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)

Школа: Инженерная школа информационных технологий и робототехники
Направление подготовки: 09.03.04 «Программная инженерия»
Отделение школы (НОЦ): Отделение информационных технологий

## БАКАЛАВРСКАЯ РАБОТА

| Тема работы |
|---|
| Оценка рисков заёмщиков потребительского кредитования |

УДК: 004.65:005.52:005.334:336.774

Студент

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8К71 | Андреева Анастасия Борисовна | | |

Руководитель

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Губин Евгений Иванович | к.ф.-м.н. | | |

## КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Маланина Вероника Анатольевна | к.э.н. | | |

По разделу «Социальная ответственность»

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Ассистент | Черемискина Мария Сергеевна | | | |

## ДОПУСТИТЬ К ЗАЩИТЕ:

| Руководитель ООП | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Чердынцев Евгений Сергеевич | к.т.н. | | |

Томск – 2021г.

## ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

| **Код результата** | **Результат обучения (выпускник должен быть готов)** |
|---|---|
| Р1 | Применять базовые и специальные естественнонаучные и математические знания в области информатики и вычислительной техники, достаточные для комплексной инженерной деятельности. |
| Р2 | Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач. |
| Р3 | Ставить и решать задачи комплексного анализа, связанные с созданием аппаратно-программных средств информационных и автоматизированных систем, с использованием базовых и специальных знаний, современных аналитических методов и моделей. |
| Р4 | Разрабатывать программные и аппаратные средства (системы, устройства, блоки, программы, базы данных и т. п.) в соответствии с техническим заданием и с использованием средств автоматизации проектирования. |
| Р5 | Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания аппаратных и программных средств информационных и автоматизированных систем. |
| Р6 | Внедрять, эксплуатировать и обслуживать современные программноаппаратные комплексы, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды. |
| Р7 | Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности. |
| Р8 | Владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности. |

| | |
|---|---|
| P9 | Эффективно работать индивидуально и в качестве члена группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации. |
| P10 | Демонстрировать знания правовых, социальных, экономических и культурных аспектов комплексной инженерной деятельности. |
| P11 | Демонстрировать способность к самостоятельному обучению в течение всей жизни и непрерывному самосовершенствованию в инженерной профессии. |

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)

Школа: Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность): 09.03.04 «Программная инженерия»
Отделение школы (НОЦ): Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Чердынцев Е.С.
(Подпись) (Дата) (Ф.И.О.)

**ЗАДАНИЕ**
**на выполнение выпускной квалификационной работы**

В форме:

| Бакалаврской работы |
|---|

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

| Группа | ФИО |
|---|---|
| 8К71 | Андреевой Анастасии Борисовне |

Тема работы:

| Оценка рисков заёмщиков потребительского кредитования | |
|---|---|
| Утверждена приказом директора (дата, номер) | №32-2/с от 01.02.2021 г. |

| Срок сдачи студентом выполненной работы: | 11.06.2021 г. |
|---|---|

**ТЕХНИЧЕСКОЕ ЗАДАНИЕ:**

| | |
|---|---|
| **Исходные данные к работе** *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).* | Объектом проектирования в данной работе является скоринговая система, которая будет использоваться кредитными организациями для управления рисками при кредитовании физических лиц. |

| | |
|---|---|
| **Перечень подлежащих исследованию, проектированию и разработке вопросов**<br>*(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).* | 1. Обзор предметной области<br>2. Выбор средств разработки<br>3. Анализ и подготовка исходных данных для обработки<br>4. Разработка скоринговой карты с помощью алгоритма логистической регрессии<br>5. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение<br>6. Социальная ответственность |
| **Перечень графического материала**<br>*(с точным указанием обязательных чертежей)* | 1. Фрагменты кода Python<br>2. Рисунки, демонстрирующие результаты<br>3. Диаграммы рассеяния<br>4. Диаграмма Ганта |

**Консультанты по разделам выпускной квалификационной работы**
*(с указанием разделов)*

| Раздел | Консультант |
|---|---|
| Финансовый менеджмент, ресурсоэффективность и ресурсосбережение | Маланина Вероника Анатольевна |
| Социальная ответственность | Черемискина Мария Сергеевна |

| | |
|---|---|
| **Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику** | 25.01.2021 г. |

**Задание выдал руководитель:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Губин Евгений Иванович | к.ф.-м.н | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8К71 | Андреева Анастасия Борисовна | | |

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)

Школа: Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность): 09.03.04 «Программная инженерия»
Уровень образования: Бакалавр
Отделение школы (НОЦ): Отделение информационных технологий
Период выполнения: осенний / весенний семестр 2020 / 2021 учебного года

