TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
Отделение школы (НОЦ) – Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы

Подготовка исходных данных для построения кредитного скоринга

УДК 004.6.056.523:336.77.067

Студент

Группа	ФИО	Подпись	Дата
8ПМ7И	Инхиреева Татьяна Александровна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Губин Е.И.	к.ф.-м.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
старший преподаватель	Потехина Н.В.			

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Горбенко М.В.	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Губин Е.И.	к.ф.-м.н.		

Томск – 2019 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код результата	Результат обучения
Общие по направлению подготовки 09.04.04 «Программная инженерия»	
Р1	Способность проводить научные исследования, связанные с объектами профессиональной деятельности
Р2	Способность разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах
Р3	Способность составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты
Р4	Способность проектировать системы с параллельной обработкой данных и высокопроизводительные системы
Р5	Способность осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных
Р6	Способность осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем
Р7	Способность организовывать промышленное тестирование создаваемого программного обеспечения
Профиль «Технологии больших данных»/ «Big data solutions»	
Р8	Способность исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях
Р9	Способность понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях
Р10	Способность применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) – Отделение информационных технологий
Период выполнения: весенний семестр 2018 /2019 учебного года

Форма представления работы:
магистерская диссертация

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.02.2019	*Раздел 1.Объект и методы исследования*	*20*
10.03.2019	*Раздел 2. Расчеты и аналитика*	*20*
10.04.2019	*Раздел 3. Результаты проведенного исследования*	*20*
01.06.2019	*Раздел 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение*	*20*
16.05.2019	*Раздел 5. Социальная ответственность*	*20*

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

СОГЛАСОВАНО:

Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
Отделение школы (НОЦ) – Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

Студенту:

Группа	ФИО
8ПМ7И	Инхиреевой Татьяны Александровны

Тема работы:

Подготовка исходных данных для построения кредитного скоринга	
Утверждена приказом директора	№1436/с от 25.02.2019

Срок сдачи студентом выполненной работы:	06.06.2019

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	**Банковские данные о кредитоспособности заемщиков.** **Классификационная модель логистической регрессии, используемая для кредитного скоринга.**

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. Аналитический обзор литературных источников; 2. постановка задачи исследования; 3. разработка методики; 4. реализация методики; 5. выбор программного обеспечения; 6. обсуждение результатов выполненной работы; 7. финансовый менеджмент; 8. социальная ответственность; 9. заключение.
Перечень графического материала	

Консультанты по разделам выпускной квалификационной работы

Раздел	Консультант
Социальная ответственность	Горбенко Михаил Владимирович, доцент ООД ШБИП, к.т.н.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Потехина Нина Васильевна, старший преподаватель ОСГН
Обязательное приложение на английском языке	Диденко Анастасия Владимировна, доцент ОИЯ ШБИП, к.ф.н.

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Литературный обзор

Введение

Объект и предмет исследования

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Инхиреева Татьяна Александровна		

РЕФЕРАТ

Выпускная квалификационная работа 101 с., 21 рис., 23 табл., 22 источника, 3 прил.

Ключевые слова: анализ данных, подготовка данных, очистка данных, логистическая регрессия, кредитный скоринг, оценка кредитоспособности.

Объект исследования – данные о кредитоспособности заемщиков.
Предмет исследования – методика обработки данных для кредитного скоринга.

Цель работы – разработка и исследование методики обработки данных для кредитного скоринга.

В процессе исследования проводилось изучение, анализ, тестирование и сравнение различных существующих методик подготовки данных, реализация алгоритма методики с помощью таких инструментов, как Python, SAS, SAS Enterprise Miner.

В результате исследования показано, что предложенная методика обработки данных позволяет повысить точность оценки кредитоспособности. Сделан вывод о применимости метода.

Основные конструктивные, технологические и технико-эксплуатационные характеристики: разработанная методика позволяет производить оценку кредитоспособности потенциального заемщика.

Область применения: разработанная методика может применяться для повышения точности кредитного скоринга в банках.

Оглавление

Введение

Одной из главных задач, стоящих перед кредитно-финансовыми организациями, является оценка рисков выполнения заемщиком его кредитных обязательств. Анализ рисков предполагаемого заемщика производится на основе анализа анкетных данных. На практике оценка риска невозврата кредита для конкретного заемщика производится двумя способами – экспертной оценкой либо с помощью скоринговых систем [1].

Для построения скоринговой карты могут быть использованы различные методы прогнозирования, такие как нейронные сети, деревья решений, линейная и логистическая регрессия. На практике, наиболее часто используемой моделью является логистическая регрессия. В данной работе рассматривается методика подготовки данных для построения скоринговых карт на базе логистической регрессии.

Эффективность анализа данных во многом зависит от качества исходных данных, поэтому подготовка данных является очень важным шагом при анализе данных во многих областях науки и техники. Зачастую его игнорируют полностью или частично, что отрицательно отражается на результатах анализа. Полученные на этапе сбора данные обычно содержат недостатки: пропуски, дубли, недопустимые значения, невозможные комбинации значений и т.д. Данные могут иметь разный формат, обладать нежелательными для дальнейшего анализа свойствами (мультиколлинеарность, корреляция, распределение, отличное от нормального). Даже самые передовые методы не показывают хороших результатов на некачественных данных.

Актуальность работы заключается в том, что подготовка данных, несмотря на ее очевидную значимость для дальнейшего анализа, зачастую игнорируется полностью или частично – пропускаются некоторые шаги, в то время как каждый шаг увеличивает точность предсказания, и, следовательно, уменьшает финансовые потери банка при выдаче кредита. В работах [2], [3] даны

общие рекомендации по обработке данных. В [1] описывается методика построения скоринговой карты, включающая подготовку данных.

Цель данной работы – разработка и исследование методики обработки данных для кредитного скоринга.

Для достижения поставленной цели необходимо выполнить следующие **задачи**:

1) подобрать и изучить литературу по данной теме;
2) исследовать методики обработки данных для кредитного скоринга.
3) реализовать методику с помощью Python, SAS, SAS Enterprise Miner;
4) проверить правильность работы реализаций;
5) произвести сравнительный анализ результатов работы различных реализаций.

Объектом исследования в качестве тестовой задачи рассматривается данные о кредитоспособности заемщиков.

Предметом исследования является методика обработки данных для кредитного скоринга.

Практическая значимость: разработанная методика может применяться для повышения точности кредитного скоринга в банках.

По итогам исследования сделан доклад на VI международной молодежной научной конференции «Математическое и программное обеспечение информационных, технических и экономических систем».

Опубликована статья Inkhireeva T.A. Data mining classification techniques for credit scoring in banks // Математическое и программное обеспечение информационных, технических и экономических систем: материалы VI международной молодежной научной конференции, Томск, 24-26 мая 2018 г. - Томск: ТГУ, 2018 - С. 362-365

1 Объект и методы исследования

1.1 Описание метода построения скоринговой карты с использованием модели логистической регрессии.

Самым распространенным методом оценки кредитоспособности заемщика в банках является кредитный скоринг. Кредитный скоринг (скоринг заявок) представляет собой автоматизированную систему на основе предсказательной математической модели, которая использует кредитную историю банка для прогнозирования вероятности того, что потенциальный заемщик вернет кредит вовремя [4]. Прогноз строится на основе информации о кредитной истории, социальном-демографических параметрах, данных о запрашиваемом кредите. В настоящее время банки уделяют особое внимание анализу кредитных рисков в связи с учащением случаев невозврата по кредитам и мошенничества. По данным Единого федерального реестра юридически значимых сведений о фактах деятельности юридических лиц, индивидуальных предпринимателей и иных субъектов экономической деятельности (Федресурс) за 2018 год число банкротств физических лиц в России выросло на 50%, а сумма требований кредиторов превысила 762 миллиарда рублей [5]. По данным Объединенного кредитного бюро, по состоянию на 1 января 2019 года более 748 тысяч россиян были отнесены к категории потенциальных банкротов (заемщики с кредитами более 500 тысяч рублей и просрочкой платежей более 90 дней), что на 6% больше по сравнению с прошлым годом [6]. При построении скоринговой модели требуется не только определить на основе присвоенного балла, стоит ли давать заемщику кредит или нет, но и определить минимальный балл для выдачи кредита.

Большинство банков создают скоринговые модели самостоятельно, используя собственные данные, собранные за предыдущие года, либо пользуются готовые решения, базирующиеся на обобщённых данных о заемщиках нескольких банков. В обоих случаях методы построения моделей являются коммерческой тайной.

Наиболее распространенной предсказательной моделью для построения скоринговых карт является логистическая регрессия. Данная модель позволяет оценить вероятность возврата кредита для конкретного заемщика. В модели бинарной логистической регрессии целевая переменная $y \in \{0, 1\}$ подчиняется распределению Бернулли и отражает кредитоспособность заемщика.

Математически модель логистической регрессии выражает зависимость логарифма шанса от линейной комбинации независимых переменных

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1,1} + \ldots + + b_k x_{i,j} + \varepsilon_i \quad (1)$$

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

x_{ij} — значение j-ой независимой переменной;

b_0 — независимая константа модели, b_j — параметры модели;

ε_i — компонент случайной ошибки.

Уравнение (1) отражает линейную зависимость вероятности наступления просрочки по кредиту в зависимости от значений независимых переменных.

1.2 Постановка задачи

Требуется произвести подготовку данных для решения задачи бинарной классификации потенциальных заемщиков банка методом логистической регрессии. Исходными данными к работе являются исторические данные о кредитоспособности, содержащие 24 переменных, одна из которых – целевая, и 3000 наблюдений. Данная выборка сбалансирована по целевой переменной, то есть количество плательщиков и неплательщиков. Неплательщиками считаются те заемщики, которые не осуществляли запланированные выплаты по кредиту в течение 90 дней.

В табл. 1.1 приведены переменные, характеризующие заемщиков.

Таблица 1.1. Список переменных

Имя переменной	Расшифровка	Тип переменной
TITLE	Характер собственности жилья	Категориальная

CHILDREN	Количество детей	Числовая
PERS_H	Количество человек в домохозяйстве	Числовая
AGE	Возраст	Числовая
TMADD	Количество месяцев проживания на текущем месте жительства	Числовая
TMJOB1	Количество месяцев на текущей работе	Числовая
TEL	Количество контактных номеров телефона	Числовая
NMBLOAN	Количество кредитов в данном банке	Числовая
FINLOAN	Отсутствие невыплаченных кредитов	Бинарная
INCOME	Доход (в неделю в евро)	Числовая
EC_CARD	Обладание картой банка	Бинарная
INC	Заработная плата	Числовая
INC1	Разделение на 5 категорий по уровню заработной платы	Категориальная
BUREAU	Класс кредитного риска по оценке кредитного бюро	Категориальная
LOANS	Количество займов вне банка	Числовая
REGN	Регион проживания	Категориальная
CASH	Размер запрошенного кредита	Числовая
PRODUCT	Цель кредита	Категориальная
RESID	Арендатор или собственник жилья	Категориальная
NAT	Национальность	Категориальная
PROF	Индустрия	Категориальная
CAR	Тип средства передвижения	Категориальная
CARDS	Тип кредитной карты	Категориальная
GB	Целевая переменная	Бинарная

Целевой переменной является GB, принимающая значение 0, если заемщик не имеет просрочки (хороший) и 1, если заемщик имеет просрочку более 90 дней (плохой).

1.3 Выбор и описание метода решения поставленной задачи

По результатам последних опросов KDnuggets (2014 г.), 43% опрошенных лиц использует методологию анализа данных CRISP-DM, 8,5% – методологию SEMMA, 3,5% – собственную методологию организации, 27,5% – свою

собственную методологию, другими методологиями пользуется 17,5% опрошенных. Не пользуются никакой методологией 0% опрошенных [7].

Две лидирующих методологии в целом очень похожи, однако CRISP-DM заслужила большую популярность, как более полная и детальная, чем SEMMA. Каждая из этих методологий включают этап подготовки данных, имеющий в обоих случаях довольно общий рекомендательный характер, что приводит к необходимости создания более четкой и подробной методики подготовки данных для классификации с помощью логистической регрессии.

1.3.1 Методология CRISP-DM

CRISP-DM – стандартный межотраслевой процесс Data Mining, состоящий из шести этапов, организованных в виде цикла. В настоящее время является наиболее популярной методологией. Согласно стандарту CRISP-DM включает в себя следующие этапы [8]:

1. Бизнес-анализ (Business understanding) – начальная фаза, на которой происходит определение бизнес-целей и вырабатываются требования к результатам.
2. Анализ данных (Data understanding). Вторая фаза начинается со сбора данных, включает в себя описание, изучение и проверку качества данных.
3. Подготовка данных (Data preparation). Фаза подготовки данных включает формирование выборок, конструирование признаков, очистку, интеграцию и форматирование данных.
4. Моделирование (Modeling). На этапе моделирования осуществляется выбор, обучение и оценка качества моделей.
5. Оценка результатов (Evaluation) включает в себя оценку процесса, полученных результатов и определение последующих действий.
6. Внедрение (Deployment). Данный шаг предполагает внедрение модели, мониторинг и получение обратной связи.

Рисунок 1.1 – Методология CRISP-DM

1.3.2Методология SEMMA

Методология SEMMA, созданная SAS Institute, является альтернативой CRISP-DM. SEMMA состоит из пяти этапов [9]:

1. Выборка (Sample) – формирование начального набора данных для моделирования (dataset), который должен быть достаточно большим, чтобы содержать достаточную информацию для извлечения, и в то же время ограниченным, чтобы его можно было эффективно использовать.
2. Исследование (Explore) – выявление ассоциаций, визуальный и интерактивный статистический анализ, понимание данных путем обнаружения ожидаемых и непредвиденных связей между переменными, а также отклонений с помощью визуализации данных.
3. Изменение (Modify) – применение методов выбора, создания и преобразования переменных при подготовке к моделированию:

кластерный анализ, преобразование, фильтрация и замещение информации.

4. Моделирование (Model) – применение методов построения и обработки моделей интеллектуального анализа данных: искусственные нейронные сети, деревья принятия решений, регрессионный анализ и т. д.
5. Оценка (Assess) – сравнение результатов моделирования между собой и с планируемыми показателями, анализ надежности и полезности созданных моделей.

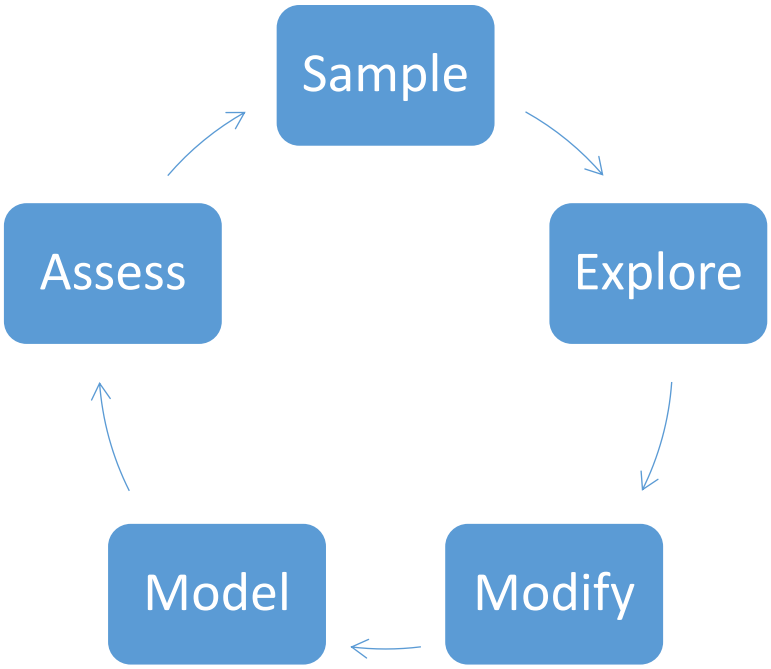

Рисунок 1.2 – SEMMA

Данная методология была разработана для использования в среде SAS Enterprise Miner, поэтому ее этапы ориентированы на возможности данного программного пакета.

Методология SEMMA скорее является сводом рекомендаций, чем сводом жестких правил, и менее детально проработана, чем CRISP-DM. Это объясняет ее меньшую популярность.

1.3.3Разработанная методология подготовки данных

1. **Разбиение данных**

Для построения адекватной модели и проверки ее точности исходные исторические данные о заемщиках необходимо разбить на две или три независимые выборки, в зависимости от количества обучаемых моделей.

Если предполагается обучение одной предсказательной модели, рекомендуется разбиение на две выборки – тренировочную и тестовую. На тренировочной выборке происходит обучение моделей, то есть оптимизация их параметров. На тестовой выборке осуществляется оценка качества модели. Соотношение числа наблюдений в тренировочной выборке и тестовой чаще всего составляет 70-80% и 30-20% соответственно. Это соотношение определяется количеством исходных данных. Модель, обученная на небольшом количестве тестовых данных, обладает большой дисперсией, т. е. ее результаты на разных выборках сильно различаются. При среднем объеме выборки (тысячи или десятки тысяч наблюдений) соотношение 70% к 30% и 80% к 20% считается стандартным решением. Если количество наблюдений измеряется сотнями тысяч, есть смысл уменьшить тренировочную выборку и увеличить тестовую, особенно если предсказательная модель требует большого объема вычислений. При небольшом объеме исходных данных (сотни-тысячи наблюдений) стоит использовать кросс-валидацию [10].

Кросс-валидация является методом разбиения исходных данных на тренировочную и тестовую выборку в условиях недостатка данных. Кросс-валидация позволяет снизить дисперсию модели. В процессе кросс-валидации данные сначала разбиваются на тренировочную и тестовую выборку. Тренировочная выборка разбивается на k частей. Обучение происходит на k-1 частях, оставшаяся часть используется для валидации [11].

При обучении нескольких моделей используется разбиение на три выборки – тренировочную, валидационную и тестовую. Этим решается проблема переобучения – получение оптимистически смещенной оценки модели. На валидационной выборке производится сравнение результатов работы нескольких моделей и выбор лучшей. Стандартным соотношением размера трех выборок является 60, 20 и 20% для данных среднего объема исходных данных.

Данные выше рекомендации о выборе соотношения размера двух выборок применимы и в данном случае [12].

Вышеперечисленные схемы разбиения данных представлены на рис. 1.3.

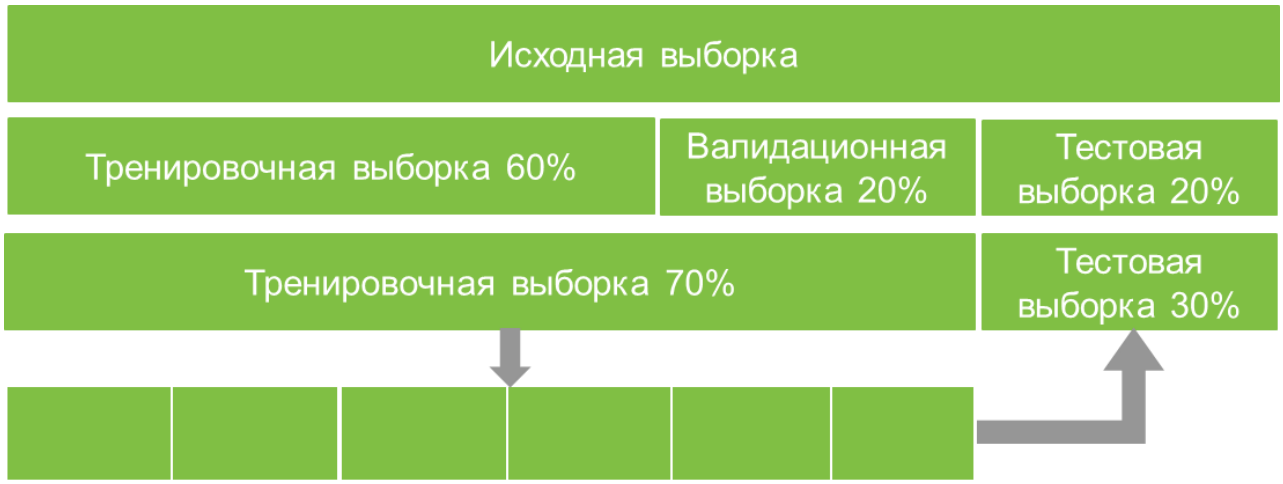

Рисунок 1.3 – Разбиение данных

При разбиении данных возникает проблема репрезентативности полученных выборок. Все выборки должны обладать тем же соотношением классов целевой переменной, что и в исходной выборке.

2. Очищение данных

Очищение данных включает в себя удаление дублирующихся записей, исправление ошибочных и противоречивых данных, обработку выбросов и пропущенных значений. Решение этих проблем может значительно улучшить качество прогнозной модели.

Дублирующиеся наблюдения

Наличие одинаковых одинаковых наблюдений влияет на коэффициенты регрессии, увеличивая дисперсию модели, поэтому дублирующиеся наблюдения должны быть найдены и удалены из анализа [2].

