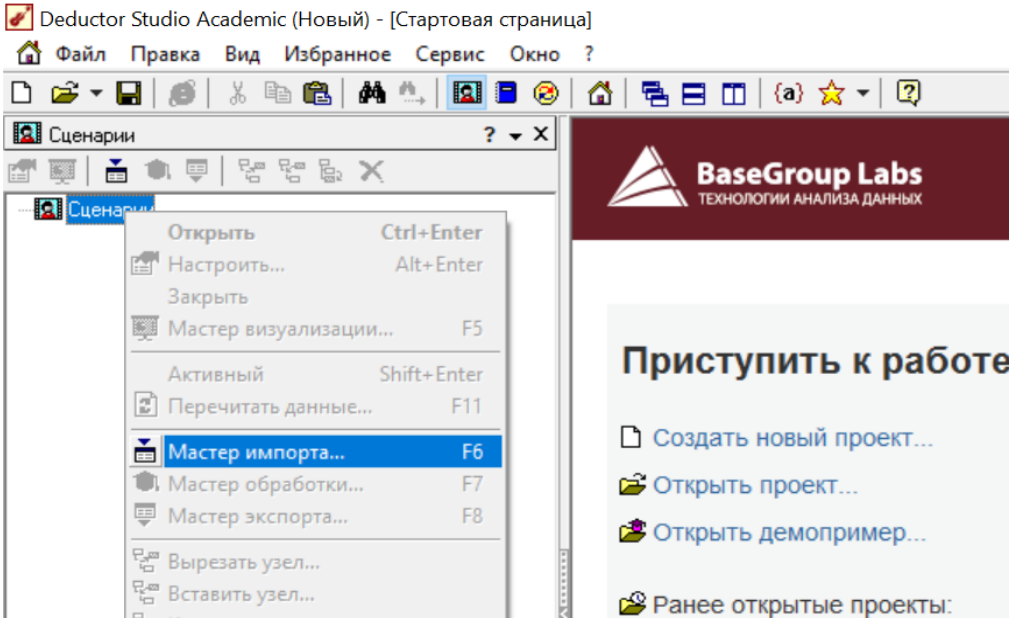


Рисунок 14 – Импорт базы данных в Deductor

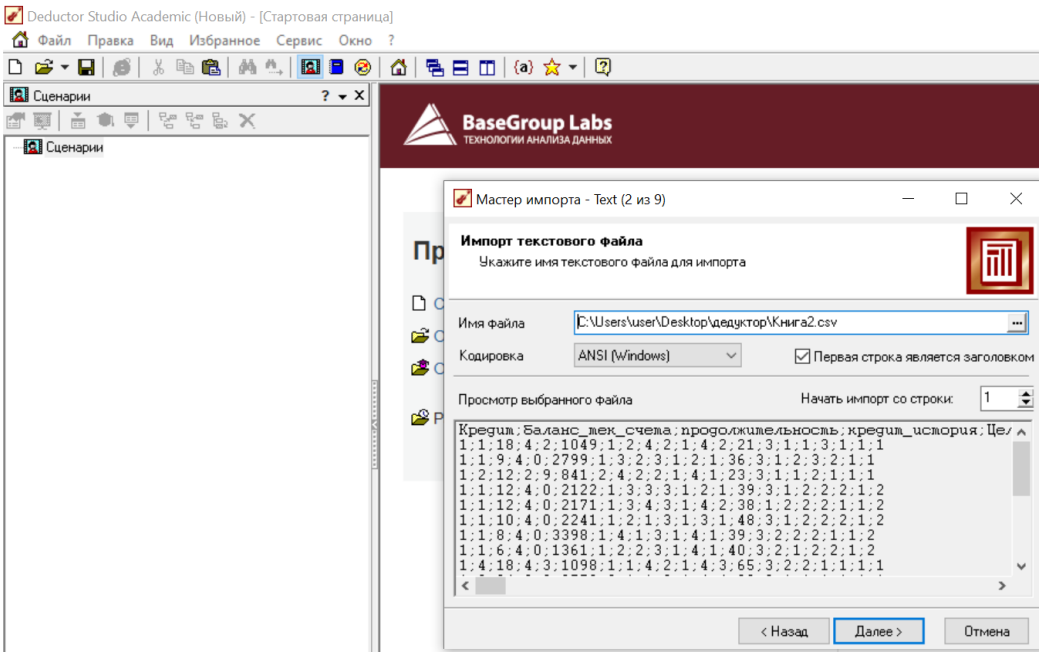


Рисунок 15 – Предварительный просмотр и настройка отображения базы данных

Следующий шаг в процессе импорта, установка параметров для корректного переноса данных. Это изображено на рисунке 16.