Форма представления работы:

| Бакалаврская работа |
|---|

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

## КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
## выполнения выпускной квалификационной работы

| | |
|---|---|
| Срок сдачи студентом выполненной работы: | 11.06.2021 г. |

| Дата контроля | Название раздела (модуля) / вид работы (исследования) | Максимальный балл раздела (модуля) |
|---|---|---|
| 16.06.2021 г. | Основная часть | 75 |
| 16.06.2021 г. | Финансовый менеджмент, ресурсоэффективность и ресурсосбережение | 15 |
| 16.06.2021 г. | Социальная ответственность | 10 |

**СОСТАВИЛ:**
**Руководитель ВКР**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Губин Евгений Иванович | к.ф.-м.н. | | |

**СОГЛАСОВАНО:**
**Руководитель ООП**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Чердынцев Евгений Сергеевич | к.т.н. | | |

**ЗАДАНИЕ ДЛЯ РАЗДЕЛА**
**«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»**

Студенту:

| Группа | ФИО |
|---|---|
| 8К71 | Андреевой Анастасии Борисовне |

| Школа | ИШИТР | Отделение школы (НОЦ) | ОИТ |
|---|---|---|---|
| Уровень образования | Бакалавриат | Направление/специальность | 09.03.04 Программная инженерия |

| **Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:** | |
|---|---|
| 1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих* | *Материальные затраты – 369 руб.*<br>*Затраты на специальное оборудование – 70053 руб.*<br>*Затраты на заработную плату – 13983,75 руб.*<br>*Затраты на отчисления во внебюджетные фонды – 4193,09 руб* |
| 2. *Нормы и нормативы расходования ресурсов* | *Бюджет проекта не более 110000 руб., в том числе затраты на оплату труда не более 20000 руб.*<br>*Значение показателя интегральной ресурсоэффективности - не менее 3.5 баллов из 5.* |
| 3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования* | *Районный коэффициент – 1,3*<br>*Коэффициент дополнительной заработной платы – 0,13*<br>*Коэффициент отчислений во внебюджетные фонды – 0,302*<br>*Коэффициент накладных расходов – 0,16* |
| **Перечень вопросов, подлежащих исследованию, проектированию и разработке:** | |
| 1. *Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения* | *1. Описание потребителей продукта.*<br>*2. Анализ конкурентных технических решений.*<br>*3. SWOT-анализ.* |
| 2. *Планирование и формирование бюджета научных исследований* | *1.Описание структуры работ в рамках научного исследования.*<br>*2. Определение трудоемкости выполнения работ и разработка графика проведения научного исследования.*<br>*3. Подсчет бюджета проекта.* |
| 3. *Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования* | *1.Оценка сравнительной эффективности исследования.* |
| **Перечень графического материала** *(с точным указанием обязательных чертежей)***:** | |
| 1. *Матрица SWOT*<br>2. *График проведения и бюджет НИ*<br>3. *Оценка ресурсной, финансовой и экономической эффективности НИ* | |

| **Дата выдачи задания для раздела по линейному графику** | 25.01.2021 |
|---|---|

**Задание выдал консультант:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| | | | | |

| | | | | |
|---|---|---|---|---|
| доцент ОСГН ШБИП ТПУ | Маланина Вероника Анатольевна | к.э.н. | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8К71 | Андреева Анастасия Борисовна | | |

# ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

| Группа | ФИО |
|---|---|
| 8К71 | Андреевой Анастасии Борисовне |

| Школа | ИШИТР | Отделение (НОЦ) | Отделение информационных технологий |
|---|---|---|---|
| Уровень образования | Бакалавриат | Направление/специальность | 09.04.03. «Программная инженерия» |

Тема ВКР:

| **Оценка рисков заёмщиков потребительского кредитования** | |
|---|---|
| **Исходные данные к разделу «Социальная ответственность»:** | |
| 1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения | Объект исследования: рабочая зона сотрудника.<br>Область применения: офисное помещение. |
| Перечень вопросов, подлежащих исследованию, проектированию и разработке: | |
| **1. Правовые и организационные вопросы обеспечения безопасности:**<br>– специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства;<br>– организационные мероприятия при компоновке рабочей зоны. | − Рабочее место при выполнении работ сидя регулируется ГОСТом 12.2.032-78<br>− Права работников на охрану труда регулируются Трудовым кодексом РФ |
| **2. Производственная безопасность:**<br>2.1. Анализ выявленных вредных и опасных факторов<br>2.2. Обоснование мероприятий по снижению воздействия | - недостаток освещенности рабочей зоны;<br>- отклонение показателей микроклимата рабочей зоны;<br>- повышенный уровень шума на рабочем месте;<br>- повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека;<br>- монотонность трудового процесса; |
| **3. Экологическая безопасность:** | Атмосфера: захоронения отходов;<br>Гидросфера: загрязнение детергентами;<br>Литосфера: загрязнение утилизированной техникой, бумагой. |
| **4. Безопасность в чрезвычайных ситуациях:** | Возможные ЧС: пожары, землетрясения, экстремальные погодные условия (очень низкая или высокая температура воздуха, снежная буря, ураган) и т.п.<br>Наиболее типичная ЧС: |

| | |
|---|---|
| | пожар. |

| | |
|---|---|
| **Дата выдачи задания для раздела по линейному графику** | **29.04.2021** |

**Задание выдал консультант:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Ассистент | Черемискина Мария Сергеевна | - | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8К71 | Андреева Анастасия Борисовна | | |

# РЕФЕРАТ

Выпускная квалификационная работа содержит 34 рисунка, 20 таблиц, 12 формул, 18 источников, 4 приложения.