Выбросы

Выбросы в данных – это аномальные значения, выделяющиеся из общей выборки. Логистическая регрессия чувствительна к выбросам, поэтому их обработка является очень важным шагом подготовки данных.

Простейший способ определения выбросов в числовой переменной – считать выбросом все наблюдения, которые не укладываются в заданные квантили. Графически этот подход реализован в виде диаграмм размаха (ящик с усами). Диаграмма размаха показывает медиану или среднее, межквартильный размах, максимальное и минимальное значение, выбросы.

Значительно труднее выявить многомерные выбросы. Двумерные выбросы можно выявить с помощью диаграммы рассеяния (точечной диаграммы). Построение точечных диаграмм для всех пар переменных помогает выявить двумерные выбросы. Наблюдения, являющиеся выбросами более высоких размерностей, можно выявить с помощью алгоритмов изолирующего леса и dbscan и многих алгоритмов кластеризации.

Выбросы в категориальных переменных могут быть обнаружены с помощью гистограмм.

При небольшом количестве выбросов можно удалить их из анализа или заменить средним, или модой. При большом количестве выбросов следует выделить их в отдельную выборку для проведения анализа, поскольку это может свидетельствовать о появлении нового феномена в данных. Помочь справиться с выбросами в числовой переменной может применение некоторых преобразований (min-max нормализация) и дискретизация [3].

Коррекция противоречивых данных

Неверные и противоречивые значения могут появиться на этапе ввода, передача и сбора данных в результате опечаток, программных ограничений (ограничение на длину переменной, ограничение размера буфера), различных форматов записи данных (Санкт-Петербург, Ст.-Петербург).

Неверными могут оказаться редкие категориальных переменных, экстремальные (рост: 250 см.) или необычные (заработная плата: -1 руб.) значения числовых переменных. Выявить такие значения можно с помощью гистограмм, ящика с усами или диаграмм рассеяния. Противоречивые данные (пол: мужской, беременность: да) можно выявить, используя релевантные

логические правила. Для выявления некоторых ошибочных и противоречивых данных может понадобиться эксперт в предметной области [3].

Неверные и противоречивые данные представляют проблему потому, что алгоритм логистической регрессии предполагает, что все исходные данные – корректные, и строит модель в соответствие с этим предположением, что приводит к неверным результатам. При выявлении неверных или противоречивых значений, необходимо исправить или удалить соответствующее наблюдение из анализа.

Обработка пропусков

Пропуски в данных могут быть обусловлены множеством причин: необходимые данные не всегда могут быть доступны (информация о клиенте), данные могут отсутствовать потому, что считались не нужными в определенный момент времени. Пропуски также могут возникнуть из-за технических проблем. Данные могут быть удалены по причине противоречивости. Многие методы анализа данных, в том числе логистическая регрессия, не способны работать с пропущенными данными, поэтому пропуски необходимо тем или иным образом устранять: удалять наблюдения, содержащие пропуски либо заполнять их.

В случае необходимости заполнения поля на этапе ввода данных, пропущенные значения кодируют некоторым заменяющим значением, выбранным так, чтобы оно не было похожим на типичное для переменной значение. Типичные заменяющие значения для пропусков в данных представлены в табл. 1.2.

Таблица 1.2 – Типичные значения, заменяющие пропущенные данные

Код	Описание
Null, пустая строка “”	для числовой или категориальной переменной
0	числовая переменная, значение которой никогда не равно нулю
-1	числовая переменная, которые принимает только положительные значения
99, 999	числовая переменная, значения которой могут быть меньше 100, 1000 и т.д.

-99, -999	числовая переменная, способная принимать отрицательные значения
U, UU	категориальная переменная
000000, XXXXXX	почтовый индекс
11.11.11	дата
00000000000	номер телефона

В зависимости от причин, породивших пропущенные данные, пропуски могут иметь различное распределение. Понимание этого распределения может помочь выбрать алгоритм заполнения пропущенных данных. Механизмы появления пропущенных данных делятся на три категории:

Missing Completely at Random (MCAR) – механизм формирования пропусков, при котором вероятность пропуска для каждой записи одинакова. В таком случае игнорирование или исключение записей, содержащих пропущенные данные, не ведет к искажению результатов.

Missing at Random (MAR) – чаще всего данные пропущены не случайно, а ввиду некоторых закономерностей. Пропуски относят к MAR, если вероятность пропуска может быть определена на основе другой имеющейся в наборе данных информации (пол, возраст, занимаемая должность, образование), не содержащей пропуски. В таком случае удаление или замена пропусков на значение «Пропуск», как и в случае MCAR, не приведет к существенному искажению результатов.

Missing not at Random (MNAR) – механизм формирования пропусков, при котором данные отсутствуют в зависимости от неизвестных факторов. MNAR предполагает, что вероятность пропуска могла бы быть описана на основе других атрибутов, но информация по этим атрибутам в наборе данных отсутствует. Как следствие, вероятность пропуска невозможно выразить на основе информации, содержащейся в наборе данных.

На практике может быть не очевидно, к какой категории отнести пропущенные данные, потому что механизм их появления может быть просто неясен. Механизм MCAR может быть выявлен с помощью t-критерия Стьюдента

или критерия Литтла [13]. Данные, содержащие менее 5% пропусков, можно считать MCAR. Для данных, содержащих от 5 до 50% пропусков, необходимо определить механизм их возникновения и в соответствие с этим выбирать стратегию их заполнения. Переменные, содержащие более 50% пропущенных значения, следует удалить из анализа. MAR и MNAR могут быть выявлены вручную, зачастую для этого требуется помощь эксперта в предметной области. Большая часть методов заполнения пропусков предполагает работу с данными MCAR и MAR, поскольку их присутствие не влияет существенным образом на результат [14].

Наиболее распространенными методами заполнения пропусков в числовых переменных являются: заполнение константой (нулем, средним, модой, медианой, последним наблюдением) [15], заполнение из распределения, заполнение с помощью модели (нейросеть, дерево решений).

Для обработки отсутствующих значений в категориальных переменных используется, оздание отдельной категории для пропущенных данных

Создание бинарной переменной-индикатора

3. Трансформация

Дискретизация

Дискретизация непрерывных переменных может быть предпочтительна, если распределение переменной мультимодально, имеет тяжелые хвосты, выбросы или пропущенные значения. В таких случаях дискретизованная версия непрерывной может упростить для анализа сложные нелинейные зависимости.

Алгоритм дискретизации числовой переменной состоит из трех шагов. Сначала значения числовой переменной разбиваются на несколько групп по квантилям. Для каждой группы вычисляется показатель вес категорий Weight of Evidence (*WOE*) по формуле

$$WOE_i = \ln\left(\frac{d_{i1}}{d_{i2}}\right),$$

где d_{i1}, d_{i2} – относительные частоты плохих и хороших заемщиков в i-й группе дискретизованной переменной; $i=1,\ldots,k$,

k – число категорий переменной.

Далее полученные показатели весов категорий анализируются, происходит объединение соседних категорий и расчет показателей *WOE* повторяется. При дальнейшем объединении категорий руководствуются следующими правилами: в каждой группе должно находиться не меньше 5% от всех валидных наблюдений переменной; не должно быть групп с количеством «плохих» или «хороших» кредитов, равным 0; процент «плохих» заемщиков и WOE должны в достаточной мере отличаться по получаемым группам; значения показателей WOE должны иметь возрастающий или убывающий тренд при переходе от одной категории к другой. При укрупнении категорий помимо статистических критериев следует руководствоваться логикой, целесообразностью и возможностью такого объединения [1].

Нормализация (масштабирование)

Принято считать, что масштабирование переменных не влияет на логистическую регрессию, однако неизбежное применение регуляризации l1 и l2 привносит необходимость масштабирования переменных перед использованием логистической регрессии.

Наиболее распространенные методы масштабирования – стандартизация и min-max нормализация. Стандартизация используется в случае приближенности распределения переменной к нормальному, в противном случае предпочтительна min-mix нормализация. Методы масштабирования представлены в табл. 1.3.

Таблица 1.3 – Методы масштабирования

Метод масштабирования	**Формула**	**Диапазон**
Min-max нормализация	$\frac{x - x_{\min}}{x_{\max} - x_{\min}}$	[0, 1]
Стандартизация	$\frac{x - \mu_x}{\sigma_x}$	в большинстве случаев [-3, 3]

Нелинейное преобразование

Обычно для преобразования числовых переменных используют следующие виды преобразований: квадратное; кубическое; квадратный корень; натуральный или десятичный логарифм; экспоненциальное; величина, обратная квадратному корню; обратная величина. При использовании степенных преобразований ко всем значениям преобразуемой переменной могут добавлять константу для преобразования нуля или отрицательных значений. Такие преобразования количественных переменных могут привести к максимизации их связи с зависимой целевой переменной. Необходимое преобразование подбирается эмпирически, так чтобы полученная переменная наиболее точно описывала целевую переменную. Также часто используются относительные преобразования, например отношение суммы дохода к сумме задолженности [1].

4. Выбор переменных

Последним шагом перед непосредственным анализом данных является выбор переменных. На этапе выбора переменных отбрасываются неинформативные, избыточные переменные и переменные, которые не улучшают модель.

Мультиколлинеарность

Присутствие мультиколлинеарности в объясняющих переменных приводит к увеличению дисперсии модели логистической регрессии, получению неправильных знаков при оценке параметров модели, а также неустойчивости оценки параметров модели.

Для выявления мультиколлинеарности используется анализ корреляционной матрицы и статистика Variance Inflation Factor (VIF):

$$VIF = \frac{1}{1 - R_i^2},$$

где R_i – коэффициент детерминации регрессии i-й переменной на остальные объясняющие переменные.

Если показатель VIF больше пяти, это говорит о присутствии мультиколлинеарности. В этом случае необходимо удалить данную переменную из анализа либо использовать метод главных компонент для конструирования новых признаков вместо исходных;

Информативность

Предварительный анализ информативности объясняющих переменных, их влияния на целевую переменную помогает сократить количество рассматриваемых признаков.

Переменные, не имеющие взаимосвязи с целевой переменной (идентификационный номер клиента, фаза луны в день подачи заявки), должны быть удалены из анализа, также, как и переменные с дисперсией равной или близкой к нулю, что значит, что на исходной выборке она почти всегда принимает одно и то же значение.

Основными методами оценки информативности переменных являются критерий хи-квадрат и показатель информационного значения (IV).

1.4 Выбор и описание программной среды

Для реализации методики подготовки данных были выбраны три инструмента: Python 3 в оболочке Jupyter Notebook, SAS и SAS Enterprise Miner.

Python – язык программирования с относительно низким порогом вхождения, на данный момент является самым популярным языком для анализа данных, имеет множество открытых библиотек.

В Jupyter Notebook программный код размещается в серии ячеек – исполняемых или разметочных. Разметочные ячейки поддерживают LaTeX, что позволяет использовать в них математические выражения. Файлы, созданные в Jupyter Notebook имеют формат .ipynb, эквивалентный формату .json. Полученные файлы могут храниться в системе контроля версий.

В ходе выполнения работ были использованы следующие библиотеки:

1. NumPy (Numerical Python) – пакет, предоставляющий оптимизированные функции для работы с многомерными массивами данных.

2. Pandas (Panel Data) – специализированный пакет, основанный на NumPy для анализа панельных данных.
3. SciPy – библиотека, ориентированная на пользователей MATLAB, содержит инструменты для обработки сигналов, изображений, решения дифференциальных уравнений и других задач.
4. Matplotlib – библиотека для визуализации данных.
5. Seaborn – библиотека визуализации данных, основанная на matplotlib.
6. Scikit-learn – библиотека для машинного обучения.
7. Mlxtend – пакет, содержащий функции для анализа данных.
8. Missingno – пакет для визуализации и заполнения пропущенных значений.

SAS (Statistical Analysis System) – программный пакет, для обработки и анализа данных, лидер рынка бизнес-анализа.

SAS Enterprise Miner – пакет для анализа данных с графическим интерфейсом. Предоставляет простой способ построения моделей.

2 Расчеты и аналитика

2.1 Разработка алгоритма методологии подготовки данных

Диаграмма потока работ методологии подготовки данных представлена в приложении Б.

Исходные данные алгоритма: D – анкетные данные заемщиков, являются матрицей, содержащей m строк (наблюдений) и n столбцов (переменных).

5. Задание исходных данных:
   5.1. D – анкетные данные заемщика.
6. Разделение исходной выборки D:
   6.1. если предполагается выбор одной из нескольких моделей и n велико, разделение на тренировочную, валидационную и тестовую выборки;
   6.2. если предполагается использование одной модели или n недостаточно велико, разделение на тренировочную и тестовую выборку.
   6.3. если n очень мало, кросс-валидация.
7. Очищение:
   7.1. устранение дублирующихся строк;
   7.2. обработка выбросов:
      7.2.1. выявление одномерных выбросов;
      7.2.2. выявление многомерных выбросов;
      7.2.3. если число выбросов невелико, удаление их из анализа;
      7.2.4. если число выбросов велико, выделение их в отдельную выборку.
   7.3. коррекция противоречивых данных;
   7.4. обработка пропусков:
      7.4.1. анализ причин появления пропусков;
      7.4.2. если пропуски появились в результате ошибок при вводе данных (меньше 5% значений переменной пропущено), исключение пропусков из анализа;
      7.4.3. если в переменной содержится от 5% до 50% пропусков, выяснение причины и заполнение пропусков;
      7.4.4. если в переменной более 50% пропусков, удаление данной переменной из анализа.
8. Трансформация:
   8.1. дискретизация:
      8.1.1. разбиение значений непрерывных переменных на несколько групп по квантилям;
      8.1.2. вычисление WOE для каждой группы;

8.1.3. если одна из полученных групп гомогенна по целевой переменной или тренд WOE в различных группах меняется (с возрастающего на убывающий или наоборот), необходимо объединить соседние группы;

8.2. нормализация (масштабирование):

8.2.1. если распределение переменной близко к нормальному, стандартизация;

8.2.2. в противном случае min-max нормализация всех переменных к диапазону от 0 до 1;

8.3. нелинейное преобразование.

9. Выбор переменных:

9.1. мультиколлинеарность:

9.1.1. если VIF переменной больше пяти, необходимо удалить ее из анализа либо использовать метод главных компонент для конструирования новых признаков вместо исходных;

9.2. нормальность:

9.2.1. проверка переменных на нормальность с помощью. критериев Шапиро-Уилка, асимметрии и эксцесса, Д'Агостино и т. д., а также графическое исследование с помощью гистограмм;

9.2.2. приведение ненормально распределенных переменных к нормальному распределению.

9.3. информативность:

9.3.1. удаление переменных с почти нулевой дисперсией;

9.3.2. исключение переменных на основе критерия хи-квадрат и IV;

9.3.3. использование пошагового алгоритма включения или исключения переменных.

2.2 Программная реализация методики

Алгоритм предложенной методологии реализован в трех программных средах.

Все программные реализации считывают исходные данные из файла в формате .xls.

- Реализация в Python сделана с помощью веб-оболочки Jupyter Notebook. Состоит из одного файла в формате .ipynb.
- Реализация в SAS состоит из нескольких файлов в формате .sas, содержащих реализацию отдельных функций.

– Реализация в SAS Enterprise Miner содержит 589 файлов, автоматически генерируемых при исполнении программной реализации.

3 Результаты проведенного исследования

Рассмотрим результаты применения предложенной методики на данных о заемщиках.

3.1 Разделение исходной выборки

Поскольку количество исходных данных невелико (3000 наблюдений) и используется только одна предсказательная модель, исходная выборка разделена на две – тренировочную и тестовую в соотношении 70% к 30% соответственно, с использованием кросс-валидации. Полученные выборки стратифицированы по целевой переменной. По рис. 3.1 видно, что баланс классов в полученных выборках сохранен.

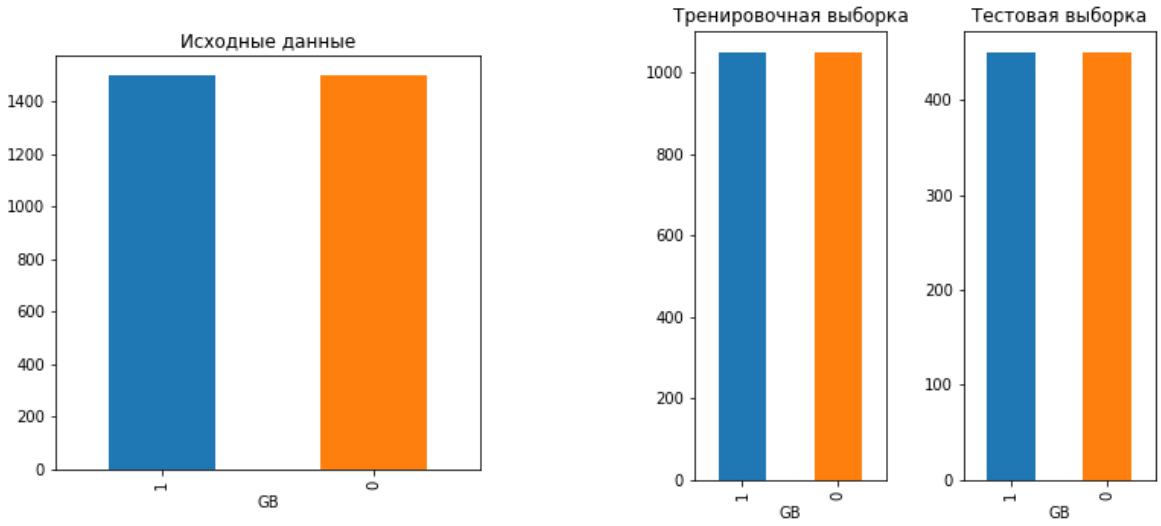

Рисунок 3.1 – Разделение данных

3.2 Очищение данных

Удаление дублирующихся строк

В данной выборке отсутствуют дублирующиеся строки.

Обработка выбросов

Анализ одномерных выбросов в числовых переменных с помощью диаграмм размаха (рис. 3.2) и гистограмм (рис.3.3) показывает, что переменные CHILDREN, PERS_H, INCOME, CASH содержат значения, которые могут быть ошибкой либо выбросом в данных. Переменные LOANS и AGE содержат выбросы, а в переменных TMADD и TMJOB1 пропущенные значения

закодированы значением 999. Экстремально большие значения переменных CASH и INCOME с большой вероятностью являются закодированными пропусками в данных, поскольку не согласуются с другими данными, что отражено на рис.3.4. Данные значения в переменных TMADD, TMJOB1, CASH и INCOME в количестве 93, 34 11 и 1 шт. соответственно были заменены на пропуски. Также из анализа удалено наблюдение с экстремальными значениями по переменным CHILDREN, PERS_H, которые согласуются между собой (23 ребенка и 25 проживающих в доме), и, следовательно, являются выбросом, а не ошибкой или кодом для пропуска, но могут сильно повлиять на коэффициенты регрессии.

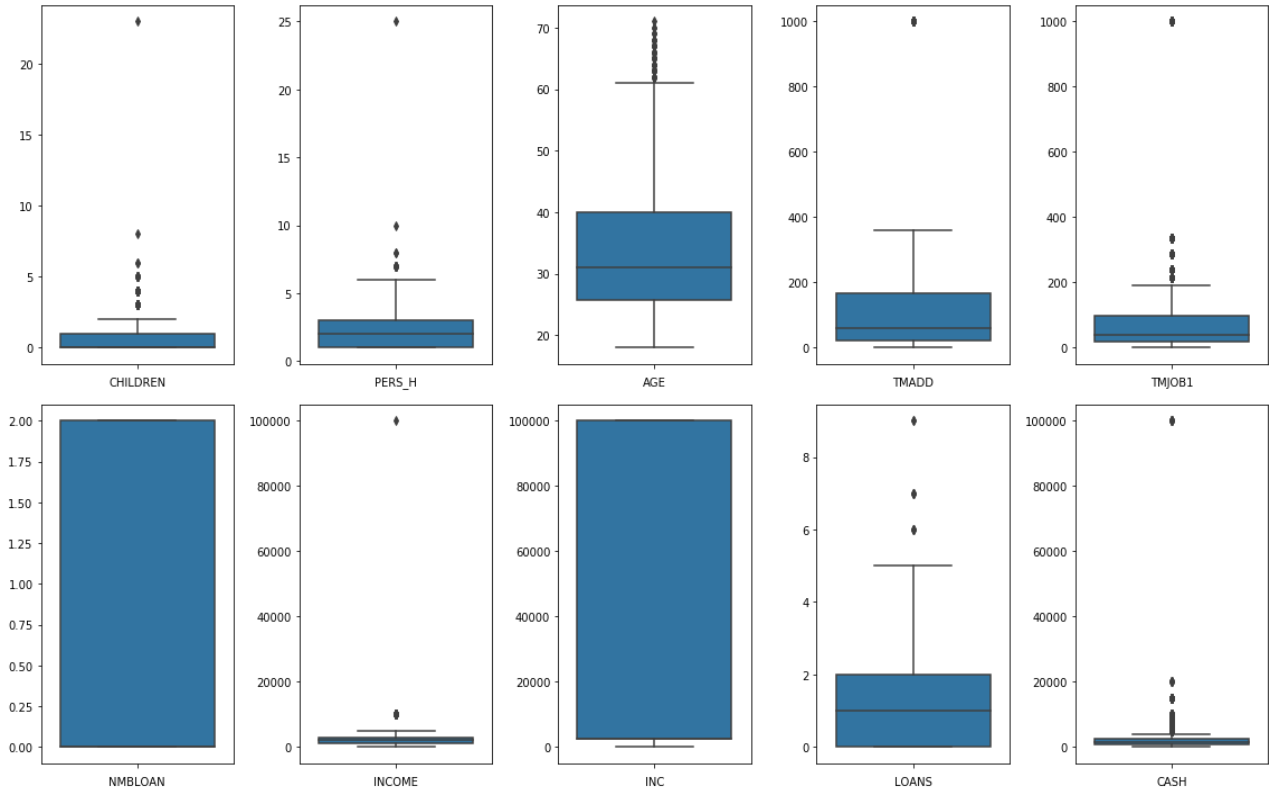

Рисунок 3.2 – Диаграммы размаха исходных данных

По рис. 3.3 можно сделать предположение, что ни одна из объясняющих переменных не распределена нормально.