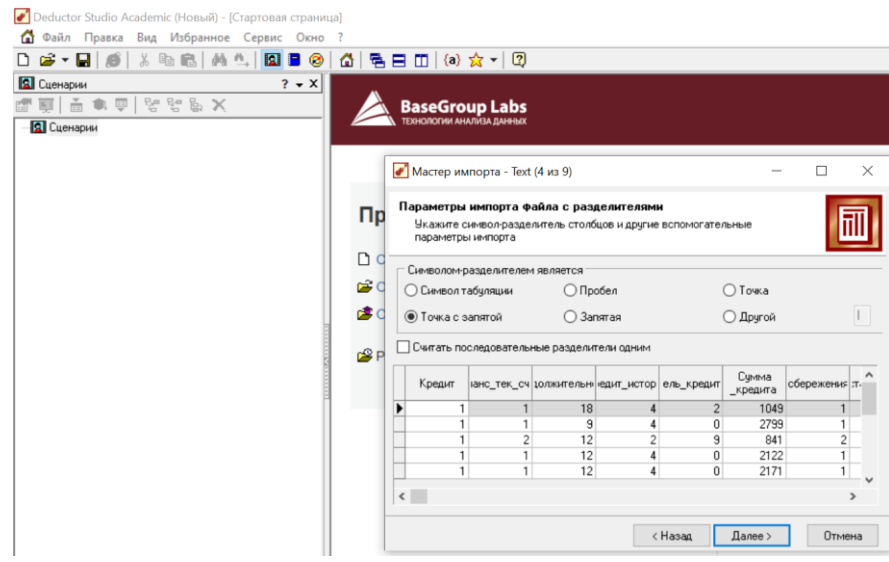


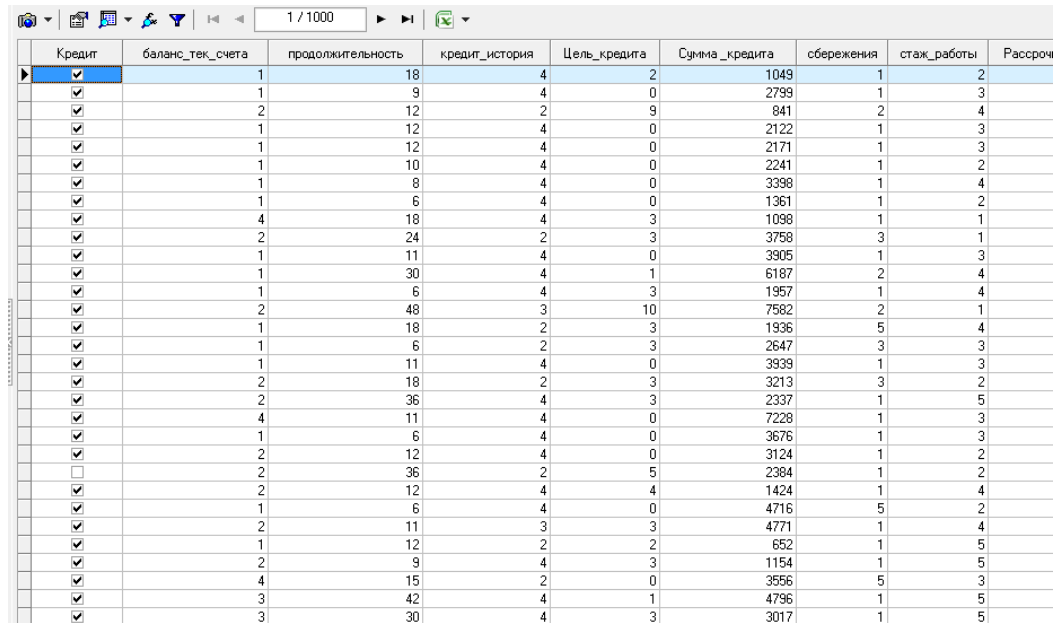
Рисунок 16 – Параметры импорта базы данных

На шаге импорта данных указываем параметры для столбцов согласно рисунку 17. Всем показателям задается тип «Вещественный», кроме показателя «Кредит», ему присваиваем тип данных «Логический».



Рисунок 17 – Настройка импорта данных

Фрагмент данных после импорта показан на рисунке 18.



Кредит	баланс_тек_счета	продолжительность	кредит_история	Цель_кредита	Сумма_кредита	сбережения	стаж_работы	Рассроч
☒	1	18	4	2	1049	1	2	
☒	1	9	4	0	2799	1	3	
☒	2	12	2	9	841	2	4	
☒	1	12	4	0	2122	1	3	
☒	1	12	4	0	2171	1	3	
☒	1	10	4	0	2241	1	2	
☒	1	8	4	0	3398	1	4	
☒	1	6	4	0	1361	1	2	
☒	4	18	4	3	1098	1	1	
☒	2	24	2	3	3758	3	1	
☒	1	11	4	0	3905	1	3	
☒	1	30	4	1	6187	2	4	
☒	1	6	4	3	1957	1	4	
☒	2	48	3	10	7582	2	1	
☒	1	18	2	3	1936	5	4	
☒	1	6	2	3	2647	3	3	
☒	1	11	4	0	3939	1	3	
☒	2	18	2	3	3213	3	2	
☒	2	36	4	3	2337	1	5	
☒	4	11	4	0	7228	1	3	
☒	1	6	4	0	3676	1	3	
☒	2	12	4	0	3124	1	2	
☐	2	36	2	5	2384	1	2	
☒	2	12	4	4	1424	1	4	
☒	1	6	4	0	4716	5	2	
☒	2	11	3	3	4771	1	4	
☒	1	12	2	2	652	1	5	
☒	2	9	4	3	1154	1	5	
☒	4	15	2	0	3556	5	3	
☒	3	42	4	1	4796	1	5	
☒	3	30	4	3	3017	1	5	

Рисунок 18 – Фрагмент таблицы данных

Далее переходим непосредственно к анализу этих данных в Deductor.

3.5.1 Дерево решений

Это очень популярный и современный алгоритм классификации в области интеллектуального анализа данных и машинного обучения. Дерево решений – это древовидная диаграмма, представляющая модель классификации или регрессии.

Он делит набор данных на все меньшие и меньшие наборы данных (которые содержат экземпляры с одинаковыми значениями), в то же время непрерывно разрабатывается соответствующее дерево решений. Дерево построено, чтобы показать, как и почему один выбор может привести к другому с помощью ветвей.

Среди преимуществ использования деревьев решений: знание предметной области не требуется; их легко понять; этапы классификации дерева решений очень просты и быстры.

При построении деревьев решений особое внимание уделяется следующим вопросам: выбору критерия атрибута, по которому пойдет разбиение, остановки обучения и отсечения ветвей.

Выходным параметром является Кредит (Кредитоспособность), а входными все остальные параметры. Результат представлен на рисунке 19.

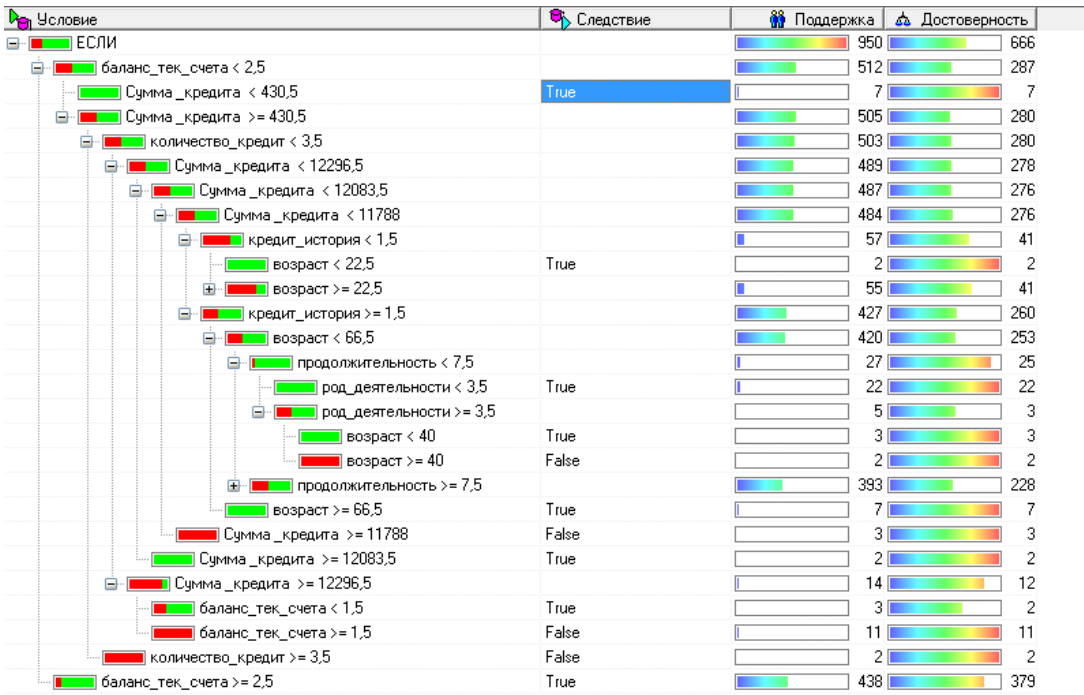


Рисунок 19 – Дерево решений

Отсюда видно, что если выполняется условие Баланс_тек_счета $\geq 2,5$ (3 или 4) (Баланс текущего счета), то с вероятностью 86,5% можно утверждать о хорошей кредитоспособности.