Ключевые слова: скоринг, скоринговая карта, кредитование, анализ данных, подготовка данных, машинное обучение, логистическая регрессия.

Объектом исследования выпускной квалификационной работы являются исторические банковские данные о прошлых заемщиках.

Целью исследования выпускной квалификационной работы является разработка алгоритма для предварительной обработки данных с последующим анализом и построением скоринговой модели добросовестных заемщиков.

Область применения: финансовые, банковские аналитические сферы.

В результате исследования с помощью Python были проанализированы и обработаны большие данные о заемщиках. По итогу работы, алгоритм машинного обучения, использующий логистическую регрессию, выдал высокую точность прогноза при работе с проанализированными и предобработанными данными. К признакам, характеризующим заемщика, подобраны соответствующие весовые коэффициенты.

**ОГЛАВЛЕНИЕ**

## ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ И ТЕРМИНОВ

- Big Data («большие данные») - обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами.

- Выброс (англ. outlier) - в статистике результат измерения, выделяющийся из общей выборки.

- Кредит – сумма денег, предоставляемая банком или кредитной организацией заемщику на условиях, указанных в кредитном договоре.

- Скоринговая карта - это набор утвержденных банком определенных характеристик и соответствующих весовых коэффициентов (в баллах).

- CSV (от англ. Comma-Separated Values - значения, разделённые запятыми) - текстовый формат, предназначенный для представления табличных данных. Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделенных запятыми.

# ВВЕДЕНИЕ

Одновременно с ростом кредитования в России возникает проблема увеличения рисков невыполнения кредитных обязательств заемщиками (дефолтов). Рост числа невозвратных кредитов представляет для всех банковских и небанковских структур серьезную проблему.

Поскольку основой банковского кредитования является принцип получения дохода от процента за предоставление заемных средств, прибыль банкиры получают только при возврате денег надежными заемщиками, и чем у кредитной организации таких клиентов больше, тем выше доход.

В настоящее время скоринг все чаще используется для оценки кредитного риска. Применение скоринг-систем коммерческими банками или кредитными учреждениями, занимающимися кредитованием физических лиц, дает им существенные преимущества, которые позволяют улучшить качество кредитного портфеля, что положительного сказывается на финансовом состоянии организации.

Скоринг-системы – достаточно сложные информационно-технологические системы, которые постоянно требуют обновлений и улучшений. Скоринговые данные и рейтинги клиентов со временем устаревают, что требует постоянного мониторинга качества используемых моделей, которые со временем перестраиваются для соответствия текущему экономическому состоянию.

Основными составляющими скоринг-анализа являются исходные данные (анкетные данные заемщика), математические модели, заложенные в основу скоринга, технологические составляющие, использование данных БКИ, миграционных служб, полиции и др.

Сегодня, в условиях финансового и экономического кризиса, систематическая модернизация скогинговых систем – необходимая мера для работы любой кредитной организации с наиболее точной и актуальной

информацией о клиентах. Данные и рейтинги клиентов становятся устаревшими, так как люди сменяются, а социально-экономические условия могут ухудшиться или улучшиться, влияя на их кредитоспособность. Качество окончательной оценки и, в конечном итоге, эффективность оценки рисков, прибыль кредитного портфеля значительно зависят от выбора исходных данных. Большую сложность представляют составление списка параметров оценки и определения соответствующих весовых коэффициентов, так как это и задаёт алгоритм, который в конечном счете будет использоваться для оценки анкетных данных клиентов.

В связи с существующими проблемами целью данной работы явилось повышение качества оценки потенциального заемщика, снижение рисков кредитных организаций, за счет создание скоринговой модели с собственными коэффициентами и параметрами оценки, при использовании современных технологических решений.

Для достижения цели были выделены следующие задачи:

1. Обзор истории появления и развития скоринга
2. Анализ предметной области
3. Анализ анкетных данных, включающий подготовку начальной информации о клиентах банка (исключающая пропуски, ошибки, выбросы и т.д), описательную статистику
4. Использование логистической регрессии, для определения весовых коэффициентов признаков на основе исторических данных о клиентах банка;
5. Разработка и описание скоринговой карты на основе полученных данных.