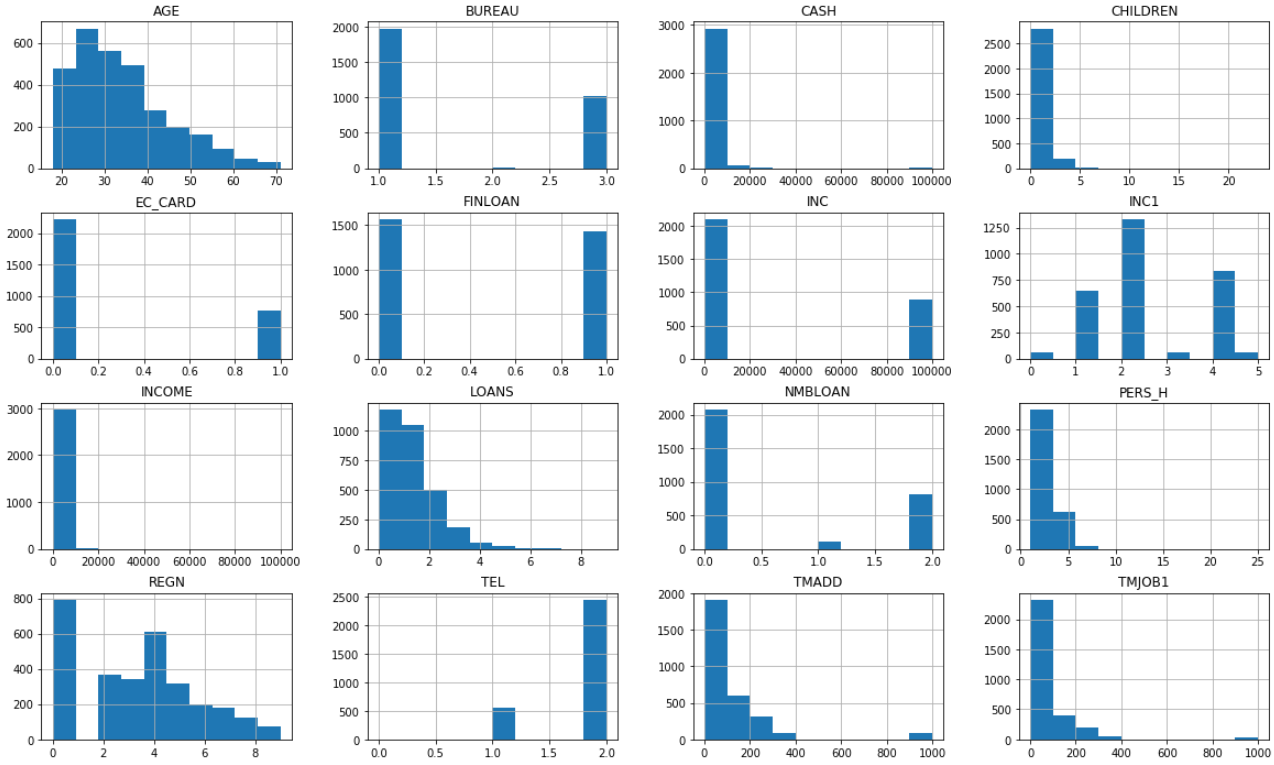

Рисунок 3.3 – Гистограммы исходных данных

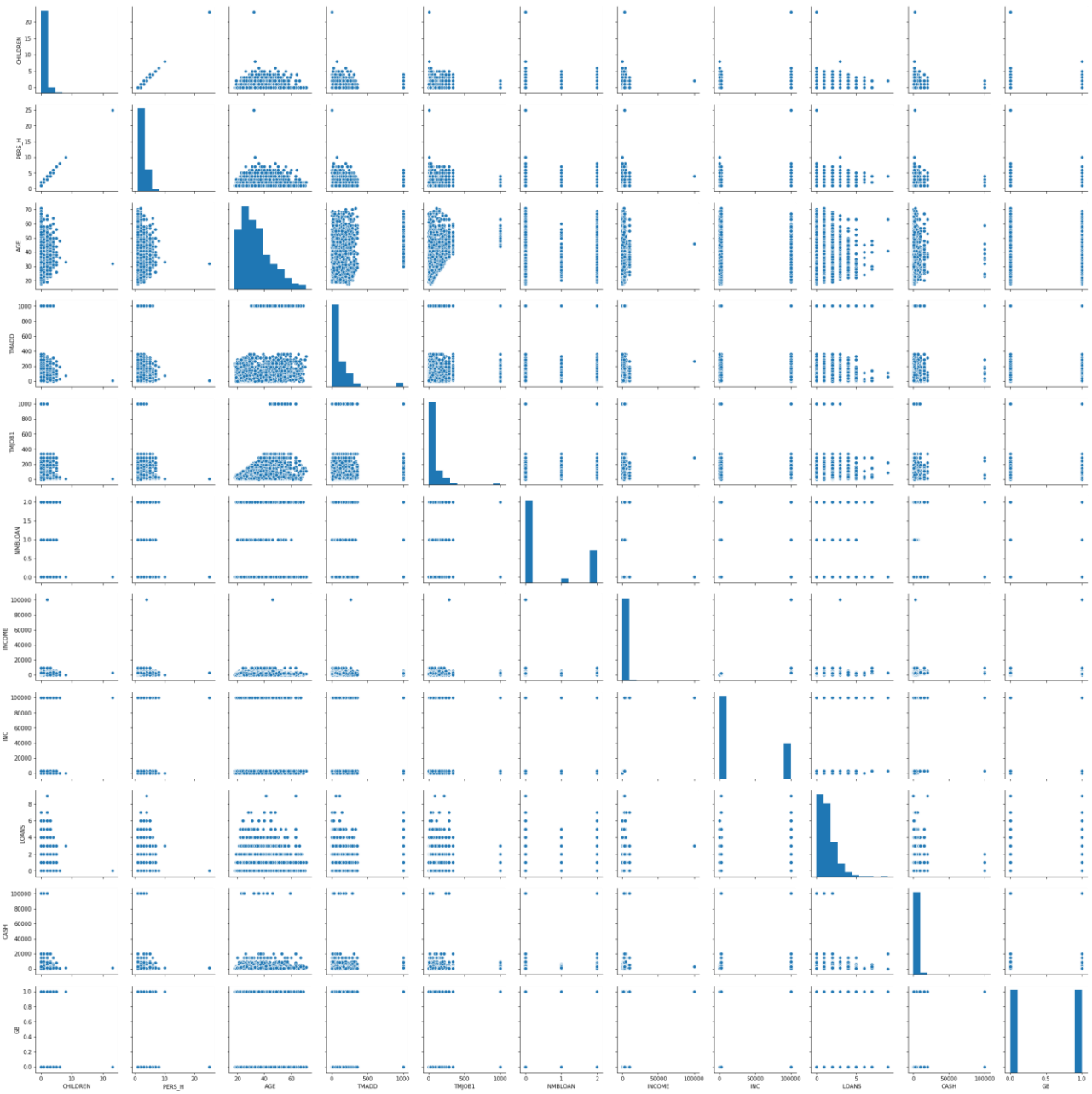

Рисунок 3.4 – Точечные диаграммы исходных данных

На рис. 3.5 представлены диаграммы размаха после удаления выбросов.

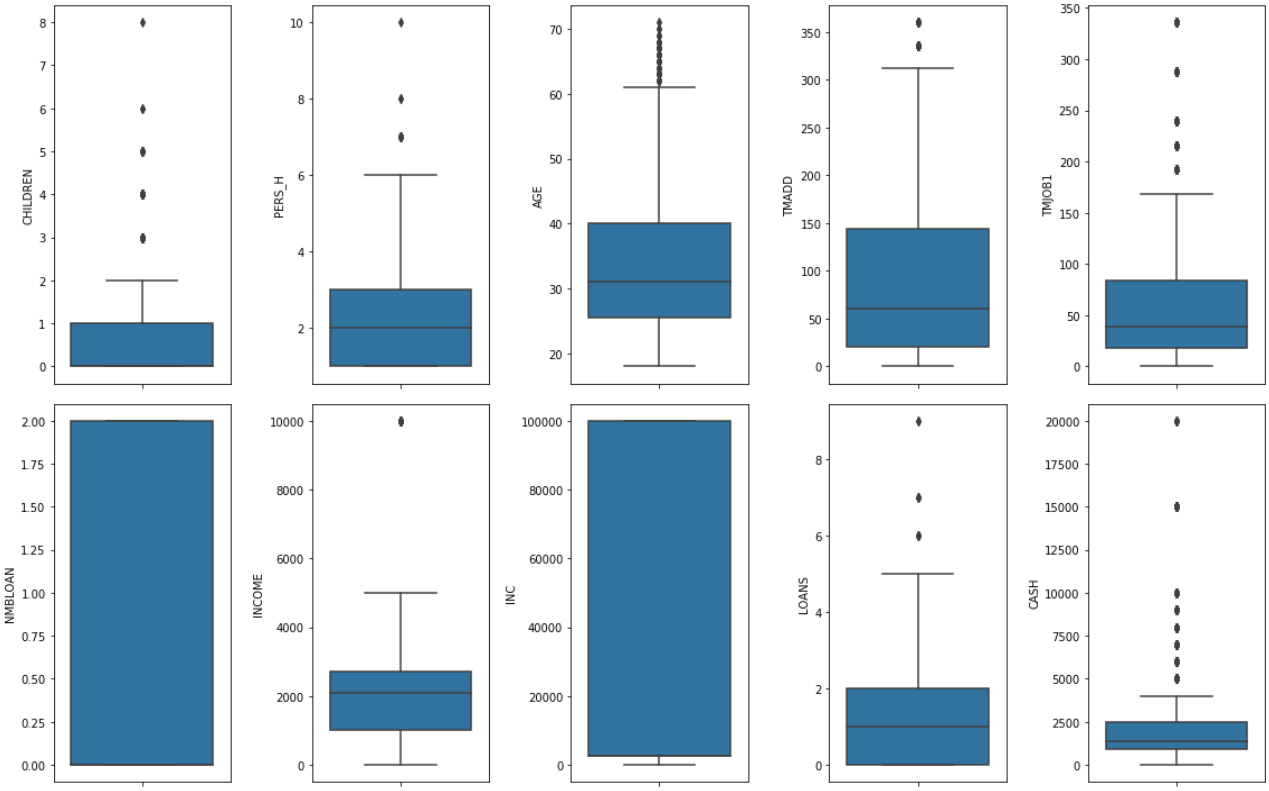

Рисунок 3.5 – Диаграмма размаха после обработки выбросов

Заметно, что на ней присутствуют и другие выбросы, но их удаление из анализа может привести к потере большого количества полезных данных.

По диаграмме частот категориальных переменных (рис. 3.6) видно, что в переменных PRODUCT, NAT, CAR, CARDS и TEL содержатся категории малой частотой.

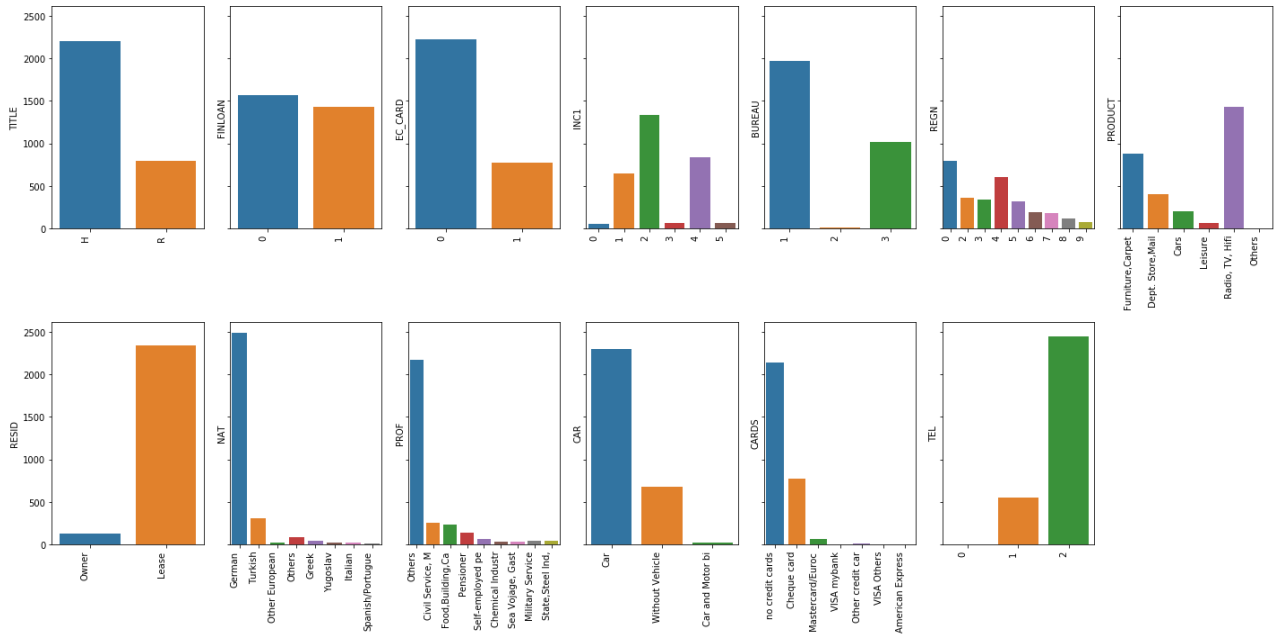

Рисунок 3.6 – Диаграмма частот исходных данных

Обработка пропусков

В исходной выборке содержится 687 пропусков. Пропуски локализованы в переменных TMADD, TMJOB, CASH, PRODUCT, RESID и PROF (рис. 3.7, 3.8). Причинами их появления скорее всего является отсутствие необходимых данных на этапе сбора. Поскольку в переменных содержится до 50% пропусков, применена стратегия их заполнения. В переменной RESID, содержащей 535 пропусков, создана отдельная категория. Отсутствующим значениям в переменных PRODUCT и PROF присвоена существующая категория Other. Отсутствующие значения числовых переменных TMADD, TMJOB, CASH заменены средним.

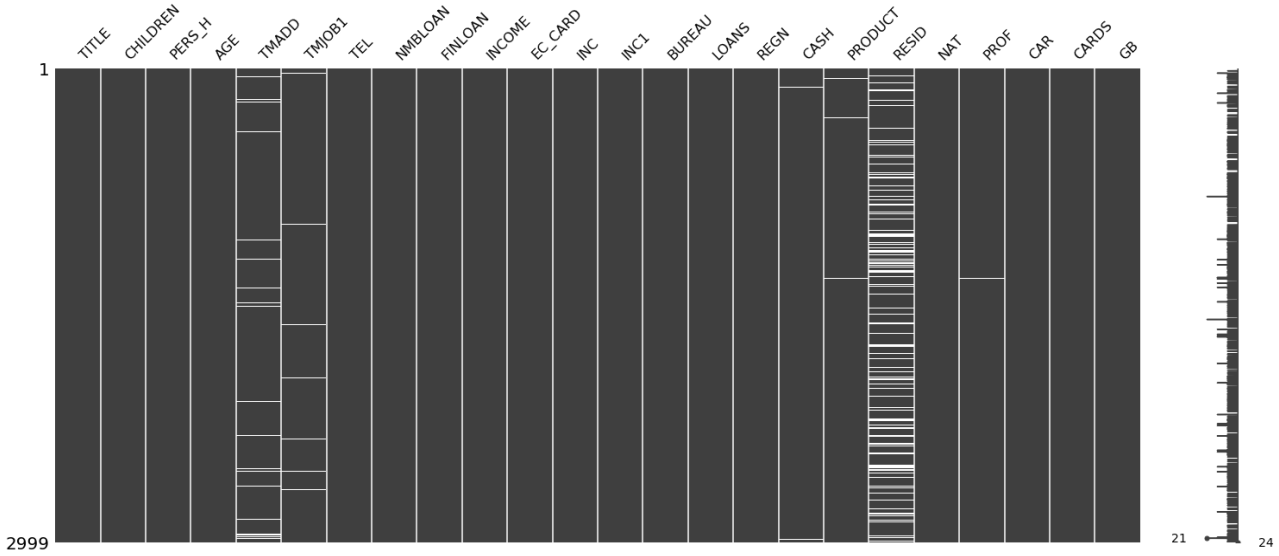

Рисунок 3.7 – Структура пропущенных данных

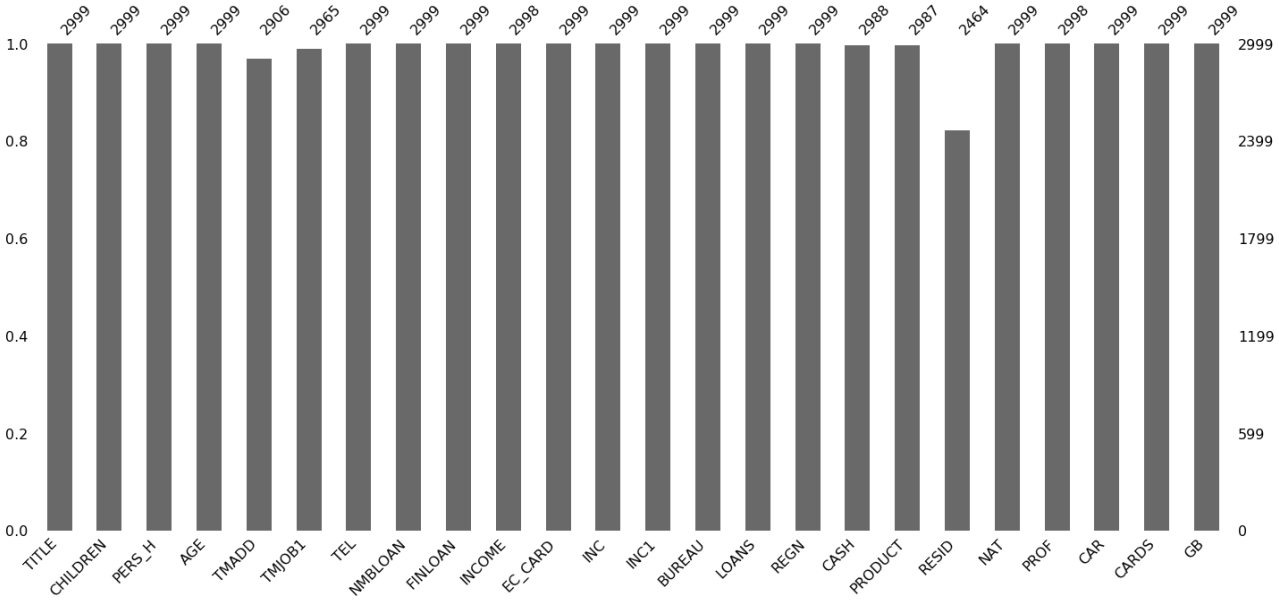

Рисунок 3.8 – Количество пропусков в данных по переменным

3.3 Трансформация данных

Дискретизация

Диапазон значений числовых переменных был разделен на пять частей по квантилям и уточнен итеративно с помощью WOE.

Масштабирование

Согласно критерию Д’Агостино (табл. 3.1) и визуальному анализу гистограмм (рис. 3.3), распределение объясняющих переменных не является

нормальным. В соответствие с этим выбран метод min-max нормализации к диапазону [0, 1].

Таблица 3.1 – Критерий Д’Агостино

Переменная	**Значение критерия**	**p-value**
TITLE	731	0
CHILDREN	663	0
PERS_H	255	0
AGE	323	0
TMADD	417	0
TMJOB1	951	0
TEL	742	0
NMBLOAN	1461	0
FINLOAN	5	0,1
INCOME	875	0
EC_CARD	643	0
INC	3615	0
INC1	752	0
BUREAU	187	0
LOANS	930	0
REGN	265	0
CASH	2445	0
PRODUCT	19	0
RESID	215	0
NAT	2319	0
PROF	1647	0
CAR	426	0
CARDS	1861	0

3.4 Выбор переменных

Мультиколлинеарность

Ни одно из рассчитанных значений VIF, как видно на рис. 3.9, не превышает 5, поэтому можно считать, что мультиколлинеарность отсутствует.