Если же выполняется условие Баланс_тек_счета $< 2,5$ и количество_кредит (Количество предыдущих кредитов в этом банке (включая текущий) $> 3,5$, с вероятностью 100% можно утверждать о наличии плохой кредитоспособности.

Работоспособность модели можно показать с помощью визуализатора «Таблица сопряженности» (рисунок 20).

Фактически	Классифицировано		
	False	True	Итого
False	209	91	300
True	21	679	700
Итого	230	770	1000

Рисунок 20 – Таблица сопряженности

Из таблицы видно, что классы классификатор определяет с хорошей точностью. Ошибочными оказались 11,2%.

При помощи визуализатора «Значимость атрибутов» можно узнать, насколько сильно зависит выходной атрибут от каждого входного. Данный визуализатор представляет собой таблицу, построенную на основе предварительно рассчитанного показателя «Значимость» на рисунке 21.

Целевой атрибут: Кредит				
№	Номер	Атрибут	Значимость, %	/
1	1	баланс_тек_счета		22,761
2	5	Сумма_кредита		13,948
3	2	продолжительность		11,350
4	13	возраст		6,922
5	4	Цель_кредита		6,477
6	17	род_деятельности		5,602
7	6	сбережения		5,279
8	3	кредит_история		4,928
9	12	собственность		3,905
10	7	стаж_работы		3,546
11	15	место_жительства		3,387
12	16	количество_кредит		2,519
13	10	поручитель		2,270
14	8	Рассрочка		2,267
15	20	доп_доход		1,577
16	11	проживание		1,430
17	19	телефон		1,352
18	9	сем_положение		0,481
19	14	Тек_кредит		0,000
20	18	иждивенцы		0,000

Рисунок 21 – Значимость атрибутов

Из таблицы видно, что наличие Кредитоспособность в первую очередь зависит от баланса текущего счета, во вторую от суммы кредита.

3.5.2 Нейронная сеть

Нейронные сети – это вычислительные системы с взаимосвязанными узлами, которые работают подобно нейронам в человеческом мозге. Используя алгоритмы, они могут распознавать скрытые закономерности и корреляции в необработанных данных, кластеризовать и классифицировать их, а со временем - постоянно изучать и совершенствовать.

Искусственные нейронные сети (ANN), часто называемые «нейронными сетями», представляют собой метафору мозга для обработки информации.

Эти модели являются биологически вдохновленными вычислительными моделями. Они состоят из взаимосвязанной группы искусственных нейронов и обрабатывают информацию с использованием вычислительного подхода.

В нейронных сетях нейроны объединяются в слои, при этом выходы нейронов предыдущего слоя являются входами нейронов следующего слоя. В каждом слое нейроны выполняют параллельную обработку данных.

Применение нейронных сетей в интеллектуальном анализе данных очень широко. Они обладают высокой приемной способностью к шумным данным и высокой точностью. Детальный анализ данных на основе нейронных сетей детально исследован. Было показано, что нейронные сети являются очень многообещающими системами во многих приложениях для прогнозирования и классификации бизнеса.

Банки используют нейронные сети для выявления мошенничества, проведения кредитного анализа и автоматизации услуг финансовых консультантов.

Выходным параметром является Кредит (Кредитоспособность), а входными все остальные параметры. Далее в настройках количество слоев и нейронов – 2 нейрона в 1 слой. Граф нейронной сети приведен рисунке 22.

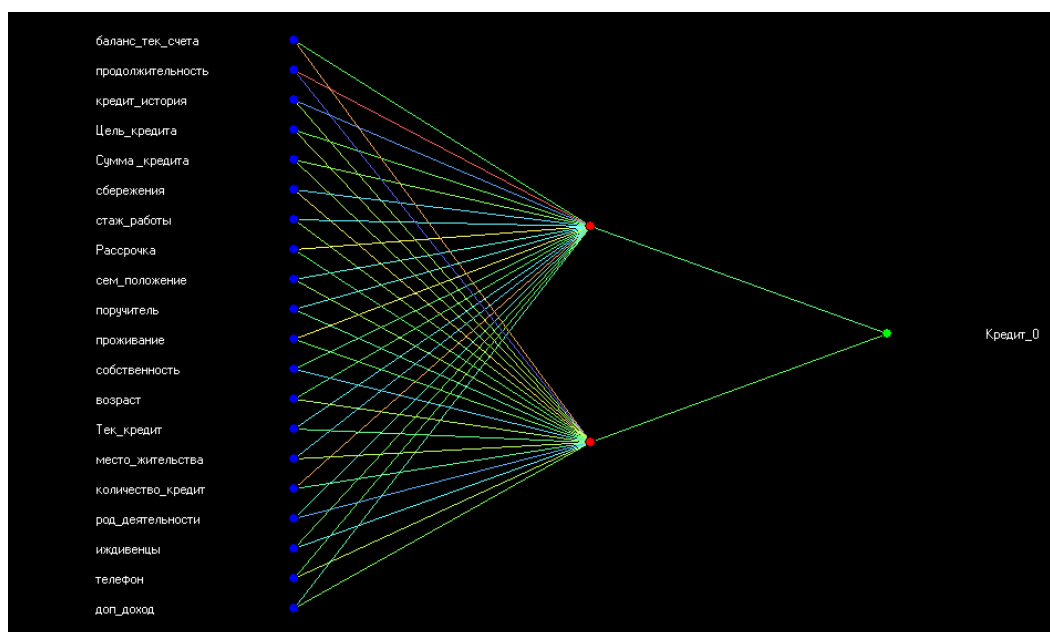


Рисунок 22 – Граф нейронной сети

Основываясь на данном результате нейронной сети получим таблицу сопряженности, представленную на рисунке 23.

Фактически	Классифицировано		
	False	True	Итого
False	111	189	300
True	12	688	700
Итого	123	877	1000

Рисунок 23 – Таблица сопряженности нейронной сети

Из таблицы видно, что классы классификатор определяет с удовлетворительной точностью. Ошибочными оказались 21%.