# ГЛАВА 1 ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

## 1.1 Понятие кредитного скоринга

Скоринг изначально представляет собой метод классификации популяции, представляющей для нас интерес, на определенные группы. Его использование происходит в том случае, если нам неизвестна характеристика, разделяющая эти группы (в нашем случае это способность погашения кредита потенциальным заемщиком), но известны связанные с ней характеристики. Идеи классификации популяции на группы были выработаны Рональдом Фишером в 1936 г. на примере растений. Однако уже в 1941 г. данный метод был впервые применен Дэвидом Дюраном при классификации кредитов на «хорошие» и «плохие».

Рост популярности кредитных карточек привел к необходимости расширить применение скоринга. Количество людей, которые ежедневно обращались в банк за кредитными картами, было настолько велико, что увеличению скорости процесса принятия решений о выдаче кредита помогла бы только его автоматизация. После того, как были введены скоринг-системы, уровень безнадежного долга снизился до 50%.

Кредитный скоринг – это математическая или статистическая модель, благодаря которой на основании информации, известной о потенциальном заемщике, банк может определить вероятность того, что заемщик погасит кредит в срок.

Скоринговая система получила свое название от английского слова «score», что в переводе означает счет или подсчет очков. Когда клиент банка запрашивает кредит, он проходит обязательное анкетирование.

## 1.2 Пример популярной скоринговой системы

Рассмотрим пример готового решения «SAS».

Здесь оценки основаны на ряде различных характеристик, таких как демография, информация о занятости и соотношение долга к доходу. Окончательный балл заявителя получается из суммы отдельных баллов.

| Characteristic | Value | Score |
|---|---|---|
| Age | 18 - 24 | 20 |
| | 25 - 45 | 35 |
| | 46+ | 50 |
| | | |
| Time at Current Address | 0 - 2 years | 20 |
| | 3 - 8 years | 30 |
| | 9+ years | 40 |
| | | |
| Residential Status | Owner | 40 |
| | Renting | 25 |
| | Other | 20 |

Рисунок 1 – Пример используемых характеристик в скоринг-системе SAS.

Окончательный результат классифицирует кандидата в определенную группу: хороших или плохих шансов. Затем он сравнивается с заранее определенной точкой отсечения, чтобы определить уровень риска заявителя. Допустим на 15 человек, получивших итоговый балл в диапазоне 520-559, приходится 1 «плохой» заемщик.

| Application Score | Good/Bad Odds | Decision |
|---|---|---|
| < 320 | 2/1 | Reject |
| 320 – 359 | 3/1 | Reject |
| 360 – 399 | 4/1 | Reject |
| 400 – 439 | 7/1 | Refer |
| 440 – 479 | 8/1 | Refer |
| 480 – 519 | 10/1 | Refer |
| 520 – 559 | 15/1 | Accept |
| 560 – 599 | 19/1 | Accept |
| 600+ | 25/1 | Accept |

Рисунок 2 – Принятие решения в SAS.

Далее сотрудник банка решает, следует ли предоставить клиенту кредит, отклонить его заявку или же установить какие-либо ограничения.

### 1.3 Преимущества применения скоринговых систем

Применение скоринг-систем коммерческими банками или кредитными учреждениями, занимающимися кредитованием физических лиц, дает им некоторые преимущества:

1. Усовершенствование качества кредитного портфеля кредитора;
2. Использование более совершенной технологии оценивания рисков;
3. Оказание положительного влияния на рыночные позиции банка и его финансовое состояние.
4. Существенное уменьшение коррупционной составляющей.

Повышение качества кредитного портфеля кредитора происходит благодаря уменьшению необоснованных отказов по заявкам на кредит. Кроме того, скоринг-система позволяет выявлять и отсеивать ненадежных или вызывающих сомнения заемщиков, таким образом, банку открывается возможность расширения клиентской базы без излишнего роста рисков, снижается количество невозвратных кредитов и просроченных платежей. Как результат, в обширной клиентской базе кредитора число надежных заемщиков становится намного выше.

Основой скоринговой системы является математический аппарат, он уменьшает затрачиваемое на принятие решения об одобрении кредита время, предоставляет более точную оценку заемщика, чем анализ данных банковским работником вручную. Предоставление точных данных автоматизированной системой помогает в построении и развитии банковского бизнеса.

Что касается финансового положения кредитной организации, применение скоринга предоставляет возможность снижения объема резервов по потерям по ссудам, которые формируются на случай возможных потерь в связи с невыполнением заемщиками своих договорных обязательств. Это приводит к более эффективносму использованию финансовых средств.

# ГЛАВА 2 ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ ДЛЯ ОБРАБОТКИ

## 2.1 Исходные информационные данные

В основе построения скоринговых карт лежат статистические модели. Именно качественная информация о заемщиках банка, качество исходных данных, в итоге является определяющим фактором точности прогнозирования и успеха разработанной скоринговой системы в целом.