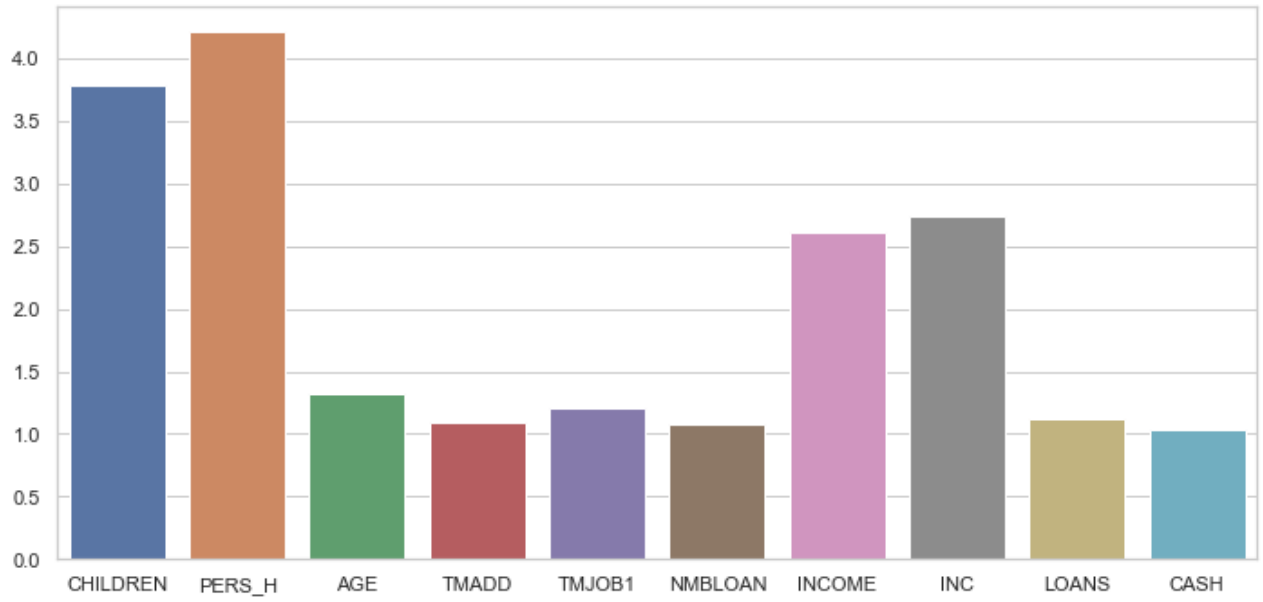

Рисунок 3.9 – Расчетное значение VIF

Для выявления корреляций между всеми объясняющими переменными произведен расчет корреляции Спирмена, представленный на рис. 3.10. Сильная положительная (от 0,7 до 0,9) корреляция наблюдается между переменными CHILDREN и PERS_H, CARDS и EC_CARD, EC_CARD и INC1, CARDS и INC1, INC и INC1. В случае CHILDREN и PERS_H, CARDS и EC_CARD очевидно наличие прямой причинно-следственной связи, в остальных случаях корреляцию можно объяснить тем, что переменные INC и INC1 были созданы с использованием переменных EC_CARD или CARDS. Для устранения сильной корреляции между переменными, необходимо удалить из анализа некоторые из данных переменных, по результатам оценки их информативности.

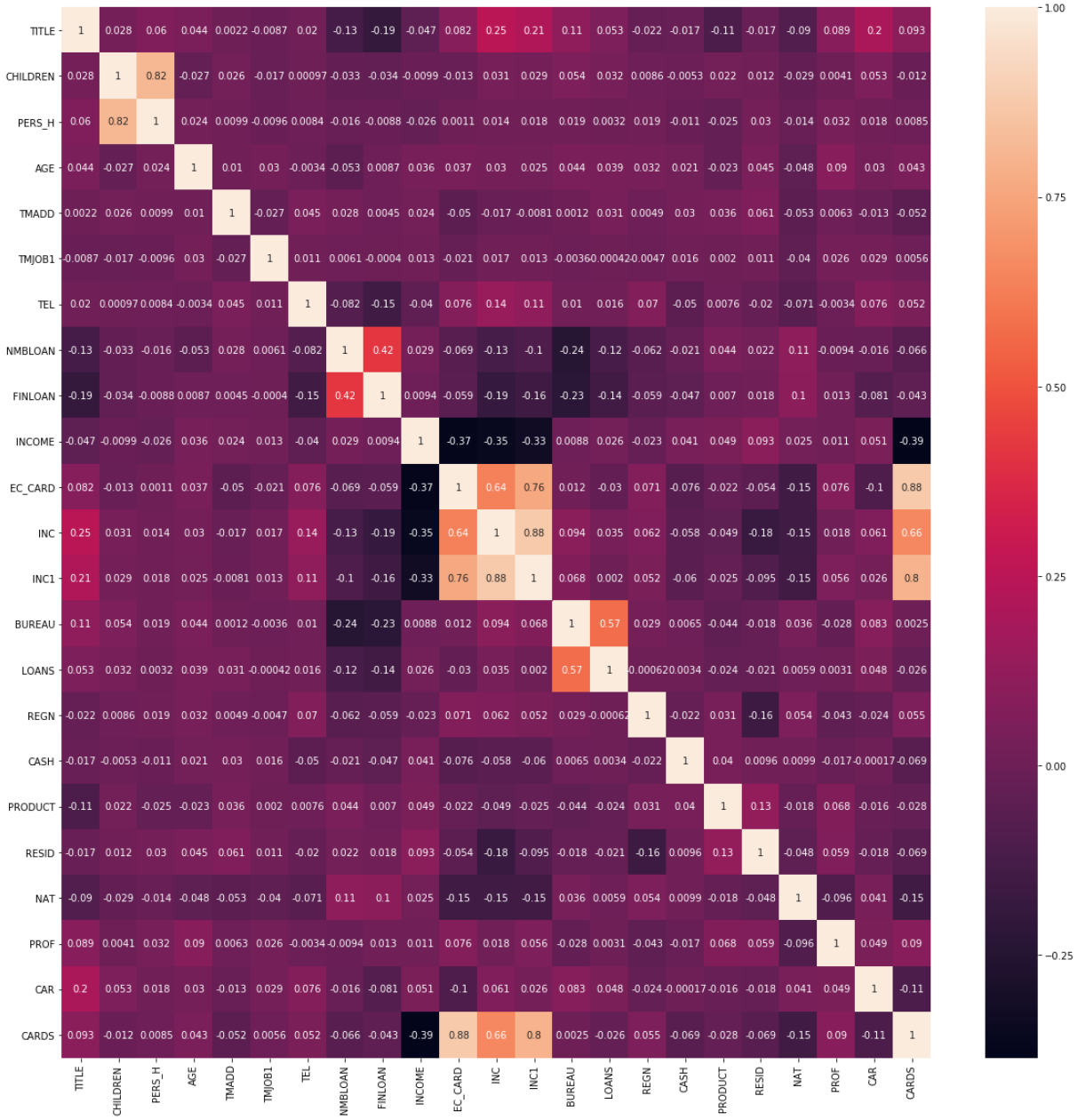

Рисунок 3.10 – Коэффициенты корреляции

Информативность

Согласно критерию хи-квадрат (рис. 3.11), наиболее информативными категориальными переменными являются CARDS и EC_CARD. Поскольку между ними наблюдается сильная корреляция, из анализа удалена переменная EC_CARD как менее информативная.

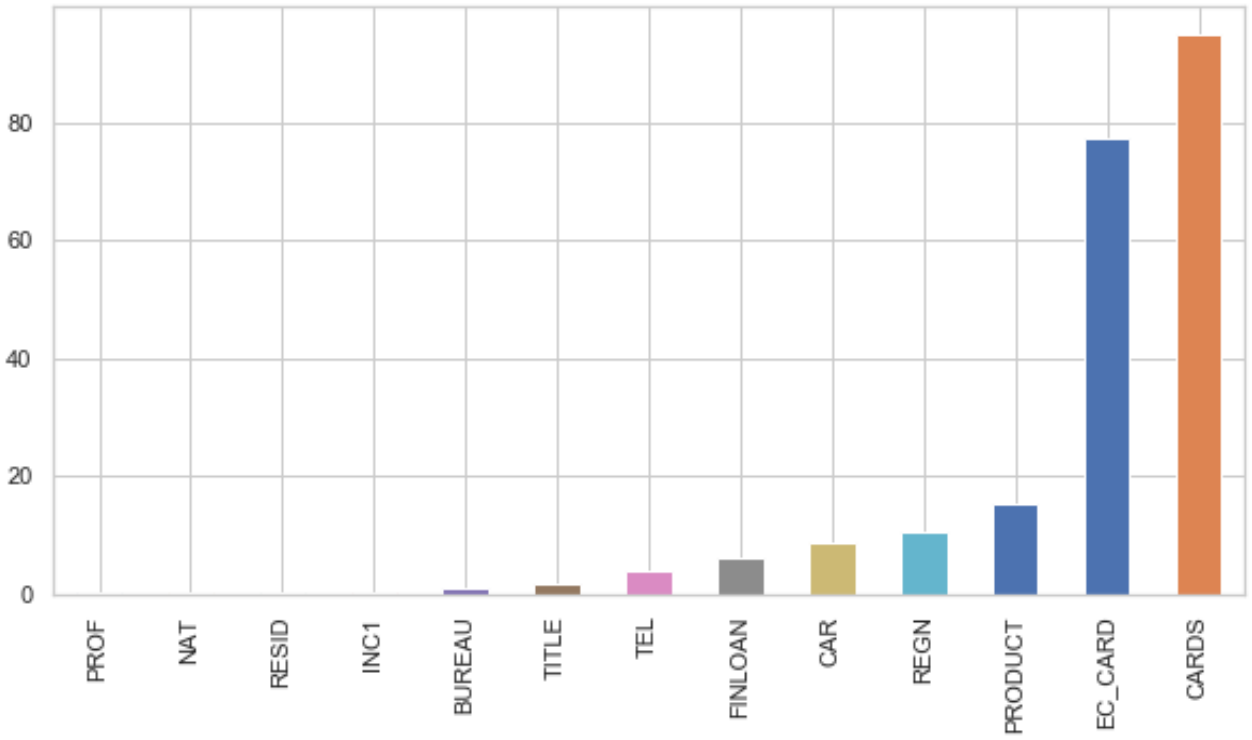

Рисунок 3.11 – Расчетное значение критерия хи-квадрат

Согласно критерию Фишера (рис. 3.12) наиболее информативными числовыми переменными являются AGE, TMJOB1 и PERS_H. Удаление ни одной из переменных не требуется, так как гипотеза об отсутствии взаимосвязи между независимыми переменными и зависимой переменной не подтверждена.

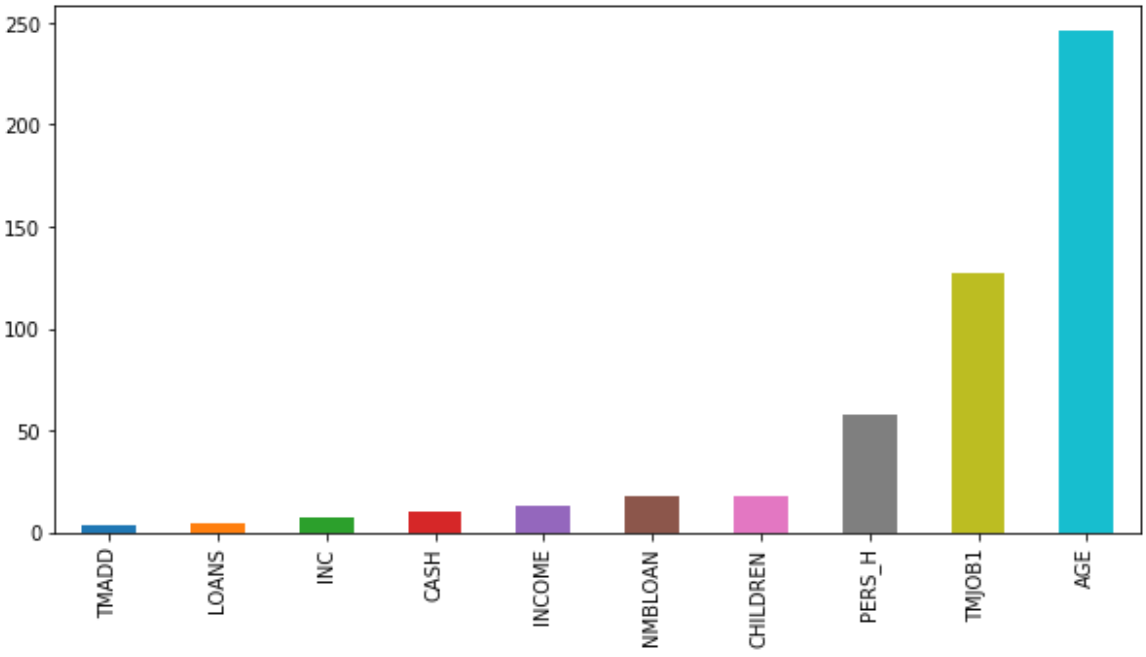

Рисунок 3.12 – Расчетное значение критерия Фишера

В результате применения алгоритма рекурсивного удаления признаков с кросс-валидацией, установлено оптимальное количество признаков – 21 (рис. 3.13).

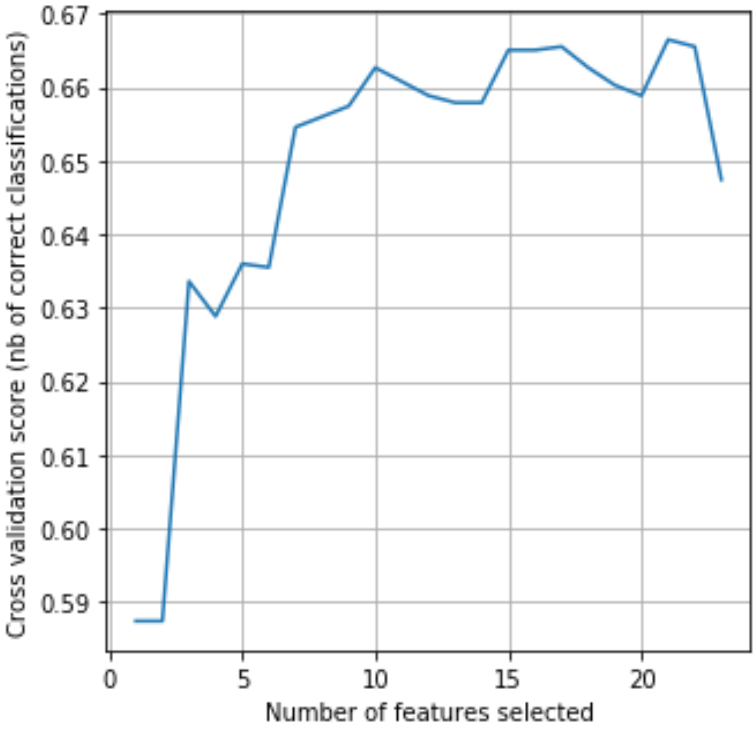

Рисунок 3.13 – Алгоритм рекурсивного удаления признаков

Аналогичная реализация методики произведена в пакете SAS Enterprise Miner. Карта последовательности этапов методики приведена приложении В.

Листинги реализации методики в пакете SAS приведены в приложении Б, результаты работы приведены в приложении В.

Оценка результата влияния методики подготовки данных на точность скоринга произведена сравнением результатов классификации логистической регрессии без подготовки данных с результатами после подготовки данных. Сравнение результатов приведено в табл. 3.2.

Таблица 3.2 – Сравнение результатов

	Python	**SAS**	**SAS Enterprise Miner**
Отсутствие подготовки данных	59%	59%	57%

Присутствие подготовки данных	69%	73%	75%
Прирост точности	10%	14%	18%

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ7И	Инхиреева Татьяна Александровна

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Магистратура	Направление/специальность	Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	Оклады участников проекта: – Оклад руководителя 33 664 руб. – Оклад инженера 21 760 руб. Тариф на электроэнергию 5,257 руб./кВт·час.
2. Нормы и нормативы расходования ресурсов	– Нормы амортизации 33,3%
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	– Ставки налоговых отчислений во внебюджетные фонды 30%. – Районный коэффициент по г. Томску 1,3.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Оценка коммерческого и инновационного потенциала НТИ	Анализ потенциальных потребителей проекта Оценка готовности проекта к коммерциализации и выбор метода коммерциализации Диаграмма Исикавы Проведение SWOT-анализа
2. Разработка устава научно-технического проекта	Постановка целей проекта, определение ожидаемого результата и критериев приемки проекта, рабочая группа проекта
3. Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок	Планирование этапов разработки программы, определение трудоемкости, построение диаграммы Ганта. Расчет сметы затрат на выполнение проекта.
4. Определение ресурсной, финансовой, экономической эффективности	Определение ресурсной и финансовой эффективности проекта
Перечень графического материала:	
1. Карта сегментирования рынка услуг по построению моделей кредитного скоринга *2. Диаграмма Исикавы* *3. Календарный график работ* *4. Структура затрат на проект*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОСГН ШБИП	Потехина Н.В.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата

4 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Подготовка данных для проведения кредитного скоринга, несмотря на ее очевидную значимость для дальнейшего анализа, зачастую игнорируется полностью или частично – пропускаются некоторые шаги, в то время как каждый шаг увеличивает точность предсказания, и, следовательно, уменьшает финансовые потери банка при выдаче кредита.

Целью данного раздела является определение перспективности научно-исследовательского проекта, разработка механизма управления и сопровождения конкретных проектных решений на этапе реализации.

Достижение цели обеспечивается решением задач:

- разработка общей экономической идеи проекта, формирование концепции проекта;
- организация работ и планирование работ;
- описание потенциальных потребителей;
- разработка диаграммы Исикавы;
- проведение SWOT-анализа;
- оценка готовности проекта к коммерциализации;
- формирование сметы затрат;
- выявление рисков и способов их смягчения;
- определение ресурсной и финансовой эффективности проекта.

4.1 Предпроектный анализ

4.1.1Потенциальные потребители результатов исследования

Для анализа потребителей результатов исследования рассмотрен целевой рынок и проведено его сегментирование.

Целевым рынком проекта являются большинство финансовых организаций, предоставляющих услуги кредитования и компании из сферы

информационных технологий, предоставляющие услуги построения скоринговых моделей.

Наиболее важными критериями сегментирования для предложенной методики являются отрасль компании и ее размер. Карта сегментирования, построенная в соответствии с данными критериями, представлена на рис. 4.1.

		Размер компании		
		Крупные	Средние	Малые
Отрасль компании	Финансы, кредитование			
	Информационные технологии			

Рисунок 4.1 – Карта сегментирования рынка услуг по построению моделей кредитного скоринга.

SAS Institute BaseGroup Labs

По карте сегментирования видно, что на рынке услуг существует незанятая ниша – малые компании из сферы информационных технологий. При разработке маркетинговой стратегии можно считать данный сегмент целевым.

4.1.2Диаграмма Исикавы

Диаграмма причины-следствия Исикавы позволяет графически проанализировать и сформировать причинно-следственные связи. Она может быть использована для выявления причин возникновения проблемы, анализа и структурирования процессов на предприятии и оценки причинно-следственных связей.

Проблемной областью анализа являются недостатки в данных, которые снижают точность кредитного скоринга и приводят банк к финансовым потерям.

Источниками недостатков в данных могут быть ошибки, вызванные опечатками при внесении данных в базу, различиями в форматах записи данных и дублирующиеся значения.

Повлиять на результат скоринга также могут помехи – пропущенные значения или выбросы (нетипичные значения) в данных.

Некоторые методы классификации, используемые для построения скоринговых карт, требуют данные, распределенные по нормальному закону. Присутствие корреляции и мультиколлинеарности может привести к искажению результатов скоринга.

Диаграмма Исикавы, построенная в соответствии с выявленными источниками и причинами недостатков в данных, приведена на рис. 4.2.