3.5.3 Карты Кохонена

Самоорганизующиеся карты Кохонена – мощный самообучающийся механизм кластеризации, позволяющий отобразить результаты в виде компактных и удобных для интерпретации двумерных карт.

Это нейронная модель, которая представляет собой метод визуализации данных. Этот метод имеет любые характеристики: количество классов задается пользователем, и он адаптирован к большим данным, для проецирования многомерного входного пространства в выходное пространство, обычно одномерное или двумерное.

Алгоритм обучения Кохонена – это итерационная процедура, которую можно рассматривать как алгоритм классификации, который группирует наблюдения по классам, сохраняя топологию пространства наблюдений.

Данный обработчик используется для поиска закономерностей в больших массивах данных. Это позволяет проводить разведочный анализ данных, отличающийся от классических статистических процедур, в ходе которых проверяется некоторый набор выдвинутых гипотез.

Преимуществом самоорганизующихся карт Кохонена является то, что при обучении используется метод обучения без учителя, то есть результат обучения зависит только от структуры входных данных.

Разбить исходный набор данных на подмножества и количество кластеров равным 2 т.к. кредит: хороший или плохой. Результат построения представлен на рисунке 24.

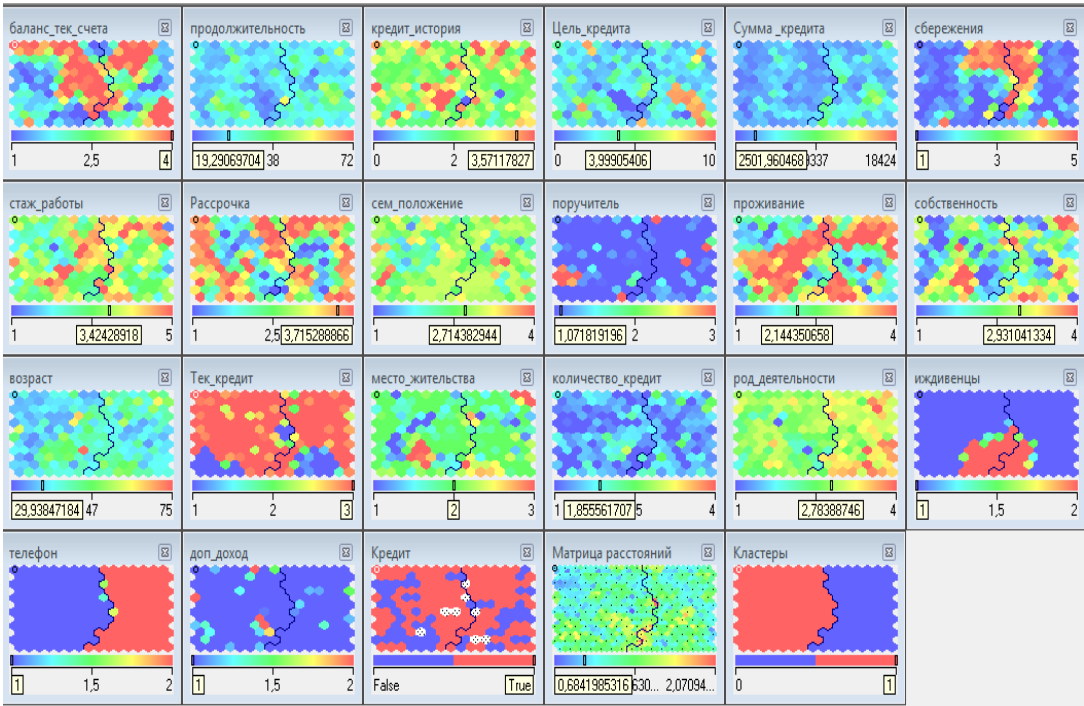


Рисунок 24 – Карты Кохонена

На нижней карте показано расположение определившихся кластеров. Работоспособность модели можно показать с помощью визуализатора «Таблица сопряженности» (рисунок 25).

Фактически	Классифицировано		
	False	True	Итого
False	176	124	300
True	85	615	700
Итого	261	739	1000

Рисунок 25 – Таблица сопряженности для Карт Кохонена

Из таблицы видно, что кластеры классификатор определяет с удовлетворительной точностью. Ошибочными оказались 21%.

3.5.4 Логистическая регрессия

Логистическая регрессия используется для нахождения линейных отношений между целью и одним или несколькими предикторами. Данный анализ позволяет моделировать отношения между зависимой переменной и одной или несколькими независимыми переменными. В интеллектуальном анализе данных этот метод используется для прогнозирования значений с учетом конкретного набора данных.

Регрессия является одним из самых популярных типов методов анализа данных, используемых в бизнесе, маркетинге на основе данных, финансовом прогнозировании.

Логистическая регрессия предназначена для получения прогноза непрерывных числовых переменных. Кроме того, реализованный в аналитической платформе Deductor доработанный вариант линейной регрессии позволяет также решать и задачу классификации.

Выходным параметром является Кредит (Кредитоспособность), а входными все остальные параметры, как это показано на рисунке 26.