Разработка скоринговой модели осуществляется посредством анализа прошлого кредитного опыта. Объем информации должен быть достаточен, количество данных может варьироваться в зависимости от конкретных моделей. Исходные данные зачастую содержат внутренние данные анкет заемщиков банка, однако могут также содержать внешние данные кредитных историй. Всё это сотни тысяч записей, поэтому можно утверждать, что наиболее долгим и трудоемким этапом процесса анализа больших объемов данных является именно процесс сбора и подготовки исходных данных. Порой он может занять 80% всего времени. Однако с появлением современного программного обеспечения, использующего статистические методики, временные и финансовые затраты на данном этапе могут значительно снизиться, а эффективность и качество конечных результатов повыситься.

В идеале модели скоринга должны использоваться для тех же кредитных продуктов, рыночного сектора и экономической ситуации, которые были заложены в основу данных о предыдущем кредитном опыте. Помимо этого, для разработки точной скоринг-модели исходные данные должны иметь определенную историческую давность, период, за который собираются данные. Историческая давность данных для построения модели определяется видом скоринга и видом кредитования, а также требованиям надзорных органов.

Кроме того, есть определенные типы клиентов, данные о которых также исключать из начальной информационной базы. Речь о нетипичных клиентах, таких как мошенники, сотрудники банка, VIP клиенты, умершие клиенты,

несовершеннолетние, двойные заявки, кредиты по украденным картам и др. Также из информационной базы следует исключать кредиты с аномально большими суммами, условиям погашения, отличающимися от стандартных, нетипичными целями займа. Дополнительным критерием отбора данных может быть вид кредитования, регион рынка для разрабатываемой скоринг-карты.

| Sex | DocType | Business | Education | CarStatus | AddIncome | Income | CreditSum | Age | DayOff | GB |
|---|---|---|---|---|---|---|---|---|---|---|
| Male | Driving lic | Industry | Secondary e | 1 | 3317 | 16589 | 500000 | 33 | 90+ | 1 |
| Female | other | Education | Secondary e | 0 | 912 | 4562 | 100000 | 999 | 0 | 0 |
| Female | other | Finance /E | Higher educ | 0 | 919 | 4597 | 100000 | 30 | 0 | 0 |
| | other | Retail | Secondary e | 0 | 4000 | 20000 | 280000 | 46 | 0 | 0 |
| Male | Driving lic | Industry | Secondary e | 1 | 12170 | 60850 | 500000 | 42 | 0 | 0 |
| Male | Driving lic | Industry | Secondary e | 1 | 12170 | 60850 | 500000 | 42 | 0 | 0 |
| Male | Foreign pa | Industry | Higher educ | 1 | 728 | 3640 | 100000 | 25 | 60_90 | 0 |
| Male | Drivine lic | Finance /E | Secondary e | 0 | 2468 | 12341 | 150000 | 43 | 0 | 0 |
| Male | Military ca | Industry | Secondary e | 0 | 824 | 4123 | 270000 | 36 | 0 | 0 |
| Female | other | Industry | Secondary e | 0 | 468 | 2340 | 136000 | 31 | 0 | 0 |
| Male | Drivine lic | Industry | Secondary e | 1 | 430 | 2150 | | 45 | 0 | 0 |
| Female | Driving lic | Education | Secondary e | 0 | 2468 | 12341 | 150000 | 22 | 0 | 0 |
| Female | Driving lic | Industry | Secondary e | 0 | 1170 | 5850 | 480000 | 31 | 30_90 | 0 |
| Male | Drivine lic | other | Higher educ | 0 | 1063 | 5317 | 183000 | 28 | 0 | 0 |
| Female | other | Education | Secondary e | 0 | 1027 | 5139 | 214000 | 36 | 0 | 0 |
| Female | other | Education | Secondary e | 0 | 1024 | 5123 | 15000000 | 43 | 0 | 0 |
| Male | other | Education | Higher educ | 0 | 3085 | 15429 | 200000 | 31 | более90 | 1 |
| Male | Military ca | IT and Tel | Secondary e | 1 | 1314 | 6572 | 130000 | 24 | 0 | 0 |
| Male | Drivine lic | Industry | Secondary e | 0 | 300 | 1500 | 400000 | 44 | 0 | 0 |
| Male | other | Industry | Secondary e | 0 | 0 | 0 | 100000 | 31 | 0-30 | 0 |
| Female | Driving lic | Retail | Higher educ | 1 | 3040 | 15202 | 500000 | 34 | 0-30 | 0 |
| Female | other | other | Higher educ | 0 | 1304 | 6523 | 500000 | 48 | 0 | 0 |

Рисунок 3 – Файл исходных данных с «грязными» строками.

Вот таким образом выглядит фрагмент файла (22 записи из 1501) исходных данных, характеризующих прошлых заемщиков. Требуется произвести подготовку исходных данных для дальнейшего решения задачи бинарной классификации потенциальных заемщиков банка методом логистической регрессии.

Исходными данными к работе являются исторические данные о кредитоспособности, содержащие 11 переменных, одна из которых – целевая (GB), и 1501 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.