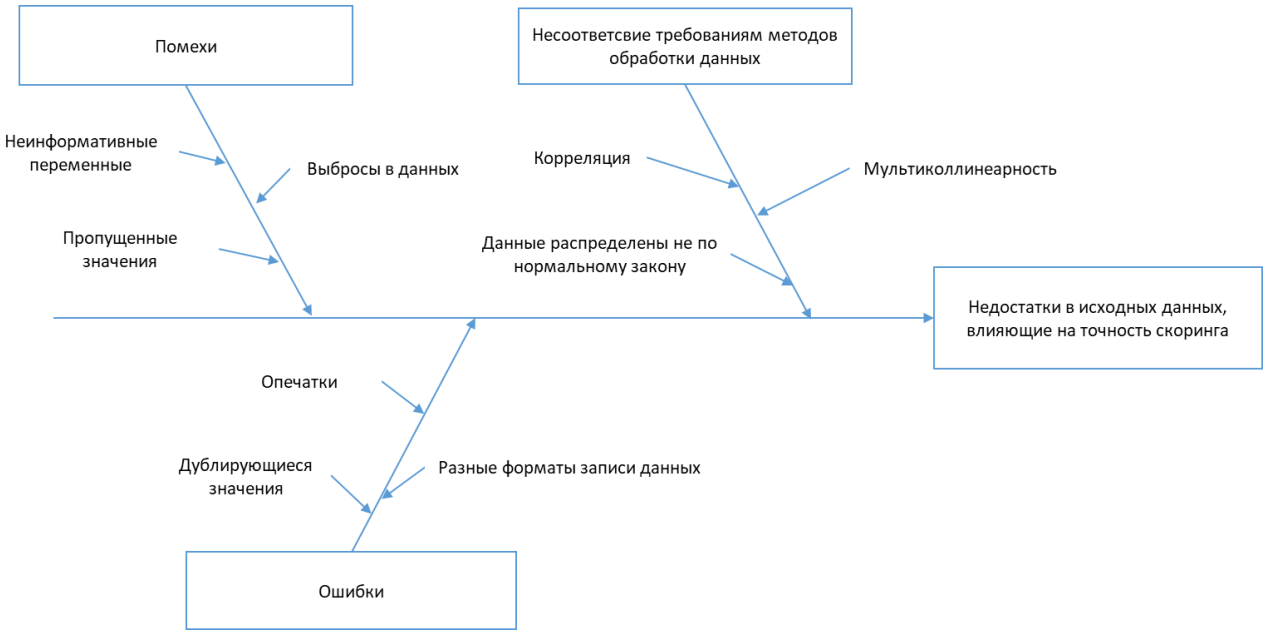

Рисунок 4.2 – Диаграмма Исикавы

Все выявленные причины являются значительными. В зависимости от использованного далее метода анализа данных, фактор «Несоответствие требованиям методов обработки данных» может включать и другие причины, например наличие категориальных переменных. Предложенная методика подготовки данных призвана устранить выявленные недостатки.

4.1.3SWOT-анализ

Для исследования внешней и внутренней среды проекта был проведен SWOT-анализ, который отражает сильные и слабые стороны разрабатываемого проекта. Сильные и слабые стороны являются факторами внутренней среды разрабатываемого проекта, (то есть то, на что сам объект способен повлиять); возможности и угрозы являются факторами внешней среды (то есть то, что может повлиять на объект извне и при этом не контролируется объектом).

Сильные стороны – это ресурсы или возможности, которыми располагает руководство проекта и которые могут быть эффективно использованы для достижения поставленных целей.

Слабые стороны – это то, что плохо получается в рамках проекта или где он располагает недостаточными возможностями или ресурсами по сравнению с конкурентами.

Возможности включают в себя любую предпочтительную ситуацию в настоящем или будущем, возникающую в условиях окружающей среды проекта.

Угроза представляет собой любую нежелательную ситуацию, тенденцию или изменение в условиях окружающей среды проекта, которые имеют разрушительный или угрожающий характер для его конкурентоспособности в настоящем или будущем. Результаты проведенного SWOT-анализ представлены в табл 4.1.

Таблица 4.1 – Матрица SWOT-анализа проекта

Сильные стороны: С1. Улучшение показателей (повышение точности) С2. Бесплатное распространение методологии С3. Реализация решения в различных программных продуктах С4. Адаптация методики подготовки данных для кредитного скоринга	**Слабые стороны:** Сл1. Ограниченное время на выполнение проекта Сл2. Ориентация на узкий сегмент рынка Сл3. Недостаточное количество данных для проведения полноценного тестирования
Возможности: В1. Потребность банков в качественной оценке кредитоспособности клиентов В2. Публикации о проекте в тематических журналах В3. Повышение спроса на методику подготовки данных	**Угрозы:** У1. Отсутствие широкого спроса на созданную методологию У2. Доминирование нескольких конкурентов У3. Наличие методологий-заменителей

Положительные и слабые стороны проекта, которые были выделены в ходе проведенного анализа, дают возможность спланировать необходимые

изменения, слабые стороны проекта необходимо по возможности минимизировать, опираясь прежде всего на имеющиеся сильные стороны.

Из-за негативных вариантов, связанных с угрозами низкого спроса на рынке, медленного выхода на рынок и невозможности занять свободный сегмент рынка необходимо в большей мере информировать возможного потребителя о появлении новой технологии. Для большего эффекта и сочетания с сильными сторонами и возможностями проекта в целях информирования используются такие инструменты как написание статей, выступления на научных конференциях и использование статуса ТПУ для получения доверия к новой методологии.

Ориентация на узкий сегмент рынка обусловлено наличием конкурентов и доминированием на рынке нескольких конкурентов и может привести к отсутствию широкого спроса на методологию. Ограниченное время на разработку проекта и недостаток данных для тестирования также негативным образом может повлиять на спрос.

Потребность банков в качественной оценке кредитоспособности заемщиков может помочь удовлетворить предложенная методика подготовки данных, поскольку она разработана специально для повышения точности кредитного скоринга. Расширить целевой сегмент рынка возможно за счет того, что реализация методики произведена в нескольких программных пакетах, которые используют компании из разных сегментов рынка. Недостаток данных может быть восполнен с помощью исторических данных банков. Публикации в тематических журналах способны помогут при распространении предложенной методологии.

4.2 Оценка готовности проекта к коммерциализации

Заполнение бланка оценки готовности научного проекта к коммерциализации позволяет оценить степень готовности разработки к коммерциализации и выяснить уровень собственных знаний для ее применения

на любом из этапов ее жизненного цикла. Заполненный бланк приведен в табл. 4.2.

Таблица 4.2 – Бланк оценки готовности научного проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся у разработчика знаний
1	Определен имеющийся научно-технический задел	5	4
2	Определены перспективные направления коммерциализации научно-технического отдела	5	3
3	Определены отрасли и технологии для предложения на рынке	4	3
4	Определена товарная форма научно-технического задела для представления на рынок	3	2
5	Определены авторы и осуществлена охрана их прав	3	3
6	Проведена оценка стоимости интеллектуальной собственности	1	2
7	Проведены маркетинговые исследования рынков сбыта	3	3
8	Разработан бизнес-план коммерциализации научной разработки	2	2
9	Определены пути продвижения научной разработки на рынок	2	2
10	Разработана стратегия реализации научной разработки	2	2
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	2	2
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	1	1

13	Проработаны вопросы финансирования коммерциализации научной разработки	1	1
14	Имеется команда для коммерциализации научной разработки	1	1
15	Проработан механизм реализации научного проекта	2	2
	ИТОГО БАЛЛОВ	37	33

По каждому из показателей выставляется оценка по пятибалльной шкале, при этом система измерений по каждому из направлений (степень проработанности научного проекта, уровень знаний разработчика) отличается. Система баллов приведена в табл. 4.3.

Таблица 4.3 – Система баллов

Балл	**Направление**	
	Степень проработанности научного проекта	**Уровень знаний разработчика**
1	Не проработанность	Не знаком или мало знаю
2	Слабая проработанность	В объеме теоретических знаний
3	Выполнено, но в качестве не уверен	Знаю теорию и практические примеры применения
4	Выполнено качественно	Знаю теорию и самостоятельно выполняю
5	Имеется положительное заключение независимого эксперта	Знаю теорию, выполняю и могу консультировать

Оценка готовности научного проекта к коммерциализации (или уровень имеющихся знаний) определяется суммарным количеством баллов по каждому из направлений.

Полученное значение суммарного балла 37 по степени проработанности научного проекта свидетельствует о средней перспективности разработки. Результаты оценки указывают на необходимость повышения уровня компетенций разработчика или привлечения специалистов в области

привлечения финансирования, продвижения продукта и формирование команды по коммерциализации разработки для работы над проектом в соответствующих областях.

Целью коммерциализации методики подготовки данных является как одноразовое получение финансовых ресурсов, так и обеспечение постоянного притока финансовых средств.

Выбор метода коммерциализации напрямую влияет на время продвижения товара на рынок. Задача данного раздела магистерской диссертации – выбор метода коммерциализации объекта исследования и обоснование его целесообразности.

В результате анализа методов коммерциализации, для реализации проекта выбраны два из них – инжиниринг и передача ноу-хау. Инжиниринг подходит для внедрения методики подготовки данных, поскольку эта методика позволяет обеспечить постоянный приток денежных средств от заказчиков к исполнителю. В качестве заказчика в данном случае выступают финансовые организации, предоставляющие услуги кредитования, исполнителем выступает владелец интеллектуальной собственности. Передача ноу-хау выбрана с целью распространения и дальнейшего улучшения предложенной методики.

4.3 Инициация проекта

Группа процессов инициации состоит из процессов, которые выполняются для определения нового проекта или новой фазы существующего. В рамках процессов инициации определяются изначальные цели и содержание и фиксируются изначальные финансовые ресурсы. Определяются внутренние и внешние заинтересованные стороны проекта, которые будут взаимодействовать и влиять на общий результат научного проекта.

4.3.1Цели и результаты проекта

В данном разделе приводится информация о заинтересованных сторонах проекта, иерархии целей проекта, а также критериях достижения целей. Под заинтересованными сторонами проекта понимаются лица или организации,

которые активно участвуют в проекте или интересы которых могут быть затронуты как положительно, так и отрицательно в ходе исполнения или в результате завершения проекта. Информация по заинтересованным сторонам проекта представлена в табл. 4.4.

Таблица 4.4 – Заинтересованные стороны проекта

Заинтересованные стороны проекта	**Ожидания заинтересованных сторон**
ОИТ ТПУ	Учебное пособие с описанием предложенной методики подготовки данных.
Инхиреева Т.А.	Написание выпускной квалификационной работы магистра

В табл. 4.5 представлена информация о иерархии целей проекта и критериях достижения целей, включая цели в области ресурсоэффективности и ресурсосбережения.

Таблица 4.5 – Иерархия целей проекта и критерии их достижения

Цели проекта:	Разработка методики подготовки данных для кредитного скоринга
Ожидаемые результаты проекта:	Документирование методики подготовки данных для кредитного скоринга и ее реализация на тестовом примере
Критерии приемки результатов проекта:	Повышение точности кредитного скоринга на тестовом примере на 10%
Требования к результату проекта:	Формализованное описание всех этапов подготовки данных в отчете
	Реализация в программных пакетах SAS, SAS Enterprise Miner, Python
	Повышение точности кредитного скоринга на тестовом примере

В табл. 4.6 представлена рабочая группа разработки, определена роль и основные функции каждого участника в разработке.

Таблица 4.6 – Рабочая группа разработки

№ п/п	ФИО, основное место работы, должность	Роль в разработке	Функции	Трудо-затраты, час.
1	Губин Евгений Иванович, ТПУ, кандидат физико-математических наук	Научный руководитель	Утверждение основных разделов, выдача заданий к исполнению, координирование деятельности исполнителя	42
2	Инхиреева Т.А.	Исполнитель	Исполнение поставленных задач	702
ИТОГО:				744

Данный раздел отражает тот факт, что выполняемая работа имеет довольно большой объём. Заинтересованные стороны проекта ожидают достаточно высококачественные результаты, которые необходимо достичь исполнителю.

4.4 Планирование управления научно-техническим проектом

4.4.1Организация и планирование работ

При организации процесса реализации конкретного проекта необходимо рационально планировать занятость каждого из его участников и сроки проведения отдельных работ. Полный перечень проводимых работ, определение их исполнителей и рациональная продолжительность приведены в табл. 4.7.

Таблица 4.7 – Перечень работ и загрузка исполнителей

Название	Длительность, дни	Дата начала работ	Дата окончания работ	Состав участников
Постановка целей и задач, получение исходных данных	1	10.01.2019	11.01.2019	Губин Е.И.

Составление и утверждение ТЗ	6	11.01.2019	17.01.2019	Губин Е.И., Инхиреева Т.А.
Разработка календарного плана	8	17.01.2019	25.01.2019	Губин Е.И., Инхиреева Т.А.
Подбор и изучение материалов по тематике	20	25.01.2019	14.02.2019	Губин Е.И., Инхиреева Т.А.
Обсуждение литературы	10	14.02.2019	24.02.2019	Губин Е.И., Инхиреева Т.А.
Разработка методики подготовки данных	19	24.02.2019	15.03.2019	Инхиреева Т.А.
Реализация методики подготовки данных	19	15.03.2019	03.04.2019	Инхиреева Т.А.
Проведение исследования	19	03.04.2019	22.04.2019	Инхиреева Т.А.
Оформление отчета	20	22.04.2019	12.05.2019	Инхиреева Т.А.
Оформление графического материала	8	12.05.2019	20.05.2019	Инхиреева Т.А.
Подведение итогов	10	20.05.2019	30.05.2019	Инхиреева Т.А.
Итого:	140	10.01.2019	30.05.2019	

Расчет продолжительности этапов работ при выполнении проекта является важным этапом, так как мы можем определить трудоемкость проводимых работ, а трудовые затраты составляют основную часть стоимости проекта.

Трудоемкость – это максимально допустимые затраты труда в человеко-днях на выполнение проекта с учетом организационно технических

мероприятий, обеспечивающих наиболее рациональное использование выделенных ресурсов.

Существуют разные методы расчета продолжительности этапов работы, в рамках данного проекта используется экспертный способ. Он предполагает генерацию необходимых количественных оценок специалистами конкретной предметной области, опирающимися на их профессиональный опыт и эрудицию.

Для определения вероятных (ожидаемых) значений продолжительности работ $t_{ожс}$ применяется следующая формула:

$$t_{ожс} = \frac{3t_{\min} + 2t_{\max}}{5}, \quad (4.1)$$

где t_{min} – минимальная продолжительность работы, дн.; t_{max}– максимальная продолжительность работы, дн.

Ожидаемая трудоемкость выполнения первого этапа работы

$$t_{ожс1} = \frac{3 \cdot 1 + 2 \cdot 1}{5} = 1.$$

Аналогичным образом посчитана ожидаемая трудоёмкость выполнения для всех остальных работ. Расчеты $t_{ожс\,i}$ занесены в табл. 4.8.

Для выполнения перечисленных в табл. 4.7 работ требуются специалисты:

– программист – исполнитель (И);

– научный руководитель (НР).

Для построения линейного графика необходимо рассчитать длительность этапов в рабочих днях, а затем перевести ее в календарные дни. Расчет продолжительности выполнения каждого этапа в рабочих днях ($T_{РД}$) ведется по формуле:

$$T_{РД} = \frac{t_{ожс}}{K_{ВН}} K_{Д}, \quad (4.2)$$

где $t_{ожс}$ – продолжительность работы, дн.; $K_{ВН}$ – коэффициент выполнения работ, учитывающий влияние внешних факторов на соблюдение предварительно определенных длительностей; $K_{Д}$ – коэффициент, учитывающий

дополнительное время на компенсацию непредвиденных задержек и согласование работ ($К_Д$ = 1,2).

Продолжительность выполнения первого этапа в рабочих днях составляет

$$T_{РД1} = \frac{1}{1} \cdot 1,2 = 1,2.$$

Аналогичным образом посчитана продолжительность выполнения каждого этапа в рабочих днях. Расчеты $T_{РД}$ занесены в табл. 4.8.

Расчет продолжительности этапа в календарных днях ведется по формуле:

$$T_{КД} = T_{РД} k_{КАЛ}, \quad (4.3)$$

где $T_{КД}$ – продолжительность выполнения этапа в календарных днях; $T_К$ – коэффициент календарности, позволяющий перейти от длительности работ в рабочих днях к их аналогам в календарных днях, и рассчитываемый по формуле:

$$k_{КАЛ} = \frac{T_{КАЛ}}{T_{КАЛ} - T_{ВД} - T_{ПД}}, \quad (4.4)$$

где $T_{КАЛ}$ – календарные дни ($T_{КАЛ}$ = 365), $T_{ВД}$ – выходные дни ($T_{ВД}$ = 52), $T_{ПД}$ – праздничные дни ($T_{ПД}$ = 10). По формуле (4.4) рассчитаем:

$$k_{КАЛ} = \frac{365}{365 - 52 - 10} = 1,2.$$

Проведем расчет продолжительности первого этапа в календарных днях

$$T_{КД} = 1,2 \cdot 1,2 = 1,44 \approx 1.$$

В табл. 4.8 приведен расчет определения продолжительности этапов работ и их трудоемкости по исполнителям, занятым на каждом этапе. По показанию полученных величин трудоёмкости этапов по исполнителям построен календарный график работ (рис. 4.3).

Таблица 4.8 – Трудозатраты на выполнение проекта

Этап	**Исполнители**	**Продолжительность работ, дни**			**Трудоемкость работ по исполнителям чел.–дн.**			
					$T_{РД}$		$T_{КД}$	
		t_{min}	t_{max}	$t_{ож}$	НР	И	НР	И
Постановка целей и задач,	НР	1	1	1	1	0	1	0

получение исходных данных								
Составление и утверждение ТЗ	НР, И	3	5	4	1	5	1	6
Разработка календарного плана	НР, И	5	7	6	1	7	1	8
Подбор и изучение материалов по тематике	НР, И	12	18	14	0	17	0	20
Обсуждение литературы	НР, И	6	8	7	1	8	1	10
Разработка алгоритма реализации метода	И	11	17	13	1	16	1	19
Реализация разработанного алгоритма	И	11	17	13	0	16	0	19
Проведение исследования	И	11	17	13	0	16	0	19
Оформление расчетно-пояснительной записки	И	12	18	14	0	17	0	20
Оформление графического материала	И	5	7	6	0	7	0	8
Подведение итогов	И	6	8	7	2	8	3	10
			Итого:	100,8	7	117	8	139

Этап	НР	И
1	1	0
2	1	6
3	0	20
4	1	8
5	1	10
6	1	19
7	0	19
8	0	19
9	0	20
10	0	8
11	3	10

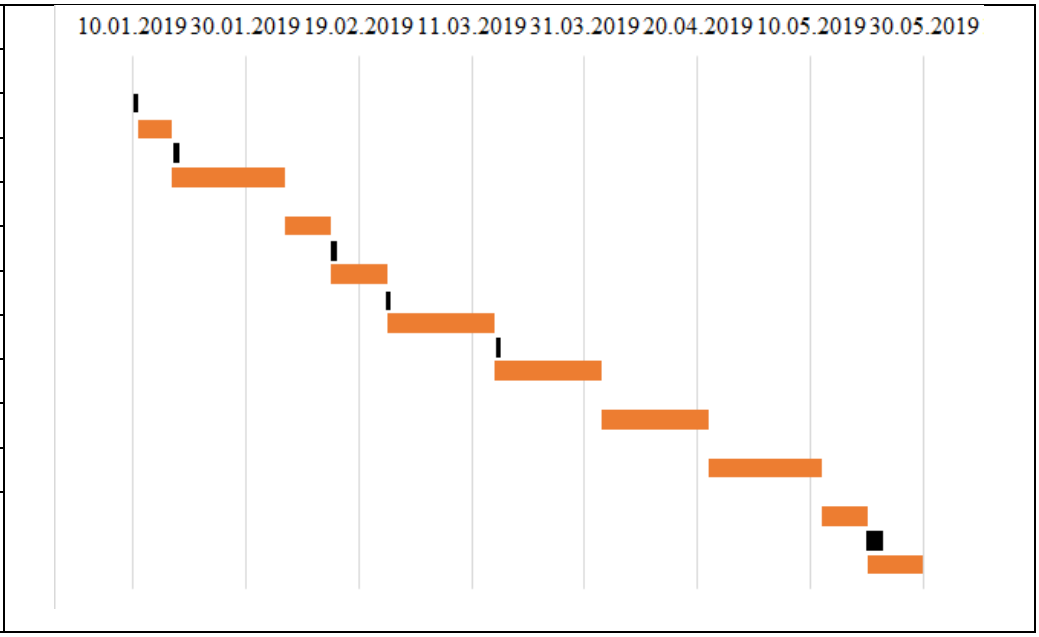

Рисунок 4.3 – Календарный график работ

 – И,  – НР.

4.4.2Расчет сметы затрат на выполнение проекта

В состав затрат на создание проекта включается величина всех расходов, необходимых для реализации комплекса работ, составляющих содержание данной разработки. Расчет сметной стоимости ее выполнения производится по следующим статьям затрат:

- материальные затраты;
- затраты на электроэнергию;
- затраты на основную заработную плату;
- отчислены во внебюджетные фонды;
- амортизационные отчисления;
- накладные расходы.

Материальные затраты

К материальным затратам относятся: приобретаемые со стороны сырье и материалы, покупные материалы, канцелярские принадлежности, картриджи и т.п.

Таблица 4.9 – Материальные затраты

Наименование.	**Единица измерения**	**Количество**	**Цена за ед., руб.**	**Затраты, руб.**
Краска для принтера	шт.	1	600	600
Бумага для принтера (500 листов)	пачка	2	255	510
Ручка шариковая	шт.	2	30	60
Текстовыделитель	шт.	4	135	540
Итого, руб.				1710

Материальные расходы составили 1710 рублей.