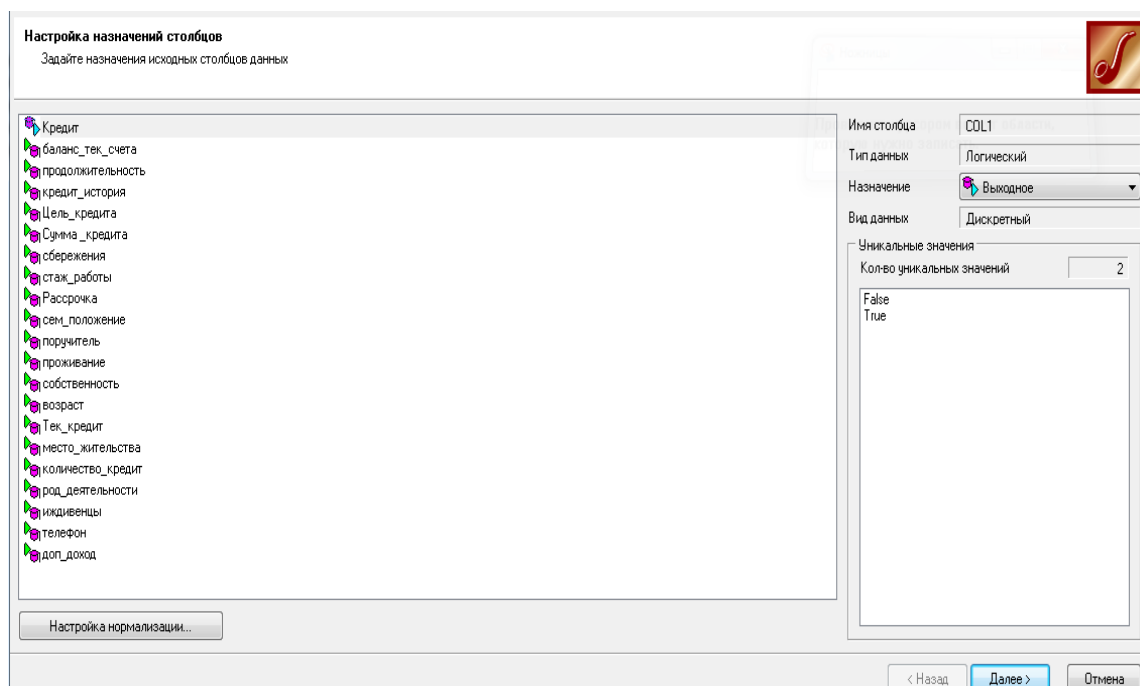


Рисунок 26 – Настройка входных/выходных параметров Логистической регрессии

Задаем настройки разбиения Обучающего и тестового множества, рисунок 27.

Способ разделения исходного множества данных Случайно

Столбец для разделения исходного множества

Множество	Размер		Порядок сортировки
	В процентах	В строках	
☒ Обучающее	90,00	900	По возрастанию
☒ Тестовое	10,00	100	По возрастанию
ИТОГО:	100,00	1000	

Рисунок 27 – Способы разделения исходных данных

Устанавливаем параметры отбора переменных, как это показано на рисунке 28.

Отбор переменных в регрессионные модели

Настройте параметры отбора переменных в регрессионные модели

Метод отбора переменных

☐ Полное включение (Enter)
☐ Прямой отбор (Forward)
☐ Обратное исключение (Backward)
☒ Последовательный отбор (Stepwise)

Вероятность F включения переменной 0,05

Вероятность F удаления переменной 0,1

Рисунок 28 – Параметры отбора переменных

Из таблицы сопряженности видно, что классы классификатор определяет с удовлетворительной точностью. Ошибочными оказались 26,8%. (рисунок 29)

	Классифицировано			
Фактически	False	True	Итого	
False	94	206	300	
True	62	638	700	
Итого	156	844	1000	

Рисунок 29 – Таблица сопряженности для Логистической регрессии

На рисунке 30 представлен отчет о регрессии.

Регрессия "1.1"							
Множ. коэффициент корреляции, R	Коэффициент детерминации, R^2	Скоррект. коэффициент детерминации	Стандартное отклонение	Размер выборки	Метод отбора переменных		
0,4157	0,1728	0,1710	0,4166	900	Прямой отбор		
Таблица дисперсионного анализа (ANOVA)							
Источник	Сумма квадратов, SS	Число степеней свободы, df	Средние квадраты, MS	F-критерий	Значимость		
Регрессия	32,5224	2	16,2612	93,6980	0,0000		
Ошибки	155,6732	897	0,1735				
Сумма	188,1956	899					
Коэффициенты регрессии							
	Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t-критерий	Значимость	Доверительный интервал (95%)	
	Значение	Ошибка				Значение	Ошибка
(Константа)	0,5364	0,0415		12,9378	0,0000	0,4551	0,6178
"баланс_тек_счета" (X0)	0,1265	0,0111	0,3471	11,3969	0,0000	0,1047	0,1483
"продолжительность" (X1)	-0,0077	0,0012	-0,2037	-6,6875	3,990E-11	-0,0100	-0,0055

Рисунок 30 – Отчет о регрессии

Из отчета видно, что в уравнение включено всего две переменных: Баланс текущего счета и продолжительность.

Коэффициент детерминации равен 0,172, что говорит о низком качестве рассмотренной модели.

Далее на рисунке 31 представлены ROC- кривая, которая графически отображает чувствительность (процентное отношение неплательщиков, правильно спрогнозированных как неплательщики) в зависимости от 1-й специфичности (процентное отношение тех, кто ошибочно был классифицирован как неплательщик), или отношение истинного положительного показателя по сравнению с ложным положительным показателем.

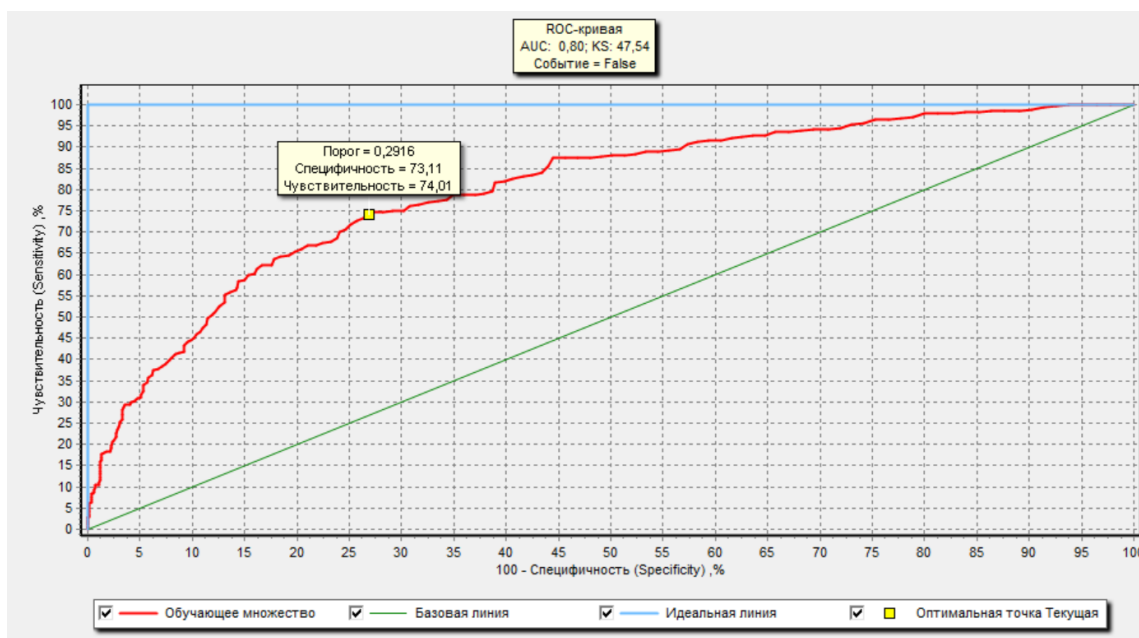


Рисунок 31 – ROC-кривая

Базовая линия указывает, что в каждой точке линии процент неплательщиков, правильно спрогнозированных как неплательщики (истинно положительный показатель), эквивалентен проценту неплательщиков, ошибочно отнесенных к неплательщикам (ложноположительный показатель).