## 2.2 Общая методика подготовки исходных данных для построения прогнозных моделей классификации

Как вы могли заметить из приведенного фрагмента файла исходных данных, строки могут содержать пустые или ошибочные значения атрибутов, подозрительные значения, также строки могут и вовсе дублироваться. Для того чтобы получить более точный прогноз платежеспособности заемщиков на выходе, требуется подготовить исходные данные, исключив всё вышеперечисленное и, потеряв как можно меньше исторической информации. Это очень важный этап анализа данных, так как здесь работает принцип, если в наборе данных содержатся некорректные данные на входе, будут плохие результаты в конце.

Для наглядности приведу тот же фрагмент файла исходных данных с выделенными ячейками с отсутствующими данными, ячейками с ошибками и выбросами, а также выделенными дублирующими строками.

| Sex | DocType | Business | Education | CarStatus | AddIncome | Income | CreditSum | Age | DayOff | GB |
|---|---|---|---|---|---|---|---|---|---|---|
| Male | Driving lic | Industry | Secondary e | 1 | 3317 | 16589 | 500000 | 33 | 90+ | 1 |
| Female | other | Education | Secondary e | 0 | 912 | 4562 | 100000 | 999 | 0 | 0 |
| Female | other | Finance /E | Higher educ | 0 | 919 | 4597 | 100000 | 30 | 0 | 0 |
| | other | Retail | Secondary e | 0 | 4000 | 20000 | 280000 | 46 | 0 | 0 |
| Male | Driving lic | Industry | Secondary e | 1 | 12170 | 60850 | 500000 | 42 | 0 | 0 |
| Male | Driving lic | Industry | Secondary e | 1 | 12170 | 60850 | 500000 | 42 | 0 | 0 |
| Male | Foreign pa | Industry | Higher educ | 1 | 728 | 3640 | 100000 | 25 | 60_90 | 0 |
| Male | Drivine lic | Finance /E | Secondary e | 0 | 2468 | 12341 | 150000 | 43 | 0 | 0 |
| Male | Military ca | Industry | Secondary e | 0 | 824 | 4123 | 270000 | 36 | 0 | 0 |
| Female | other | Industry | Secondary e | 0 | 468 | 2340 | 136000 | 31 | 0 | 0 |
| Male | Drivine lic | Industry | Secondary e | 1 | 430 | 2150 | | 45 | 0 | 0 |
| Female | Driving lic | Education | Secondary e | 0 | 2468 | 12341 | 150000 | 22 | 0 | 0 |
| Female | Driving lic | Industry | Secondary e | 0 | 1170 | 5850 | 480000 | 31 | 30_90 | 0 |
| Male | Drivine lic | other | Higher educ | 0 | 1063 | 5317 | 183000 | 28 | 0 | 0 |
| Female | other | Education | Secondary e | 0 | 1027 | 5139 | 214000 | 36 | 0 | 0 |
| Female | other | Education | Secondary e | 0 | 1024 | 5123 | 15000000 | 43 | 0 | 0 |
| Male | other | Education | Higher educ | 0 | 3085 | 15429 | 200000 | 31 | более90 | 1 |
| Male | Military ca | IT and Tel | Secondary e | 1 | 1314 | 6572 | 130000 | 24 | 0 | 0 |
| Male | Drivine lic | Industry | Secondary e | 0 | 300 | 1500 | 400000 | 44 | 0 | 0 |
| Male | other | Industry | Secondary e | 0 | 0 | 0 | 100000 | 31 | 0-30 | 0 |
| Female | Driving lic | Retail | Higher educ | 1 | 3040 | 15202 | 500000 | 34 | 0-30 | 0 |
| Female | other | other | Higher educ | 0 | 1304 | 6523 | 500000 | 48 | 0 | 0 |

Рисунок 4 – Проблемы файла исходных данных.

Этапы подготовки исходных данных для построения прогнозных моделей включают в себя следующие шаги:

1. проверку исходных данных на ошибки (mistakes);
2. на отсутствие данных (missing);
3. на выбросы данных (outliers);
4. на наличие дублирующих строк (duplicates);
5. на проверку исходных атрибутов на мультиколлинеарность;
6. трансформация исходных данных в цифровой формат (векторизация)
7. выбор целевой переменной.

В качестве базовых эксперты предлагают использовать следующие рекомендации для «отчистки» исходных данных при наличии отсутствующих или ошибочных данных:

- Если количество отсутствующих данных не превышает 5%, то при сохранении репрезентативности, эти строки с пропусками могут быть удалены.
- Если количество отсутствующих данных превышает 50%, то данный атрибут возможно удалить из дальнейшего анализа.
- Если количество отсутствующих данных находится в интервале 5% - 50%, то для численных атрибутов возможно несколько вариантов замены отсутствующих значений: средним значением, медианой, ближайшим соседом и др. Для категориальных атрибутов возможно использовать значения, наиболее часто встречающиеся либо ближайших соседей.