Расчет затрат на электроэнергию

Данный вид расходов включает в себя стоимость материалов, покупных изделий полуфабрикатов и других материальных ценностей, расходуемых непосредственно в процессе выполнения работ над объектом проектирования, а также затраты на электроэнергию, потраченную в ходе выполнения проекта на работу используемого оборудования,

Затраты на электроэнергию рассчитываются по формуле:

$$C_{эл.об.} = P_{об} \cdot t_{об} \cdot Ц_Э, \quad (4.5)$$

где $P_{ОБ}$ – мощность, потребляемая оборудованием, кВт; $Ц_Э$ – тариф на 1 кВт·час; $t_{об}$ – время работы оборудования, час.

В Томском политехническом университете $Ц_Э$ = 5,257 руб./кВт·час с учетом налога на добавленную стоимость.

Время работы оборудования вычисляется на основе итоговых данных табл. 4.8 для инженера ($T_{РД}$) из расчета, что продолжительность рабочего дня равна 8 часов.

$$t_{об} = T_{РД} \cdot K_t, \quad (4.6)$$

где $K_t \leq 1$– коэффициент использования оборудования по времени, равный отношению времени его работы в процессе выполнения проекта к $T_{РД}$, определяется исполнителем самостоятельно.

Мощность, потребляемая оборудованием, определяется по формуле:

$$P_{об} = P_{ном.} \cdot K_C, \quad (4.7)$$

где $P_{ном.}$ – номинальная мощность оборудования, кВт; $K_C \leq 1$ –коэффициент загрузки, зависящий от средней степени использования номинальной мощности. Для технологического оборудования малой мощности $K_C = 1$.

Расчет затраты на электроэнергию для технологических целей представлены в табл. 4.10.

Таблица 4.10 – Затраты на электроэнергию технологическую

Наименование оборудования	**Время работы оборудования $t_{об}$,час**	**Потребляемая мощность $P_{об}$, кВт**	**Затраты $Э_{ОБ}$, руб.**

Персональный компьютер	117*8	0,04	198,5

Расчет основной заработной платы

Смета затрат на оплату труда в большинстве случаев составляет наибольшую часть себестоимости проектных работ. Среднедневная тарифная заработная плата ($ЗП_{дн\text{-}т}$) рассчитывается по формуле:

$$ЗП_{дн-т} = \frac{МО}{24,91}, \quad (4.8)$$

учитывающей, что в году 299 рабочих дней и, следовательно, в месяце в среднем 24,91 рабочих дня (при шестидневной рабочей неделе).

Расчеты затрат на основную заработную плату приведены в табл. 4.11. Затраты времени по каждому исполнителю в рабочих днях с округлением до целого взяты из табл. 4.8. Для учета в ее составе районной надбавки используется коэффициент: $К_р$ = 1,3.

Таким образом, для перехода от тарифной (базовой) суммы заработка исполнителя, связанной с участием в проекте, к соответствующему полному заработку (зарплатной части сметы) необходимо первую умножить на районный коэффициент.

Таблица 4.11 – Расчет затрат на основную заработную плату

Исполнитель	Оклад, руб./мес.	Среднедневная ставка, руб./раб. день	Затраты времени, раб. дни	Коэффициент	Фонд з/платы, руб.
НР	33664,00	1351,43	7	1,3	12294,10
И	21760,00	873,54	117	1,3	132865,43
Итого:					145159,53

Расчет дополнительной заработной платы

В данную статью включаются выплаты, предусмотренные законодательством о труде, такие как оплата очередных и дополнительных

отпусков, оплата времени, связанного с выполнением государственных и общественных обязанностей, выплата вознаграждения за выслугу лет и т.д.

Дополнительная заработная плата рассчитана исходя из 12% от основной заработной платы работников, непосредственно участвующих в выполнении темы. Расчет затрат на заработную представлен в табл. 4.12.

Таблица 4.12 – Затраты на заработную плату

Заработная плата	**НР**	**И**
Основная заработная плата	12294,10	132865,43
Дополнительная заработная плата	1475,29	15943,85
Итого:	13769,39	148809,28

$$C_{зп} = 13769,39 + 148809,28 = 162578,67 \text{ руб.}$$

Расчет отчислений во внебюджетные фонды

В данной статье расходов отражаются обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников.

Величина отчислений во внебюджетные фонды определяется исходя из следующей формулы:

$$C_{внеб} = k_{внеб} * C_{зп} \quad (4.9)$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.).

В соответствии с Федеральным законом установлен размер страховых взносов равный 30%.

Итак, в нашем случае:

$$C_{внеб.} = 162578,67 * 0,3 = 48773,60 \text{ руб.}$$

Расчет амортизационных отчислений

В данном разделе рассчитывается амортизация используемого оборудования (ПК) за время выполнения проекта. Первоначальная стоимость ПК

составляет 50000 рублей. Срок полезного использования для машин офисных код 330.28.23.23 составляет 2-3 года. Планируется использование ПК для исследования в течение 4 месяцев. Норма амортизации:

$$A_H = \frac{1}{n} \cdot 100\% = \frac{1}{3} \cdot 100\% = 33,33\%.$$

Годовые амортизационные отчисления:

$$C_Г = 50000 \cdot 0,33 = 16500 \text{ рублей.}$$

Сумма амортизации основных средств за период выполнения проекта:

$$C_{ам} = 16500 \cdot \frac{4}{12} = 5500 \text{ рублей.}$$

Расчет накладных расходов

В статье «Накладные расходы» отражены расходы на выполнение проекта, которые не учтены в предыдущих статьях, к ним относятся содержание оргтехники, услуги связи, представительные расходы и другие. Их следует принять равными 16% от суммы всех предыдущих расходов:

$$C_{н.} = (C_{мат} + C_{эл.об} + C_{зп} + C_{внеб.} + C_{ам}) \cdot 0,16 \qquad (4.10)$$

Найдем прочие расходы по формуле (4.10), учитывая данные полученные выше:

$C_{н.}$ = (1710 + 198,50 + 162578,67+ 48773,60 + 5500,00) · 0,16 = 35001,72 руб.

Расчет общей себестоимости разработки

Проведя расчет по всем статьям сметы затрат на проект, можно определить общую себестоимость. Смета затрат на проект представлена в табл. 4.13.

Таблица 4.13 – Смета затрат на проект

Статья затрат	Сумма, руб	в %
Материальные затраты	1710	0,7
Затраты на электроэнергию	198,50	0,1
Заработная плата	162578,67	64,1

Отчисления во внебюджетные фонды	48773,60	19,2
Амортизационные отчисления	5500,00	2,2
Накладные расходы	35001,72	13,8
Итого:	253762,49	100

Структура расходов наглядно изображена на рис. 4.4.

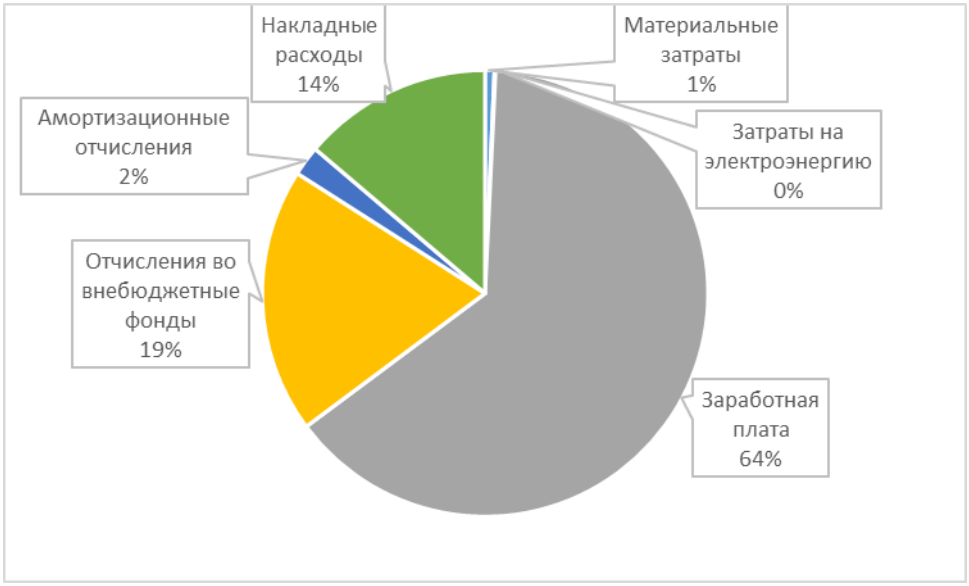

Рисунок 4.4 – Структура затрат на проект

Большая часть расходов (64%) приходится на заработную плату.

4.4.3Реестр рисков проекта

Риск – это возможность наступления некоторого неблагоприятного события, влекущего за собой возникновение различного рода потерь. Единой классификации рисков проекта не существует. Можно выделить следующие основные группы рисков, присущие практически всем проектам: политические, экономические, социальные, технологические, экологические, финансовые, организационные, маркетинговые, кадровые, технические. В табл. 4.14 представлены основные группы риска, присущие проекту, их потенциальное воздействие и условия наступления. Произведена оценка вероятности риска по шкале вероятности риска и шкале оценки уровня потерь, предложены мероприятия по снижению рисков.

Таблица 4.14 – Реестр рисков

Риск	Потенциальное воздействие	Вероятность наступления (1-5)	Влияние риска (1-5)	Уровень риска	Способы смягчения риска	Условия наступления
Организационные	Увеличение сроков работ, остановка работ	3	3	Средний	Выделение ответственного со стороны заказчика для контроля сроков и результатов	Недостаточная поддержка со стороны руководства заказчика
Маркетинговые	Недостижение рыночной доли	4	5	Высокий	Разработка уникального торгового предложения	Недооценка возможностей конкурентов
Технические	Отсутствие необходимости в данной методике подготовки данных	1	5	Низкий	Расширение методологии подготовки данных на другие предметные области	Изменение технологий принятия решений о кредитоспособности заемщика в банках

Анализ табл. 4.14 показывает, что наиболее опасным для проекта является маркетинговый риск. Организационный риск на среднем уровне. Технический риск, хоть и влечет за собой тяжелые для проекта последствия, довольно маловероятен.

4.5 Определение ресурсной и финансовой эффективности проекта

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный показатель финансовой эффективности научного исследования получают в ходе оценки бюджета затрат научного исследования. Для этого наибольший интегральный показатель реализации технической задачи

принимается за базу расчета (как знаменатель), с которым соотносится финансовые значения.

Интегральный финансовый показатель разработки определяется как:

$$I_{финр} = \frac{Ф_p}{Ф_{max}}, \quad (4.11)$$

где $I_{финр}$ – интегральный финансовый показатель разработки, $Ф_p$– стоимость i-го варианта исполнения, $Ф_{max}$ – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

Максимальная стоимость составляет 350000 рублей, следовательно:

$$I_{финр} = \frac{253762,49}{350000} = 0,7$$

Полученная величина интегрального финансового показателя разработки составила 0,7, что отражает соответствующее численное удешевление стоимости разработки в разах.

Интегральный показатель ресурсоэффективности исполнения объекта исследования можно определить следующим образом:

$$I_p = \sum a \cdot b \quad (4.12)$$

где I_p – интегральный показатель ресурсоэффективности для i-го варианта исполнения разработки, a – весовой коэффициент, b – бальная оценка, устанавливается экспертным путем по выбранной шкале оценивания, n – число параметров сравнения.

Расчет интегрального показателя ресурсоэффективности приведен в табл. 4.15.

Таблица 4.15 – Оценка характеристик исполнения проекта

Объект исследования / **Критерии**	**Весовой коэффициент параметра**	**Оценка выполнения**
1. Трудоемкость	0,20	3
2. Точность вычислений	0,35	5
3. Потребность в ресурсах памяти	0,10	5
4. Надежность	0,20	5
ИТОГО	1	

$$I_p = 3*0,20+5*0,35+5*0,10+5*0,20 = 3,85;$$

Данные, полученные при оценке конкурентоспособности, позволяют сделать вывод, что разработка является перспективной и привлекательной для инвесторов и потребителей.

Общая длительность проектирования и разработки программного продукта составила 140 дней.

Общий бюджет проекта составляет 253762,49 рублей. Он включает в себя затраты на заработную плату работников, материальные затраты, затраты на электроэнергию, отчисления на внебюджетные фонды, амортизационные затраты и накладные расходы.

Разработка может представлять интерес для кредитно-финансовых организаций, предоставляющих услуги кредитования и компаний из сферы информационных технологий, предоставляющих услуги построения скоринговых моделей.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ПМ7И	Инхиреева Татьяна Александровна

Школа	ИШИТР	Отделение (НОЦ)	ОИТ
Уровень образования	Магистратура	Направление/специальность	Программная инженерия

Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Методика кредитного скоринга. Применяется для оценки кредитоспособности потенциальных заемщиков в банках.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	– специальные правовые нормы трудового законодательства при работе с компьютером и орг. техникой (Трудовой кодекс РФ, СанПиН 2.2.2/2.4.1340-03 Гигиенические требования к персональным электронно-вычислительным машинам и организации работы); – требования к организации рабочих мест пользователей (ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования», ГОСТ 12.2.061-81 «ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам»).
2. Производственная безопасность 2.1 Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	Вредные и опасные факторы: – отклонение показателей микроклимата; – превышение уровня шума; – отсутствие или недостаток естественного света; – недостаточная освещенность рабочей зоны; – умственное перенапряжение; – повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека;
3. Экологическая безопасность	– анализ воздействия объекта на литосферу, гидросферу и атмосферу (отходы, связанные с утилизацией вышедшего из строя ПК, люминесцентных ламп и др.); – разработка решений по обеспечению экологической безопасности.

4. Безопасность в чрезвычайных ситуациях:	– типичная ЧС – пожар. – разработка превентивных мер по предупреждению пожара; – разработка действий в результате пожара и мер по ликвидации последствий.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Горбенко М.В.	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Инхиреева Татьяна Александровна		

5 Социальная ответственность

Введение

Объект исследования – данные о кредитоспособности заемщиков.

Выпускная квалификационная работа представляет собой составление методики обработки данных для кредитного скоринга. В ходе выполнения работы проводилось изучение, анализ, тестирование и сравнение различных существующих методик подготовки данных, реализация алгоритма методики. Расчеты и алгоритмы произведены с помощью Python, SAS, SAS Enterprise Miner. Разработанная методика может применяться для кредитного скоринга в банках.

В разделе будут рассмотрены опасные и вредные факторы, оказывающие влияние на производственную деятельность инженера-программиста. Исследовано рабочее место программиста и помещение, в котором он находится. Предполагаемое место работы – компьютерный класс Кибернетического центра ТПУ. Основные средства работы – персональный компьютер и локальная вычислительная сеть с выходом в Интернет. Рассмотрены воздействия объекта исследования на окружающую среду, правовые и организационные вопросы, а также мероприятия в чрезвычайных ситуациях.

5.1 Правовые и организационные вопросы обеспечения безопасности

5.1.1Специальные правовые нормы трудового законодательства

Отношения межу работником и работодателем регулируются трудовым кодексом РФ. Согласно трудовому кодексу РФ, нормальная продолжительность рабочего времени не может превышать 40 часов в неделю.

Порядок исчисления нормы рабочего времени на определенные календарные периоды (месяц, квартал, год) в зависимости от установленной продолжительности рабочего времени в неделю определяется федеральным органом исполнительной власти, осуществляющим функции по выработке

государственной политики и нормативно-правовому регулированию в сфере труда.

Продолжительность ежедневной работы (смены) не может превышать:

Для работников, занятых на работах с вредными и (или) опасными условиями труда, где установлена сокращенная продолжительность рабочего времени, максимально допустимая продолжительность ежедневной работы (смены) не может превышать:

при 36-часовой рабочей неделе – 8 часов;

при 30-часовой рабочей неделе и менее – 6 часов.

Продолжительность рабочего дня или смены, непосредственно предшествующих нерабочему праздничному дню, уменьшается на один час.

Ночное время – время с 22 часов до 6 часов. Продолжительность работы (смены) в ночное время сокращается на один час без последующей отработки.

В течение рабочего дня (смены) работнику должен быть предоставлен перерыв для отдыха и питания продолжительностью не более двух часов и не менее 30 минут, который в рабочее время не включается. Правилами внутреннего трудового распорядка или трудовым договором может быть предусмотрено, что указанный перерыв может не предоставляться работнику, если установленная для него продолжительность ежедневной работы (смены) не превышает четырех часов.

Всем работникам предоставляются выходные дни (еженедельный непрерывный отдых).

Организация-работодатель выплачивает заработную плату работникам. Возможно удержание заработной платы только в случаях, установленных ТК РФ ст. 137. В случае задержки заработной платы более чем на 15 дней, работник имеет право приостановить работу, письменно уведомив работодателя.

Обработка персональных данных работника может осуществляться исключительно в целях обеспечения соблюдения законов и иных нормативных правовых актов, содействия работникам в трудоустройстве, получении образования и продвижении по службе, обеспечения личной безопасности

работников, контроля количества и качества выполняемой работы и обеспечения сохранности имущества.

Все персональные данные работника следует получать у него самого. Если персональные данные работника возможно получить только у третьей стороны, то работник должен быть уведомлен об этом заранее и от него должно быть получено письменное согласие. Работодатель должен сообщить работнику о целях, предполагаемых источниках и способах получения персональных данных, а также о характере подлежащих получению персональных данных и последствиях отказа работника дать письменное согласие на их получение.

Законодательством РФ запрещена дискриминация по любым признакам и принудительный труд.

5.1.2Организационные мероприятия при компоновке рабочей зоны

При работе с персональным компьютером очень важную роль играет соблюдение правильного режима труда и отдыха. В противном случае у работника отмечаются значительное напряжение зрительного аппарата с появлением жалоб на неудовлетворенность работой, головные боли, раздражительность, нарушение сна, усталость и болезненные ощущения в глазах, в пояснице, в области шеи и руках.

При восьмичасовой рабочей смене на ВДТ и ПЭВМ перерывы в работе должны составлять от 10 до 20 минут каждые два часа работы (ТОИ Р-45-084-01). В перерывах, рекомендуется проводить комплекс упражнений для глаз (СанПиН 2.2.2/2.4.1340-03).

5.1.3Эргономические требования к рабочему месту оператора ПЭВМ

Проектирование рабочих мест, снабженных видеотерминалами, относится к числу важных проблем эргономического проектирования в области вычислительной техники.

Организация рабочего места программиста или оператора регламентируется следующими нормативными документами: ГОСТ 12.2.032-78 ССБТ, ГОСТ 12.2.033-78 ССБТ, СанПиН 2.2.2/2.4.1340-03 и рядом других.

Главными элементами рабочего места программиста или оператора являются стол, кресло, дисплей, клавиатура и мышь. Основным рабочим положением является положение сидя.

Для комфортной работы стол должен удовлетворять следующим условиям:

- высота стола должна быть выбрана с учетом возможности сидеть свободно, в удобной позе, при необходимости опираясь на подлокотники;
- нижняя часть стола должна быть сконструирована так, чтобы программист мог удобно сидеть, не был вынужден поджимать ноги;
- поверхность стола должна обладать свойствами, исключающими появление бликов в поле зрения программиста;
- конструкция стола должна предусматривать наличие выдвижных ящиков (не менее 3 для хранения документации, листингов, канцелярских принадлежностей);
- высота рабочей поверхности рекомендуется в пределах 680-760 мм. Высота поверхности, на которую устанавливается клавиатура, должна быть около 650 мм.

Рабочий стул (кресло) должен быть подъемно-поворотным и регулируемым по высоте и углам наклона сиденья и спинки, а также регулируемым по расстоянию спинки от переднего края сиденья. Конструкция стула должна обеспечивать:

- ширину и глубину поверхности сиденья не менее 400 мм;
- поверхность сиденья с закругленным передним краем;
- регулировку высоты поверхности сиденья в пределах 400-550 мм и углов наклона вперед до 15° и назад до 5°;
- высоту опорной поверхности спинки 300 ± 20 мм, ширину не менее 380 мм и радиус кривизны горизонтальной плоскости 400 мм;

- угол наклона спинки в вертикальной плоскости в пределах 0 ± 30°;
- регулировку расстояния спинки от переднего края сиденья в пределах 260-400 мм;
- стационарные или съемные подлокотники длиной не менее 250 мм и шириной 50-70 мм;
- регулировку подлокотников по высоте над сиденьем в пределах 230 ± 30 мм и внутреннего расстояния между подлокотниками в пределах 350-500 мм.

Работающий за ПЭВМ должен сидеть прямо, опираясь в области нижнего края лопаток на спинку кресла, не сутулясь, с небольшим наклоном головы вперед (до 5-7°). Предплечья должны опираться на поверхность стола, снимая тем самым статическое напряжение плечевого пояса и рук.

Положение экрана определяется:

- расстоянием считывания (0,6…0,7 м);
- углом считывания, направлением взгляда на 20° ниже горизонтали к центру экрана, причем экран перпендикулярен этому направлению.