3.6 Работоспособность полученных моделей

Работоспособность полученных моделей можно проверить при помощи инструмента визуализации «Что-если». В широком смысле в экономике и бизнесе это процесс изучения и оценки альтернативных стратегий. В Data

Mining это подход, позволяющий экспериментально исследовать поведение аналитических моделей (нейронных сетей, деревьев решений и т.д.) для различных входных воздействий.

Иными словами, данный тип анализа позволяет моделировать любые гипотетические ситуации, которые могут быть сколь угодно далеки от реальности, но представлять интерес с точки зрения перспективы.

Для проверки модели дерева решений необходимо вручную ввести какое-либо значение для переменной. Так как баланс текущего счета является самым значимым атрибутом, установим его в значения 2 и 3 и сравним выходные поля при этих значениях. На рисунке 32 представлены результаты при разных значениях.

Входные	
9.0 баланс_тек_счета	3
9.0 продолжительно...	18
9.0 кредит_история	4
9.0 Цель_кредита	2
9.0 Сумма_кредита	1049
9.0 сбережения	1
9.0 стаж_работы	2
9.0 Рассрочка	4
9.0 сем_положение	2
9.0 поручитель	1
9.0 проживание	4
9.0 собственность	2
9.0 возраст	21
9.0 Тек_кредит	3
9.0 место_жительства	1
9.0 количество_кред...	1
9.0 род_деятельности	3
9.0 иждивенцы	1
9.0 телефон	1
9.0 доп_доход	1
Выходные	
0/1 Кредит	True

Входные	
9.0 баланс_тек_счета	2
9.0 продолжительно...	18
9.0 кредит_история	4
9.0 Цель_кредита	2
9.0 Сумма_кредита	1049
9.0 сбережения	1
9.0 стаж_работы	2
9.0 Рассрочка	4
9.0 сем_положение	2
9.0 поручитель	1
9.0 проживание	4
9.0 собственность	2
9.0 возраст	21
9.0 Тек_кредит	3
9.0 место_жительства	1
9.0 количество_кред...	1
9.0 род_деятельности	3
9.0 иждивенцы	1
9.0 телефон	1
9.0 доп_доход	1
Выходные	
0/1 Кредит	False

Рисунок 32 – Сравнение выходных параметров

Как видно из рисунка, условие если баланс_тек_счета > 2,5, тогда кредитоспособность хорошая. А при баланс_тек_счета < 2,5 кредитоспособность плохая.

Для проверки работоспособности нейросети, можно сравнить ее с Картами Кохена. Чтобы сравнить две модели, нужно выбрать поле с необычным для него значением самого значимого атрибута. Пускай Баланс текущего счета =2, и Сумма_кредита=1000. На рисунке 33 слева расположен результат нейросети, а справа Карты Кохена.

Поле	Значение
Входные	
9.0 баланс_тек_с...	2
9.0 продолжительно...	18
9.0 кредит_история	4
9.0 Цель_кредита	2
9.0 Сумма_кредита	1000
9.0 сбережения	1
9.0 стаж_работы	2
9.0 Рассрочка	4
9.0 сем_положение	2
9.0 поручитель	1
9.0 проживание	4
9.0 собственность	2
9.0 возраст	21
9.0 Тек_кредит	3
9.0 место_жительства	1
9.0 количество_кред...	1
9.0 род_деятельности	3
9.0 иждивенцы	1
9.0 телефон	1
9.0 доп_доход	1
Выходные	
0/1 Кредит	False

Входные	
9.0 баланс_тек_счета	2
9.0 продолжительно...	18
9.0 кредит_история	4
9.0 Цель_кредита	2
9.0 Сумма_кредита	1000
9.0 сбережения	1
9.0 стаж_работы	2
9.0 Рассрочка	4
9.0 сем_положение	2
9.0 поручитель	1
9.0 проживание	4
9.0 собственность	2
9.0 возраст	21
9.0 Тек_кредит	3
9.0 место_жительства	1
9.0 количество_кред...	1
9.0 род_деятельности	3
9.0 иждивенцы	1
9.0 телефон	1
9.0 доп_доход	1
Выходные	
0/1 Кредит	False

Рисунок 33 – Сравнение результатов нейросети и кластеризации

Несмотря на такое необычное значение параметра нейросеть и Карты Кохена определили наличие плохой кредитоспособности правильно, что наглядно демонстрирует, как работают данные модели и используют значимость атрибутов.

3.7 Пример расчета кредитоспособности для потенциального заемщика

Расчет будет проводится на клиенте, который является потенциальным заемщиком. Вначале нам необходимо узнать все данные клиента с помощью анкетирования.

В таблице 3 собраны данные клиента, после анкетного опроса.

Таблица 3 – Данные о клиенте

ФИО	Иванов Иван Иванович
Дата рождения	01.01.1981
Паспортные данные	11 22 334566, 01.01.2001, г. Москва
Телефон моб. /домашний	+79123456789, 123456
Адрес по месту регистрации	г. Рязань, ул. Московская 1-2
Адрес по месту жительства	г. Москва, ул. Рязанская 2-1
Срок проживания	10 лет
Основание для проживания	съемная квартира
Пол	мужской
Семейное положение	женат
Количество детей	3
Образование	высшее
Наименование организации	ООО «ПромСервис»
Должность	руководитель отдела
Тип занятости	постоянная занятость
Род занятий	наемный сотрудник
Стаж работы	5 лет
Юр. адрес организации	г. Москва, ул. Первомайская 10, офис 2
Рабочий телефон	987654
Основной доход	75000
Доп. доход	отсутствует
Наличие автомобиля	отсутствует
Наличие недвижимости	отсутствует
Сумма кредита	300000
Срок кредита	8 месяцев
Цель кредита	другое

Далее собираем и структурируем все данные в одну таблицу. Результат представлен на рисунке 34.

баланс_т	продолжит_кредит_и	Цель_кре	Сумма_к	сбережен	стаж_раб	Рассрочка	сем_пол	поручите	проживае	собствен	возраст	Тек_кред	место_жи	количество	род_деят	иждивен	телефон	доп_доход
отсутствует	8 мес.	погашен_и_другое	300000	отсутствует	5 лет	35%	женат	отсутствует	10 лет	отсутствует	39	отсутствует	съемная	отсутствует	работающ	3	79123456789	отсутствует

Рисунок 34 – Очищенные данные

Затем следующий шаг преобразование данных к единой базисной системе. Таким образом у нас получаются преобразованные и готовые для анализа данных. Результат представлен на рисунке 35.

баланс_тек_счета	продолжит	кредит_и	Цель_кре	Сумма_к	сбережен	стаж_рабо	Рассрочка	сем_поло	поручите	прожива	собствен	возраст	Тек_кред	место_жи	количество	род_деят	иждивен	телефон	доп_доход
1	8	4	0	300000	1	4	1	3	1	4	1	39	3	2	2	2	1	1	2

Рисунок 35 – Преобразованные данные

Следующий этап, непосредственно анализ в Deductor методом Деревья решений. На рисунке представлены результаты анализа.