### 2.3 Выбор технологий

Для подготовки исходных данных был выбран язык программирования Python. Выбор был сделан в пользу данного языка, так как он бесплатный, и располагает множеством библиотек, предназначенных для машинного обучения. Помимо этого, у Python простой синтаксис и удобочитаемость, что

способствует быстроте тестирования сложных алгоритмов машинного обучения. Всё вышеперечисленное объясняет процветающее сообщество, поддерживаемое совместными инструментами, такими как, например, Google Colab.

В качестве среды разработки были использованы:

- PyCharm – это кросс-платформенная среда разработки, которая совместима с Windows, macOS, Linux. Мною использовалась платная версия PyCharm Professional Edition.
- JupyterLab – это интерактивная среда разработки для работы с блокнотами, кодом и данными.

В данной работе использованы такие Python библиотеки, как Pandas, scikit-learn и matplotlib.

### 2.3.1 Pandas для извлечения и подготовки данных

Pandas представляет собой известную библиотеку, которая предоставляет высокоуровневые структуры данных, простые в использовании и интуитивно понятные.

Данная библиотека содержит множество встроенных методов группировки, комбинирования и фильтрации данных, методов анализа временных рядов. Помимо этого, библиотека умеет извлекать и осуществлять операции с данными из различных источников, среди них базы данных SQL, файлы CSV, Excel и JSON.

### 2.3.2 Scikit-learn для работы с классическими алгоритмами машинного обучения

Scikit-learn является популярной библиотекой для машинного обучения. Она поддерживает большое количество алгоритмов обучения, как контролируемых, так и неконтролируемых. Среди них есть такие алгоритмы

линейная регрессия, логистическая регрессия, деревья принятия решений, кластеризация, k-means и другие.

Библиотека Scikit-learn основана на двух главных Python библиотеках – NumPy и SciPy. В ней содержится набор алгоритмов для распространенных задач машинного обучения и добычи данных, в их числе регрессия, кластеризация и классификация.

### 2.3.3 Matplotlib для визуализации данных

Matplotlib используется для извлечения пользы из всех имеющихся данных. Данная стандартная библиотека Python применяется для создания графиков. Библиотека Matplotlib низкоуровневая и требует много команд для генерации хорошо выглядящих фигур или графиков.

Однако данная библиотека достаточно гибкая. Используя ее команды, можно создать практически любой график, строить разнообразные диаграммы, будь то гистограммы и диаграммы рассеяния или же графики с не-декартовыми координатами.

Также библиотека поддерживает GUI-бэкенд во всех операционных системах, экспортирует графики в общеизвестных форматах (PDF, SVG, JPG, PNG, BMP, GIF).

## 2.4 Анализ и предобработка данных

Согласно описанию рассматриваемой задачи, данные содержат информацию о клиентах, запрашивающих кредит. Всего информации представлено о 1501 клиенте, число признаков о клиенте – 11.

Последний столбец содержит целевую функцию GB (good=0, bad=1) цифры 0 и 1, соответствующие тому, хороший это или плохой заемщик (т.е. 0 – отсутствие у заемщика задолженности более 90 дней, 1 – ее наличие).

Для удобства обращения к требуемым признакам, зададим столбцам имена, применив следующую команду:

```
data.columns = ['A' + str(i) for i in range(1, 11)] + ['GB']
```

Результат представлен на рисунке 5.

```
          A1              A2                          A3  ...   A9  A10  GB
0       Male  Driving licence                    Industry  ...   33  90+   1
1     Female            other                   Education  ...  999    0   0
2     Female            other  Finance /Banking/Insurance  ...   30    0   0
3        NaN            other                      Retail  ...   46    0   0
4       Male  Driving licence                    Industry  ...   42    0   0
...      ...              ...                         ...  ...  ...  ...  ..
1496    Male  Drivine licence                      Retail  ...   56    0   0
1497    Male  Drivine licence                    Industry  ...   32    0   0
1498  Female            other                    Industry  ...   40    0   0
1499  Female  Driving licence              IT and Telecoms  ...   30    0   0
1500    Male  Drivine licence                    Industry  ...   30    0   0

[1501 rows x 11 columns]
```

Рисунок 5 – Данные, с заданными именами признаков

Расшифровка заданных имен и соответствующих им признаков представлена в таблице ниже:

Таблица 1 – Заданные имена признаков

| **Заданное имя признака в программе** | **Признак** |
|---|---|
| A1 | Sex |
| A2 | DocType |
| A3 | Business |
| A4 | Education |
| A5 | CarStatus |
| A6 | AddIncome |
| A7 | Income |
| A8 | CreditSum |
| A9 | Age |
| A10 | DayOff |
| GB | Good-Bad |

Выбрасываем характеристику DayOff или A10, так как она является поясняющей для целевой функции GB. Информация о будущей просрочке потенциального заемщика нам не будет известна заранее.