Должна также предусматриваться возможность регулирования экрана:

- по высоте +3 см;
- по наклону от -10° до +20° относительно вертикали;
- в левом и правом направлениях.

Рабочее место в аудитории № 105 КЦ ТПУ отвечает данным требованиям.

5.2 Производственная безопасность

Согласно ГОСТ 12.0.003-2015, опасные и вредные факторы по характеру происхождения делятся на следующие группы:

- физические;
- химические;
- психофизиологические;

– социально-экономические;

– биологические.

Перечень опасных и вредных факторов, влияющих на оператора ПЭВМ, представлен в табл. 5.1.

Таблица 5.16 – Вредные и опасные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ		Нормативные документы
	Исследование	Разработка	
1. Отклонение показателей микроклимата	+	+	1. СанПиН 2.2.4.548- 96; 2. СНиП 21-01-97; 3. СанПиН 2.2.2/2.4.1340-03; 4. СП 52.13330.2011; 5. ГОСТ Р 12.1.019- 2009 ССБТ;
2. Превышение уровня шума	+	+	
3. Отсутствие или недостаток естественного света	+	+	
4. Недостаточная освещенность рабочей зоны	+	+	
5. Умственное перенапряжение	+	+	
6. .Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	+	+	

5.2.1Анализ опасных и вредных производственных факторов

Отклонение показателей микроклимата

Значимым физическим фактором является микроклимат рабочей зоны (температура, влажность и скорость движения воздуха).

Температура, относительная влажность и скорость движения воздуха влияют на теплообмен и необходимо учитывать их комплексное воздействие. Нарушение теплообмена вызывает тепловую гипертермию, или перегрев.

Оптимальные нормы температуры, относительной влажности и скорости движения воздуха производственных помещений для работ, производимых сидя и не требующих систематического физического напряжения (категория Iа),

приведены в табл. 5.2, в соответствии с СанПиН 2.2.2/2.4.1340-03 и СанПиН 2.2.4.548-96.

Таблица 5.2

Нормы температуры, относительной влажности и скорости движения воздуха

Период года	Категория работы	Температура, С	Относительная влаж. воздуха, %	Скорость движения воздуха, не более м/с
Холодный	Ia	22-24	40-60	0,1
Теплый	Ia	23-25	40-60	0,1

Допустимые микроклиматические условия установлены по критериям допустимого теплового и функционального состояния человека на период 8-часовой рабочей смены. Они устанавливаются в случаях, когда по технологическим требованиям, техническим и экономически обоснованным причинам не могут быть обеспечены оптимальные величины.

Допустимые величины показателей микроклимата на рабочих местах представлены в табл. 5.3.

Таблица 5.3 – Допустимые величины показателей микроклимата

Период года	Категория работы	Температура воздуха, °С	Относительная влаж. воздуха, %	Скорость движения воздуха, не более м/с
Холодный	Ia	20-25	15-75	0,1
Теплый	Ia	21-28	15-75	0,1-0,2

Для обеспечения установленных норм микроклиматических параметров и чистоты воздуха на рабочих местах и в помещениях применяют вентиляцию. Общеобменная вентиляция используется для обеспечения в помещениях соответствующего микроклимата. Периодически должен вестись контроль влажностью воздуха. В летнее время при высокой уличной температуре должны использоваться системы кондиционирования.

В холодное время года предусматривается система отопления. Для отопления помещений используются водяные системы центрального отопления.

Превышение уровня шума

Шум – колебания различной физической природы, отличающиеся сложностью спектральной и временной структуры. Шум создает значительную нагрузку на нервную систему человека, оказывая на него психологическое воздействие. Шумовой фон провоцирует увеличение содержания в крови гормонов стресса, таких как, норадреналин и адреналин, кортизол. Шум способен замедлять реакцию человека и угнетать центральную нервную систему (ЦНС), вызывая изменения скорости пульса и дыхания, а также провоцирует возникновение сердечно - сосудистых заболеваний, гипертонических болезней и язвы желудка.

Уровень звука на рабочих местах, связанных с творческой деятельностью, научной деятельностью, программированием, преподаванием и обучением не должен превышать 50 дБА согласно СН 2.2.4/2.1.8.562–96.

Для исследуемого объекта (компьютерный зал) основными источниками шумов являются составляющие компьютера:

- Вентилятор блока питания
- Вентилятор кулера центрального процессора.
- Вентилятор на высокопроизводительной видеокарте.
- Дополнительный вентилятор в корпусе системного блока.
- Звуки нажатия клавиш пальцами пользователя, шуршания мыши по коврику.

Меры, которые необходимо принять, для того чтобы помещение было менее зашумленным – это обеспечить нормальную вентиляцию системного блока. Для охлаждения необходимо оборудовать со стороны вентиляционных отверстий хотя бы 20-30 см свободного пространства. Не загромождать оборудование посторонними предметами, которые снижают теплоотдачу, прочищать вентиляционные отверстия от пыли пылесосом.

Освещение

Освещение рабочего места – важнейший фактор создания нормальных условий труда. Освещению следует уделять особое внимание, так как при работе наибольшее напряжение получают глаза.

Освещение делится на естественное, искусственное и совмещенное. Совмещенное сочетает оба вида освещения.

В компьютерных залах, где расположено рабочее место программиста, используется совмещенное освещение.

Рекомендуемые соотношения яркостей в поле зрения следующие (СНиП 23-05-95):

- между рабочими поверхностями не должно превышать 1:3 – 1:5;
- между рабочими поверхностями и поверхностями стен и оборудования – 10:1.

Освещённость на рабочем месте должна соответствовать характеру зрительной работы, который определяется наименьшим размером объекта различения, контрастом объекта с фоном и характеристикой фона.

Освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300-500 лк (СНиП 23-05-95, СанПиН 2.2.2/2.4.1340-03). Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк. Следует ограничивать прямую блесткость от источников освещения, при этом яркость светящихся поверхностей (окна, светильники и др.), находящихся в поле зрения, должна быть не более 200 кд/м^2. Показатель ослепленности для источников общего искусственного освещения в производственных помещениях должен быть не более 20.

Умственное перенапряжение

Умственное перенапряжение вызывается информационной нагрузкой. Чтобы его избежать, необходимо устраивать небольшие перерывы в течение рабочего дня продолжительностью не более 5 минут. При умственной работе, по

сравнению с физической, потребление мозгом кислорода увеличивается в 15-20 раз. Если для умственной работы требуется значительное нервно-эмоциональное напряжение, то возможны значительные изменения кровяного давления, пульса. Продолжительная умственная работа может привести к сердечно-сосудистым и некоторым другим заболеваниям. Рабочее место позволяет делать перерывы в течение дня.

Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека

Проходя через организм человека, электрический ток оказывает термическое, электролитическое и биологическое действие.

Первое заключается в нагреве и ожогах различных частей и участков тела человека, второе — в изменении состава (разложение) и свойств крови и других органических жидкостей. Биологическое действие электрического тока выражается в раздражении и возбуждении живых тканей организма и в нарушении протекания в нем различных внутренних биоэлектрических.

Во время использования средства вычислительной техники или другими периферийными устройствами оператор должен осторожно обращаться с электропроводкой, аппаратами и приборами и всегда помнить, что, если не придерживаться правил безопасности, то это может угрожать здоровью и жизни человека.

Согласно ГОСТ Р 12.1.019-2009, для обеспечения защиты от поражения электрическим током, применяют следующие способы:

- защитное заземление;
- зануление;
- защитное отключение;
- изоляцию нетоковедущих частей;
- контроль изоляции;
- средства индивидуальной защиты;
- использование устройств бесперебойного питания.

Технические способы и средства применяют раздельно или в сочетании друг с другом так, чтобы обеспечивалась оптимальная защита.

Организационные мероприятия включают (ГОСТ Р 12.1.019-2009):

- проверку знаний правил безопасности и инструкций в соответствии с занимаемой должностью применительно к выполняемой работе с присвоением соответствующей квалификационной группы по электробезопасности;
- осуществление допуска к проведению работ;
- организацию надзора за проведением работ;
- установление рациональных режимов труда.

Чтобы избежать поражения электрическим током, необходимо выполнять следующие правила по ГОСТ Р 12.1.019-2009:

1. Необходимо постоянно следить на своем рабочем месте за исправным состоянием электропроводки, выключателей, штепсельных розеток, при помощи которых оборудование включается в сеть, и заземления. При обнаружении неисправности немедленно обесточить электрооборудование, оповестить администрацию. Продолжение работы возможно только после устранения неисправности.

2. Для исключения поражения электрическим током запрещается:

- часто включать и выключать компьютер без необходимости;
- прикасаться к экрану и к тыльной стороне блоков компьютера;
- работать на средствах вычислительной техники и периферийном
- оборудовании мокрыми руками;
- работать на средствах вычислительной техники и периферийном оборудовании, имеющих нарушения целостности корпуса, нарушения изоляции проводов, неисправную индикацию включения питания, с признаками электрического напряжения на корпусе
- класть на средства вычислительной техники и периферийное оборудование посторонние предметы.

3. Запрещается под напряжением очищать от пыли и загрязнения электрооборудование.

4. Ремонт электроаппаратуры производится только специалистами-техниками с соблюдением необходимых технических требований.

Во всех случаях поражения человека электрическим током немедленно вызывают врача. До прибытия врача нужно, не теряя времени, приступить к оказанию первой помощи пострадавшему (ГОСТ Р 12.1.019-2009).

5.2.2Обоснование мероприятий по защите от действия опасных и вредных факторов

Требования к помещениям для работы с ПЭВМ

В соответствии с основными требованиями к помещениям для эксплуатации ПЭВМ (СанПиН 2.2.2/2.4.1340-03) эти помещения должны иметь естественное и искусственное освещение. Площадь на одно рабочее место пользователей ПЭВМ с ВДТ на базе электронно-лучевой трубки (ЭЛТ) должна составлять не менее 6 м^2 и с ВДТ на базе плоских дискретных экранов (жидкокристаллические, плазменные) 4,5 м^2.

Для внутренней отделки интерьера помещений с ПЭВМ должны использоваться диффузионно-отражающие материалы с коэффициентом отражения от потолка – 0.7-0.8; для стен – 0.5-0.6; для пола – 0.3-0.5.

Основным документом, определяющим условия труда на персональных ЭВМ, являются «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы». Санитарные нормы и правила СанПиН 2.2.2/2.4.1340-03, которые были введены 30 июня 2003 года.

В кабинете, который является местом работы инженера-программиста, параметры микроклимата находятся в пределах нормы.

Для оценки соблюдения ПДУ шума необходим контроль (измерения и оценка). В случае превышения уровней необходимы мероприятия по защите от действия шума (защита временем, расстоянием, экранирование источника, либо рабочей зоны, замена оборудования, использование СИЗ).

Приведем расчет искусственного освещения для прямоугольного помещения, размерами: длина A = 5 м, ширина B = 6 м, высота H = 4 м, количество ламп N = 12 шт.

Определим расчетную высоту подвеса светильников над рабочей поверхностью (h) по формуле:

$$h = H - h_p - h_c, \tag{5.1}$$

где H – высота потолка в помещении, м; h_p – расстояние от пола до рабочей поверхности стола, м; h_c – расстояние от потолка до светильника, м.

Вычислим расчетную высоту подвеса светильников над рабочей поверхностью по формуле 5.1 для компьютерной аудитории кафедры программной инженерии:

$$h = 4 - 0{,}8 - 0{,}01 = 3{,}19 \text{ м.}$$

Индекс помещения определяется по формуле (5.2):

$$i = \frac{S}{h(A+B)}, \tag{5.2}$$

где S – площадь помещения, м^2; A – длина комнаты, м; B – ширина комнаты, м; h – высота подвеса светильников, м.

Индекс помещения для компьютерной аудитории кафедры программной инженерии:

$$i = \frac{30}{3{,}19(5+6)} = 0{,}83.$$

Исходя из того, что потолок в помещении чистый бетонный, а также свежепобеленные стены без окон, согласно методическим указаниям, примем коэффициенты отражения от стен ρ_c=*70%* и потолка ρ_n=*50%*. По таблице коэффициентов использования светового потока для соответствующих значений i, ρ_c, ρ_n, примем η=0,29.

Освещенность помещения рассчитывается по формуле:

$$E_\phi = \frac{n \cdot \eta \cdot \Phi}{S \cdot k_з \cdot z}; \tag{5.3}$$

где $Ф$ – световой поток светильника, лм; S – площадь помещения, м^2; $k_з$ – коэффициент неравномерности освещения; n – число светильников; η – коэффициент использования светового потока.

Коэффициент запаса k учитывает запыленность светильников и их износ. Для помещений с малым выделением пыли $k = 1{,}5$. Поправочный коэффициент z – это коэффициент неравномерности освещения. Для люминесцентных ламп $z = 1{,}1$. В помещении находятся светильники ЛВО 4×18 CSVT, с люминесцентными лампами типа L 18W/640 с потоком $F = 1200$ лм. Учитывая все параметры, рассмотренные выше, найдем освещенность (формула 5.3):

$$E_ф = \frac{48 \cdot 0,29 \cdot 1200}{30 \cdot 1,5 \cdot 1,1} = 337 \text{ лк.}$$

В рассматриваемом помещении освещенность должна составлять 300 лк согласно СНиП 23-05-95. В данном помещении освещенность находится в пределах нормы, следовательно дополнительные источники света не нужны.

Согласно правилам устройства электроустановок, компьютерный зал по степени опасности поражения электрическим током можно отнести к классу помещений без повышенной опасности.

5.3 Экологическая безопасность

В ходе выполнения работ по исследованию, разработке и дальнейшей эксплуатации методики не происходит выбросов загрязняющих веществ в атмосферу и гидросферу. Однако, люминесцентные лампы, применяющиеся для освещения рабочих мест, содержат ртуть, чрезвычайно токсичный металл, который может вызвать загрязнение литосферы, гидросферы и атмосферы и нанести тяжкий вред здоровью.

Их эксплуатация требует осторожности и четкого выполнения инструкции по обращению с данным отходом (код отхода 35330100 13 01 1, класс опасности – 1 (ФККО). В данной лампе содержится опасное вещество ртуть в газообразном состоянии.

Замену ртутьсодержащей лампы осуществляет лицо, ответственное за сбор и хранение ламп (обученное по электробезопасности и правилам обращения с отходом). Отработанные люминесцентные лампы сдаются на полигон токсичных отходов для захоронения. Запрещается сваливать отработанные люминесцентные лампы с мусором (Постановление Правительства РФ от 3 сентября 2010 г. N 681).

Бытовой мусор помещений организаций несортированный, образованный в результате деятельности работников предприятия (код отхода 91200400 01 00 4). Агрегатное состояние отхода твердое; основные компоненты: бумага и древесина, металлы, пластмассы и др. (Федеральный классификационный каталог отходов). Для сбора мусора рабочее место оснащается урной. При заполнении урны, мусор выносится в контейнер бытовых отходов. Предприятие заключает договор с коммунальным хозяйством по вывозу и размещению мусора на организованных свалках.

5.4 Безопасность в чрезвычайных ситуациях

5.4.1Анализ вероятных ЧС

В принципе, перечень возможных ЧС на объекте исследования может быть достаточно широк. Ограничиваясь местоположением объекта и условиями его эксплуатации, его можно представить следующим (ориентировочным) вариантом:

- наводнение;
- удар молнии;
- пожар на объекте;
- взрыв.

Наиболее вероятная ЧС в рассматриваемом помещении – пожар. Помещение, в котором велась работа, по степени пожаробезопасности относится к категории Д – негорючие вещества и материалы в холодном состоянии.

Рабочее место программиста должно соответствовать требованиям ФЗ Технический регламент по ПБ и норм пожарной безопасности (НПБ 105-03) и удовлетворять требованиям по предотвращению и тушению пожара по ГОСТ 12.1.004-91 и СНиП 21-01-97.

5.4.2Анализ причин, которые могут вызвать ЧС

Пожар в помещении оператора может возникнуть вследствие причин неэлектрического и электрического характера.

К причинам неэлектрического характера относятся халатное и неосторожное обращение с огнем (курение, оставление без присмотра нагревательных приборов).

К причинам электрического характера относятся: короткое замыкание, перегрузка проводов, большое переходное сопротивление, искрение, статическое электричество.

Короткое замыкание может возникнуть вследствие ошибки при проектировании, старения изоляции, увлажнения изоляции, механической перегрузки.

5.4.3Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС

К мерам по предупреждению пожара согласно относятся такие профилактические мероприятия, как (N 123-ФЗ):

- соблюдение эксплуатационных норм оборудования;
- обучение персонала правилам техники безопасности;
- издание противопожарных инструкций, планов эвакуации.

Пожарная защита должна обеспечиваться применением средств пожаротушения, а также применением автоматических установок пожарной сигнализации.

Должны быть приняты следующие меры противопожарной безопасности:

- обеспечение эффективного удаления дыма;

– обеспечение правильных путей эвакуации;
– наличие огнетушителей и пожарной сигнализации;
– соблюдение всех противопожарных требований к системам отопления и кондиционирования воздуха.

Согласно Правилам пожарной безопасности, в Российской Федерации ППБ 01-2003 (п. 16) в зданиях и сооружениях (кроме жилых домов) при единовременном нахождении на этаже более 10 человек должны быть разработаны и на видных местах вывешены планы (схемы) эвакуации людей в случае пожара.

Помещение, в котором выполнялась работа, входит в общий план эвакуации этажа, который предусматривает выход из всех помещений этажа в основной или запасной эвакуационные выходы здания. Эвакуация проводится согласно плану эвакуации, который выставлен на всеобщее обозрение в нескольких местах на каждом этаже (рис. 5.1).

В каждом кабинете установлен углекислотный огнетушитель ОУ-2 и табличка с указанием лица, ответственного за пожарную безопасность.

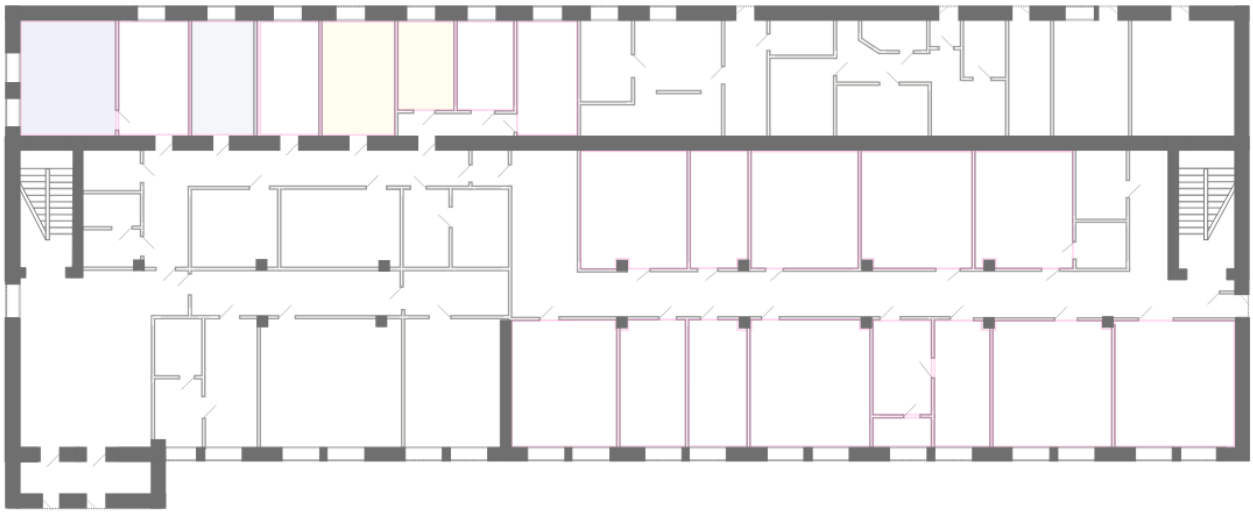

Рисунок 5.1 – План эвакуации при пожаре

Необходимыми действиями в результате возникшей ЧС и мерами по ликвидации её последствий являются (N 123-ФЗ):

1. Передать сигнал «Тревога» голосом, задействовать систему оповещения людей о пожаре.

2. Сообщить по телефону 01, с сотового 010 адрес объекта, место возникновения пожара, свою фамилию. Сообщить по телефону 03, с сотового 030 адрес объекта, что случилось, информацию о пострадавших, свою фамилию, оказать помощь пострадавшим.
3. Открыть все эвакуационные выходы, направить людей к эвакуационным выходам согласно знакам направления движения.
4. Отключить от электропитания оборудование, механизмы и т.п., обесточить помещение.
5. По возможности принять меры по тушению пожара используя средства противопожарной защиты.
6. По возможности предотвратить развитие аварии, обозначить место аварии.

Заключение

В данном разделе рассмотрены основные вопросы соблюдения прав персонала на труд, выполнения правил к безопасности труда, промышленной безопасности, экологии и ресурсосбережения. Установлено, что рабочее место исполнителя удовлетворяет требованиям безопасности и гигиены труда во время реализации проекта, а также вредное воздействие объекта исследования на окружающую среду не превышает норму.