Поле	Значение
Входные	
9.0 баланс_тек_счета	1
9.0 продолжительно...	8
9.0 кредит_история	4
9.0 Цель_кредита	0
9.0 Сумма_кредита	300000
9.0 сбережения	1
9.0 стаж_работы	4
9.0 Рассрочка	1
9.0 сем_положение	3
9.0 поручитель	1
9.0 проживание	4
9.0 собственность	1
9.0 возраст	39
9.0 Тек_кредит	3
9.0 место_жительства	2
9.0 количество_кред...	2
9.0 род_деятельности	2
9.0 иждивенцы	1
9.0 телефон	1
9.0 доп_доход	2
Выходные	
0/1 Кредит	True

Рисунок 36 – Результаты анализа

Как показано на рисунке, данному заемщику банк может выдать кредит.

Выводы

Данные были протестированы методами интеллектуального анализа данных, а именно использовались методы – деревья решений, нейронная сеть, Карты Кохонена и Логистическая регрессия.

В результате проведенного анализа, по полученным моделям можно сделать выводы:

- наиболее работоспособная модель «Дерево решений»;
- наименее всего допущено ошибок при классификации;
- наиболее значимым параметром при определении кредитоспособности является баланс текущего счета и сумма кредита.

Таким образом, на основании анализа дерева решений показали наиболее качественную оценку, определив большее количество правильной классификации. Поэтому именно этот метод следует использовать в данном алгоритме.

С помощью полученного алгоритма была рассчитана кредитоспособность клиента, результат анализа определил потенциального клиента как «хорошего» и была одобрена выдача кредита.

Заключение

В результате выполнения диссертационной работы были проанализированы возможные методы и методики оценки кредитоспособности физических лиц. Были описаны достоинства и недостатки каждой методики и метода. Были изучены методы ИАД, используемых в кредитном скоринге.

Были собраны данные потенциальных заёмщиков, для дальнейшего нахождения оптимального метода классифицирования. В результате исследования наилучший результат показал метод Деревья решений, а также была доказана эффективность применения ИАД в оценке кредитоспособности физических лиц.

На основании этого анализа был разработан алгоритм, на основе методов интеллектуального анализа данных, который позволит делать качественную оценку кредитоспособности клиентов.

Полученный алгоритм решает одну из ключевых задач управления кредитными рисками в банковской сфере, а именно задачу оценки кредитоспособности заемщиков.

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 клиентов от набора показателей, используемых для оценки кредитоспособности, что позволяет предупреждать возможности возникновения кризисной ситуации, сохранить устойчивое финансовое состояние и повысить эффективность предпринимательской деятельности.

Список используемых источников

1. Коновалова, Н., Кристовская И., Кудинская, М. Assessing creditworthiness in consumer credit/ Н. Коновалова, И. Кристовская, М. Кудинская //Polish Journal of Management Studies. –2017. –№13(2). – С. 90– 100.
2. Масленников, А. Анализ и оценка кредитоспособности заемщика/ А. Масленников, – 2016. – 531с.
3. Loretta, J., What Is the Point of Credit Scoring?/ J. Loretta //Business review. –1997. – С. 3-16.
4. Потребительский кредитный скоринг с помощью Data Mining [Электронный ресурс]. URL: statsoft.ru/solutions/ExamplesBase/tasks/detail. (дата обращения 25.05.2020)
5. Devaney, S. TheUsefulness of Financial Ratios as Predictors of Household Insolvency/ S. Devaney // Two Perspectives. Financial Counseling and Planing. –1994. –№5. – С. 15–24.
6. Решение SAS для создания системы кредитного скоринга в банках [электронный ресурс]. URL: <https://www.sas.com/offices/europe/russia/whitepapers/CrSc.pdf> (Дата обращения: 18.04.2020).
7. Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems/ R. A. Fisher //Annals of Eugenic. – 1936. –№7. – С. 179–188.
8. Creditworthiness [Электронный ресурс]. URL: <https://www.investopedia.com/terms/c/credit-worthiness.asp> (дата обращения 30.05.2020)
9. Hand, D. J., Henley, W. E. Statistical Classification Methods in Consumer Credit Scoring/ D. J. Hand, W. E/ Henley //Journal of the Royal Statistical Society, Series A (Statistics in Society). –1997. –№160. –С. 523–541.
10. Рынок потребительского кредитования продолжит бурный рост [Электронный ресурс]. URL: <http://minfin.com.ua/2018/03/30/738661> (дата обращения 19.06.2020)

11. Дремова, У. В. Совершенствование подходов к оценке кредитоспособности заемщиков при долгосрочном кредитовании/ У. В. Дремова // Финансы и кредит. — 2015. — № 11. — С.15–23.
12. Комаров, Д. С. Применение современных технологий для оценки кредитоспособности физических лиц / Д. С. Комаров // Молодой ученый. — 2017. — № 5 (139). — С. 177-180.
13. Ефимов А.М. Современные методы оценки кредитоспособности физических лиц [Электронный ресурс].URL: <http://www.lawmix.ru/bx/101723> (дата обращения 30.06.2020)
14. Enas, G. G., Choi, S. C. Choice of the Smoothing Parameter and Efficiency of k-nearest Neighbor Classification/ G. G. Enas, S. C. Choi //Computers and Mathematics with Applications. – 1986. –№12. – С. 235–244.
15. Семенов, С. В. Применение инноваций в кредитовании на потребительском рынке/ С. В. Семенов // Карельский научный журнал. – 2017. – №6. – С. 10-15.
16. Рзаев, Р.Р. Оценивание кредитоспособности физических лиц с применением нечеткой логики / Р.Р. Рзаев, А.А. Алиев // Проблемы управления и информатики. – 2017. – № 1. – С. 114 – 127.
17. Lin, A.S. A Note on the concordance correlation coefficient / A.S. Lin // Biometrics. – 2012. – № 56. – С. 324 – 325.
18. Сеньковская, О. С. Пути снижения кредитных рисков / О. С. Сеньковская // Азимут научных исследований: экономика и управление. – 2014. – №1. – С. 99 – 102.
19. Lin, A.S. Statistical Tools for Measuring Agreement / A.S. Lin, W. Wu. – New York: Springer, 2012. – 173 с.
20. Тельнов, Г.В. Подход к формированию итоговой оценки кредитоспособности на основе взвешенных коэффициентов оцениваемых признаков / Г.В. Тельнов // Вестник АГУ. – 2015. – №1 (154). – С. 119 – 127.

21. Заде, Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений / Заде Л.А. – М.: Мир, Математические машины и системы. –2018. –№ 2. – С. 14-17.
22. Рзаев, Р.Р. Аналитическая поддержка принятия решений в организационных системах / Р.Р. Рзаев– Saarbruchen (Germany): Palmerium Academic Publishing, 2016. – 306 с.
23. Lukasiewicz, J. On three-valued logic: Selected Works / Lukasiewicz J.; Ed. by Borkowski L. – Amsterdam: NorthHolland Publishing Company. – 1970. – №14. – С. 87 – 88.
24. SAS Credit Scoring for Banking [электронный ресурс]. URL: http://www.sas.com/ru_ru/industry/banking/credit-scoring.html (Дата обращения: 15.06.2020).
25. Андрейчиков, А.В. Андрейчикова, О. Н. Анализ, синтез, планирование решений в экономике / А.В. Андрейчиков, О.Н. Андрейчикова. – М.: Финансы и статистика, 2000. – 368 с.
26. Помазанов, М. Количественный анализ кредитного риска / М. Помазанов // Банковские технологии. – 2014. – № 2. – С. 22-28.
27. Ковалев, П.П. Пути повышения результативности кредитного риск-менеджмента в коммерческом банке: автореф. дис. ... на соиск. уч. степени и канд. экон. наук: спец. 08.00.10 – Финансы, денежное обращение и кредит. – М.: Государственный НИИ системного анализа Счетной палаты РФ. – 2016. – 24 с.
28. Крючков С.А. Оценка кредитоспособности заемщика. Основные показатели оценки [Электронный ресурс]. URL: <http://www.tusur.ru/filearchive/reports-magazine/2004-9-1/208.pdf> (Дата обращения: 19.05.2020).
29. Desay, V., Crook, J. N., Overstreet, G. A. A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment/ V. Desay, J. N. Crook, G. A. Overstreet // European Journal of Operational Research. –1996. – №95. – С. – 24–37.

30. Altman, E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy/ E. Altman //Journal of Finance. – 1968, –№23. С. 589 – 609.
31. Chatterjee, S., Barcun, S. A Nonparametric Approach to Credit Screening. Journal of American Statistical Association. –1970. – № 65. – С. 150 – 154.
32. West, D. Neural Network Credit Scoring Models/ D. West // Computers & Operations Research. – 2000. – №27. – С. 1131–1152.
33. Campbell, T. S., Dietrich, J. K. The Determinants of Default on Insured Conventional Residential Mortgage Loans/ T. S. Campbell, J. K. Dietrich // Journal of Finance. – 1983. – 38. –С. 1569–1581.
34. DataMining в системе управления знаниями. [Электронный ресурс]. URL: <http://www.smart-edu.com/upravlenie-znaniyami/izvlechenie-znaniy-dta-miningv-sisteme-upravleniya-znaniyami.html> (дата обращения 19.06.2020)
35. Макарычев, П.П. Оперативный и интеллектуальный анализ данных / П.П. Макарычев, В.Б. Механов, А.Ю. Афонин // Федер. агентство по образованию, Гос. образовательное учреждение высш. проф. образования «Пензенский гос. ун-т». – Пенза, 2010.
36. Ежемесячный дайджест «Обзор банковского сектора России» за январь 2017 г., с. 8. [Электронный ресурс]. URL: <http://www.kreditprombank.com/upload/content/510/January2017!.pdf> (дата обращения 19.06.2020)
37. Барсегян, А.А. «Методы и модели анализа данных: OLAP и Data Mining»/ А.А. Барсегян. – СПб.: БХВ-Петербург, 2014. – 336 с.
38. Дюк, В. «Data Mining. Учебный курс»/ В. Дюк. – СПб.: Питер, 2013. – 386 с.
39. Паклин, Н.Б., Орешков, В.И. «Бизнес-аналитика: от данных к знаниям»/ Н.Б. Паклин, В.И. Орешков. – СПб.: ПИТЕР, 2013. – 706 с.
40. Рудой Н. Системы оценки кредитоспособности: особенности автоматизации [Электронный ресурс] // URL:

<http://www.softlab.ru/upload/iblock/f5e/f5e673f33c266f129598fa0ed42ff71b.pdf>
(дата обращения: 12.05.2020).

41. Ли, Т. С., Чиу, С.С., Чоу, Ю.С., Лу, С.Дж. Анализ кредитоспособности клиента с использованием дерева классификации и регрессии и многомерной адаптивной регрессии / Т. С. Ли, С.С. Чиу, Ю.С. Чоу, С.Дж. Лу // Вычислительная статистика и анализ данных. – 2006. – №50. – С. 1113 – 1130.

42. Леонард, К. Дж. Разработка показателей качества кредитного скоринга для потребительских кредитных заявок/ К. Дж. Леонард // Международный журнал управления качеством и надежностью. – 1995. – №12 (4). – С. 79-85.

43. Риммер, Дж. Современные изменения в кредитном скоринге / Дж. Риммер // Credit Control. – 2005. – №26 (4). – С. 56-60.

44. Йобас, М.Б., Крук, Дж. Н., Росс, П. Кредитный скоринг с использованием нейронных и эволюционных методов / М.Б. Йобас, Дж. Н. Крук, П. Росс // Математический журнал ИМА, применяемый в бизнесе и промышленности. – 2000. – №11. – С. 111-125.

45. Вендел, С. Кредитный скоринг: лучшие практики и подходы/ С. Вендел // Обзор коммерческого кредитования. – 2003. – №18 (3). – С. 4-7.

46. Сурду, А. Анализ кредитного скоринга с использованием алгоритмов интеллектуального анализа данных/ А. Сурду // IT. – 2011. – №101. – С. 1 – 12.

47. Никулин, А.Д. Аналитическая платформа «Дедуктор» – применение в информационных системах экономики: методические указания / А.Н. Никулин, И.В. Чернышев. – Ульяновск: УлГТУ, 2012. – 37 с.

48. Коркина, Е.И. Методика оценки кредитоспособности физического лица с помощью скоринговой модели/ Е. И. Коркина // Современные научные исследования: актуальные вопросы, достижения и инновации. – 2020. – С. 191 – 195.

49. Коркина, Е.И. Обзор методов проектирования моделей кредитного скоринга/ Е. И. Коркина // Студенческие научные исследования. – 2020. – С.64 – 67.

50. Паклин Н.Б. Применение логистической регрессии в скоринге. [Электронный ресурс]. URL: <http://www.basegroup.ru/> (Дата обращения 19.06.2020).

Приложение А
ДИПЛОМЫ
(рекомендуемое)



Рисунок А.1 – Диплом за I место в XII Международной научной конференции



Рисунок А.2 – Диплом за I место в Международной научной конференции