Заключение

В ходе выполнения выпускной квалификационной работы создана методика подготовки данных для построения кредитного скоринга, которая включает в себя обязательные этапы: разбиение данных, очистка данных, трансформация данных и выбор переменных. Полученная методика реализована в программных пакетах Python, SAS, SAS Enterprise Miner. Исследование методики проводилось на примере анкетных данных заемщиков.

Проведено сравнение точности результатов, полученных в различных пакетах, и результатов классификации без подготовки данных и с применением предложенной методики подготовки данных.

Качественно во всех случаях применение методики повышает точность полученных результатов на 10-18%. Наибольшую точность (75%) демонстрирует решение, полученное с помощью SAS Enterprise Miner.

В будущем планируется исследование методики на большем количестве данных и в дальнейшем внедрение.

Список публикаций студента

1. Inkhireeva T. A. , Zimin V. P. Quasianalytical solution of inhomogeneous differential equation with cubic nonlinearity // Advances in Computer Science Research. - 2017 - Vol. 72. - p. 103-107
2. Kazakyavichyus I.S., Inkhireeva T. A. Gender recognition by voice // Электронные средства и системы управления: материалы докладов XIV Международной научно-практической конференции: в 2 ч. – Ч. 2., Томск, 28-30 Ноября 2018. - Томск: В-Спектр, 2018 - С. 282-286
3. Инхиреева Т. А. , Козловских А. В. Квазианалитическое решение неоднородного дифференциального уравнения с кубической нелинейностью // Молодежь и современные информационные технологии: сборник трудов XV Международной научно- практической конференции студентов, аспирантов и молодых ученых , Томск, 4-7 Декабря 2017. - Томск: ТПУ, 2018 - С. 43-44
4. Inkhireeva T.A. Data mining classification techniques for credit scoring in banks // Математическое и программное обеспечение информационных, технических и экономических систем: материалы VI международной молодежной научной конференции, Томск, 24-26 мая 2018 г. - Томск: ТГУ, 2018 - С. 362-365

Список использованных источников

1. Сергеевич С.А. Построение скоринговых карт с использованием модели логистической регрессии // Интернет-журнал Науковедение. 2014. Vol. 2.
2. Anshu B. Data Preprocessing Techniques for Data Mining // Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets. New Delhi, 2011. P. 6.
3. Abbott D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Indianapolis: Wiley, 2014. 427 p.
4. Полищук, Ф.С., Романов А.Ю. КРЕДИТНЫЙ СКОРИНГ: РАЗРАБОТКА РЕЙТИНГОВОЙ СИСТЕМЫ ОЦЕНКИ РИСКА КРЕДИТОВАНИЯ ФИЗИЧЕСКИХ ЛИЦ // Новые информационные технологии в автоматизированных системах. 2016. Vol. 19.
5. Федресурс. Единый федеральный реестр юридически значимых сведений о фактах деятельности юридических лиц, индивидуальных предпринимателей и иных субъектов экономической деятельности [Electronic resource] // В России за год число граждан-банкротов удвоилось. 2018.
6. Филатова Ю. Число несостоятельных граждан в России выросло в 1,5 раза, потенциальных банкротов – на 6%. Москва, 2018. 4 p.
7. Piatetsky G. Knowledge Discovery Nuggets [Electronic resource] // CRISP-DM, still the top methodology for analytics, data mining, or data science projects. 2014. P. 1. URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (accessed: 25.05.2019).
8. IBM Corporation. IBM SPSS Modeler CRISP-DM Guide. Armonk, 2011. 45 p.
9. SAS Institute Inc. Introduction to SEMMA [Electronic resource]. 2018. URL: https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=15.1&locale=en.
10. Ng A. Machine learning yearning. 5th ed. deeplearning.ai, 2018. 116 p.
11. Pedregosa F. et al. Scikit-learn: Machine Learning in {P}ython // J. Mach. Learn.

Res. 2011. Vol. 12. P. 2825–2830.

12. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 2001. Vol. 14.
13. Little R.J.A. A Test of Missing Completely at Random for Multivariate Data with Missing Values // J. Am. Stat. Assoc. Taylor & Francis, 1988. Vol. 83, № 404. P. 1198–1202.
14. RUBIN D.B. Inference and missing data // Biometrika. 1976. Vol. 63, № 3. P. 581–592.
15. Moritz S. et al. Comparison of different Methods for Univariate Time Series Imputation in R.
16. SAS Institute Inc. Building Credit Scorecards Using Credit Scoring for SAS Enterprise Miner. Cary, 2014. 21 p.
17. Zekic-Susac M., Sarlija N., Bensic M. Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models // 26th International Conference on Information Technology Interfaces, 2004. 2004. P. 265-270 Vol.1.
18. Svolba G. Data Preparation for Analytics Using SAS. SAS Institute Inc., 2015. 440 p.
19. Tischler R., Grosser T. Data Preparation - Refining Raw Data into Value. CXP Group, 2017. 43 p.
20. Huang J. et al. An Empirical Analysis of Three-Stage Data-Preprocessing for Analogy-Based Software Effort Estimation on the ISBSG Data // 2017 IEEE International Conference on Software Quality, Reliability and Security (QRS). 2017. P. 442–449.
21. Nalić J., Švraka A. Importance of data pre-processing in credit scoring models based on data mining approaches // 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018. P. 1046–1051.
22. García V., Marqués A.I., Sánchez J.S. Improving Risk Predictions by Preprocessing Imbalanced Credit Data // Neural Inf. Process. 2012. Vol. 7664.

Приложение А

(справочное)

Data preparation for credit scoring

Студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Инхиреева Татьяна Александровна		

Консультант ОИТ ИШИТР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Соколова В.В.	к.т.н.		

Консультант – лингвист ОИЯ ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИЯ ШБИП	Диденко А.В.	к.ф.н.		

Literature review

There are numerous studies and publications about building scoring cards based on logistic regression. SAS Institute, the market leader of commercial scorecard building, describes in [16] best practices on scorecard building using SAS Enterprise Miner. Authors consider the whole process of scorecard building from sampling to model assessment based on SEMMA approach. M. Zekic-Susac, N. Sarlija and M. Bensic in [17] compare logistic regression, neural networks (NNs), and CART decision trees for scoring cards building on small datasets.

Data preparation methods are discussed in details in multiple papers. G.Svolba in [18] describes data preparation from business point of view. The author gives advice on preprocessing data for analysis in SAS. R.Tischler and T.Grosser in [19] investigate the reasons behind employing data preparation in companies and their expectations. J.Huang, Y.F.Li, J.W.Keung, Y.T.Yu and W.K.Chan in [20] show that three-staged data preprocessing, that includes missing data imputation, data normalization and feature selection, drastically increases classification accuracy. Authors give advice on using advanced imputation techniques.

There are only few papers dedicated to data preparation for credit scoring with logistic regression. Unfortunately, none of them is free. There is only one paper concerning data preparation for credit scoring. In [21] J.Nalić and A.Švraka consider data preprocessing steps for scoring models based on support vectors machine, naïve Bayes, generalized linear model and decision tree algorithms. Authors suggest the following data preparation pipeline: data reduction, data aggregation, data cleaning, variables binning, data transformation. They implement the model in Oracle Data Miner. The results of the study indicate that data preprocessing significantly increases accuracy of scoring models.

Nevertheless, there are some studies on data preprocessing substeps. A number of resampling techniques for dealing with imbalanced credit dataset is studied in [22]. Garcia, Marques and Sanches find out that using resampling techniques improves the results of classification models.

Introduction

One of the most common services provided by financial institutions is loan granting. Loan portfolio risk analysis is based on individual risk assessment. In practice, there are two methods to perform credit risk assessment. The first method is expert assessment, the second one is credit scoring system. Nowadays credit scoring systems dominate over experts [1]. Credit scoring is an automated system, which predicts the probability of a potential borrower being delinquent or not. The system applies a mathematical model to historical banking data in order to determine if the borrower reimburses the credit on time. The model uses such information about a borrower as credit history, demographic profile, credit goal, etc.

A scoring card may be built with such machine learning techniques as neural networks, decision trees, linear and logistic regression. In this work we consider logistic regression-based model, which is currently the most commonly used for building scoring cards due to accuracy, interpretability and capability to estimate probability.

Quality of input data is crucial for consequent analysis, that is why data preparation is a vital step of data analysis process. However, it is often ignored, which leads to decrease of accuracy. Data obtained during data gathering step is often flawed with missing or impossible values, duplicates or inconsistent value combinations. Data may have various formats or have undesirable properties (multicollinearity, correlation, non-normal distribution). Even modern methods are unable to show good results in such a case.

Data preparation, despite its importance, is often totally or partially ignored. Analysts often omit some steps, regardless of the fact that each step increases accuracy of the credit scoring model and decreases financial loss. General advice on data preparation is presented in [2], [3]. In [1] authors also describe data preparation as a part of a scoring cards building process.

The goal of the present work is to develop and study data preparation methodology for credit scoring. In order to accomplish the goal, the following tasks must be performed:

1. To choose and read literature on given topic.
2. To study data preparation methodologies, that are used for credit scoring.
3. To implement the methodology in Python, SAS, SAS Enterprise Miner.
4. To check correctness of the implementations.
5. To compare the results of different implementations.

The object of the study as a benchmark problem is borrowers' creditworthiness data. The subject of the study is data preparation methodology for credit scoring. The methodology may be used to increase accuracy of credit scoring in banks.

Proposed data preparation methodology

Data Partition

Splitting initial historical dataset into two or three independent subsets is required for building adequate model and benchmarking. The number of splits depends on the number prediction models.

In case of one prediction model, initial data is split into train set and test set. The model parameters are estimated on the train set while test set is used for assessing model performance. Train and test set split ratio according to commonly used rule of thumbs is 70-80% to 30-20%. This ratio is defined by the initial data volume. Model trained on small dataset has great variance, which means that its accuracy on different datasets differs dramatically. For a mid-sized dataset (thousands on tens thousands of observations) standard solution is using 70-80% to 30-20% split ratio. For a big dataset the train set may be smaller (about 50-60%) and the test set bigger (50-40%) accordingly. This split ratio is considered in case of computationally intensive models. Small datasets (hundreds or thousands of observations) require cross-validation [10].

Cross-validation is a method of splitting initial dataset into train and test set in case of data insufficiency. Cross-validation starts with splitting initial dataset into train and test sets. Test set is held out, while train set is split into k folds. The model is trained on k-1 folds, the rest of data is used for validation [11].

Provided multiple models, initial data is split into train, validation and test set. This helps to deal with overfitting, which is optimistically biased model estimate. Validation set is used for models' results comparison and model selection. Standard split ratio for a mid-sized dataset in 60, 20 and 20%. The recommendations concerning split ratio selection given above also apply in this case [12].

All the data split schemes mentioned above are presented in fig. A.1.

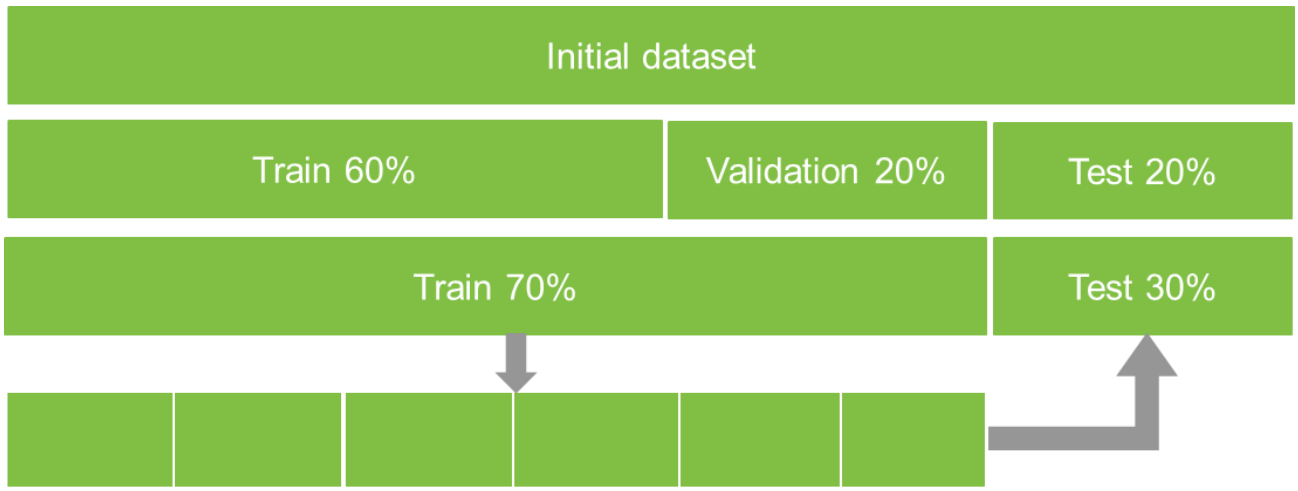

Figure A.1 – Data partition.

Splitting data always raises a problem of representativeness. All the subsets derived from initial dataset must belong to the same distribution, i.e. have the same target variable classes ratio.

Data cleaning

Data cleaning refers to removing duplicates, fixing wrong and inconsistent data, dealing with outliers and missing values. Solving those problems leads to improving quality of prediction.

Duplicates

Presence of duplicates influences regression coefficients by bringing unnecessary ambiguity into model, thus increasing model variance. That is why duplicates are removed from the dataset.

Outliers

Outliers are unusual values that are separated from the main body of the distribution, typically as measured by standard deviations from the mean or by the interquartile range. Logistic regression is sensitive to outliers, therefore fixing them is crucial for consequent analysis.

The easiest way to detect outliers is to consider all the observations behind interquartile range as outliers. This can be done visually with a box plot (whiskers plot). The box plot shows mean or median, interquartile range, minimal and maximal values and outliers. The box plot example is shown in fig. A.2.

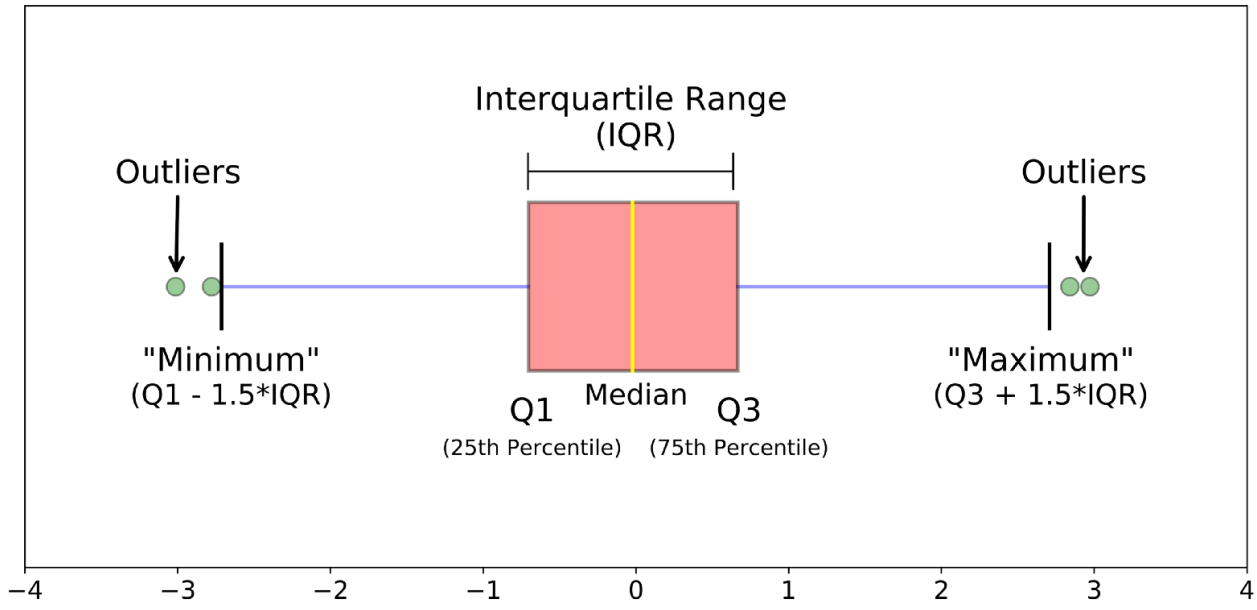

Figure A.2 – Box plot

It is considerably harder to detect multidimensional outliers. Two-dimensional outliers can be detected with a scatter plot. Constructing scatter plots for all variable pairs is an efficient way to detect all two-dimensional outliers. Outlier in the scatter plot is presented in fig. A.3. Multidimensional outliers can be detected with isolation forests, dbscan and clustering methods.

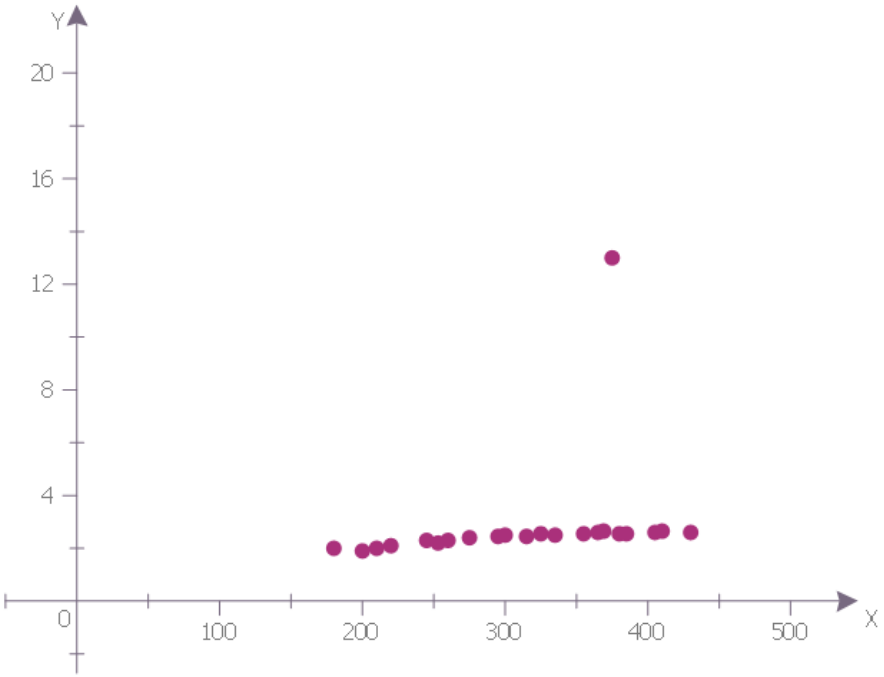

Figure A.3 – Scatter plot

Outliers in categorical variables are rare values, which can be detected with histogram. Small number of outliers is usually removed or replaced with mean or median. Separating outliers and creating a separate model for them is considered in case of multiple outliers, because it may indicate novelty in data. Data transformation of discretization is helpful for dealing with numerical outliers.

Developing an algorithm of data preprocessing methodology

An algorithm of data preparation methodology is present in Appendix B. The algorithm takes as an input D, which is borrowers' creditworthiness data. D is an m × n matrix, where m is a number of rows (observations) and n is a number of columns (variables).

10. Initial data:
    - 10.1. D is borrowers' creditworthiness data.
11. Data partition:
    - 11.1. if there are a few models, which implies model selection and n is big enough, split D into train (60%), validation (20%) and test (20%) sets;
    - 11.2. if there is only one model or n is not big enough, split D into train (70%) and test (30%).
12. Data cleaning:
    - 12.1. eliminate duplicating rows;
    - 12.2. outliers:
      - 12.2.1. one-dimensional outliers detection;
      - 12.2.2. multi-dimensional outliers detection;
      - 12.2.3. if there are only a few outliers, delete them;
      - 12.2.4. if number of outliers is big enough, subset them;
    - 12.3. inconsistent data correction;
    - 12.4. missing data:
      - 12.4.1. find the source of missing data;
      - 12.4.2. if the variable has 5% or less values missing (considered random), drop the observations;
      - 12.4.3. if the variable has 5% to 50% of missing values, find out why and fill;
      - 12.4.4. if the variable has 50% or more missing values, drop the variable.
13. Data transformation:
    - 13.1. format unification;
    - 13.2. binning:
      - 13.2.1. binning continuous variables into groups by quantiles;
      - 13.2.2. calculate WOE for each group;
      - 13.2.3. if a group is homogeneous by target variable or WOE trend in different groups changes (increase and decrease or vice versa), merge neighbor groups;
    - 13.3. scaling:
      - 13.3.1. if variable distribution is close to normal, standardization;

13.3.2. otherwise min-max normalization to {0, 1} range.

14. Feature selection:

14.1. multicollinearity:

14.1.1. if VIF is above 5, drop the variable or use principal components analysis to construct new uncorrelated features instead of initial variables;

14.2. normality:

14.2.1. normality test with Shapiro-Wilk test, Asymmetry and Excess test, D'Agostino test etc. Visual analysis with histograms;

14.2.2. transform all non-normal data into normal data.

14.3. Information capacity:

14.3.1. drop variables with near-zero variance;

14.3.2. select variables according to chi-square criterion and IV;

14.3.3. use wrapper methods for variable selection.