Выделим числовые и категориальные признаки. Фрагмент кода представлен ниже.

```
# Разделение признаков на категориальные и количественные
categorical_columns = [c for c in data.columns if
data[c].dtype.name == 'object']
numerical_columns = [c for c in data.columns if
data[c].dtype.name != 'object']
print('categorical and numerical columns respectively')
print(categorical_columns)
print(numerical_columns)
```

Результат приведен ниже. Соответственно перечислены категориальные и количественные признаки.

```
categorical and numerical columns respectively
['A1', 'A2', 'A3', 'A4', 'A5', 'GB']
['A6', 'A7', 'A8', 'A9']
```

Рисунок 6 – Числовых и категориальные признаки.

### 2.4.1 Исправление описок

Следует проанализировать категориальные признаки – перечислить уникальные значения для каждого из них, чтобы узнать есть ли описки в значениях категорий и исправить их в случае их присутствия в данных. Ниже представлен код замены значения описки 'Drivine licence'.

```
data['A2']=np.where(data['A2'] =='Drivine licence', 'Driving
licence', data['A2'])
```

```
Ordered categorical values
['Male' 'Female' nan]
['Driving licence' 'other' 'Foreign passport' 'Military card' nan]
['Industry' 'Education' 'Finance /Banking/Insurance' 'Retail' 'other'
 'IT and Telecoms']
['Secondary education' 'Higher education' 'Graduate']
['1' '0']
['1' '0']
```

Рисунок 7 – Уникальные значения категориальных признаков без описок.

Здесь nan означают пропущенные значения.

### 2.4.2 Удаление дубликатов

Следующим шагом после исправления описок в значениях датасета, будет удаление дубликатов строк данных. Для этого применим метод к набору данных метод drop_duplicates.

```
# Удаление дубликатов
print('Deleting duplicates')
data = data.drop_duplicates()
print(data.shape)
print('\r')
```

```
Deleting duplicates
(1500, 10)
```

Рисунок 8 – Размерность после удаления дубликатов.

Итогом удаления дубликатов стало сокращение количества строк на один.

### 2.4.3 Получение сводной информация о признаках

После этого можно проанализировать все признаки: количественные и категориальные. Для этого выведем сводную информацию отдельно для количественных и категориальных признаков. Для этого выполним следующие команды:

```
# Сводная информация о количественных признаках
print(data.describe())
print('\r')

# Сводная информация о категориальных признаках
```

```
print(data[categorical_columns].describe())
print('\r')
```

С помощью метода describe() получим сводную информацию для количественных признаков. По умолчанию данный метод выводит информацию только для количественных признаков. Это общее их количество (count), среднее значение (mean), стандартное отклонение (std), минимальное значение (min), медиана (50%), значения первой (25%) и третьей (75%) квартилей, максимальное значение (max) (Рисунок 9).

| | A6 | A7 | A8 | A9 |
|---|---|---|---|---|
| count | 1500.000000 | 1.500000e+03 | 1.499000e+03 | 1500.000000 |
| mean | 3044.348000 | 1.522305e+04 | 2.924419e+05 | 36.146667 |
| std | 17921.846118 | 8.960915e+04 | 4.249867e+05 | 26.480243 |
| min | 0.000000 | 0.000000e+00 | 1.930000e+02 | 19.000000 |
| 25% | 647.500000 | 3.238500e+03 | 1.500000e+05 | 28.000000 |
| 50% | 1325.000000 | 6.625500e+03 | 2.500000e+05 | 34.000000 |
| 75% | 2608.500000 | 1.304325e+04 | 4.000000e+05 | 42.000000 |
| max | 514074.000000 | 2.570370e+06 | 1.500000e+07 | 999.000000 |

Рисунок 9 – Сводная информация для количественных признаков

Также мы можем получить общую информацию по категориальным признакам (Рисунок 10).

| | A1 | A2 | A3 | A4 | A5 | GB |
|---|---|---|---|---|---|---|
| count | 1499 | 1481 | 1500 | 1500 | 1500 | 1500 |
| unique | 2 | 4 | 6 | 3 | 2 | 2 |
| top | Male | Driving licence | Industry | Secondary education | 0 | 0 |
| freq | 786 | 666 | 516 | 782 | 1011 | 1302 |

Рисунок 10 – Сводная информация по категориальным признакам

В таблице сводной информации для каждого категориального признака по порядку выводится число заполненных ячеек, количество уникальных значений, которые он принимает, наиболее часто встречающееся

значение признака и количество объектов, в которых встречается это наиболее частое значение.

### 2.4.4 Проверка на мультиколлинеарность

Чтобы проверить переменные на мультиколлинеарность, для каждой количественной переменной построим гистограмму, а для каждой пары количественных переменных построим диаграмму рассеяния. Это можно осуществить при помощи функции scatter_matrix. Все диаграммы представлены на рисунке 11.

```
scatter_matrix(data, alpha=0.7, figsize=(10, 10))
```

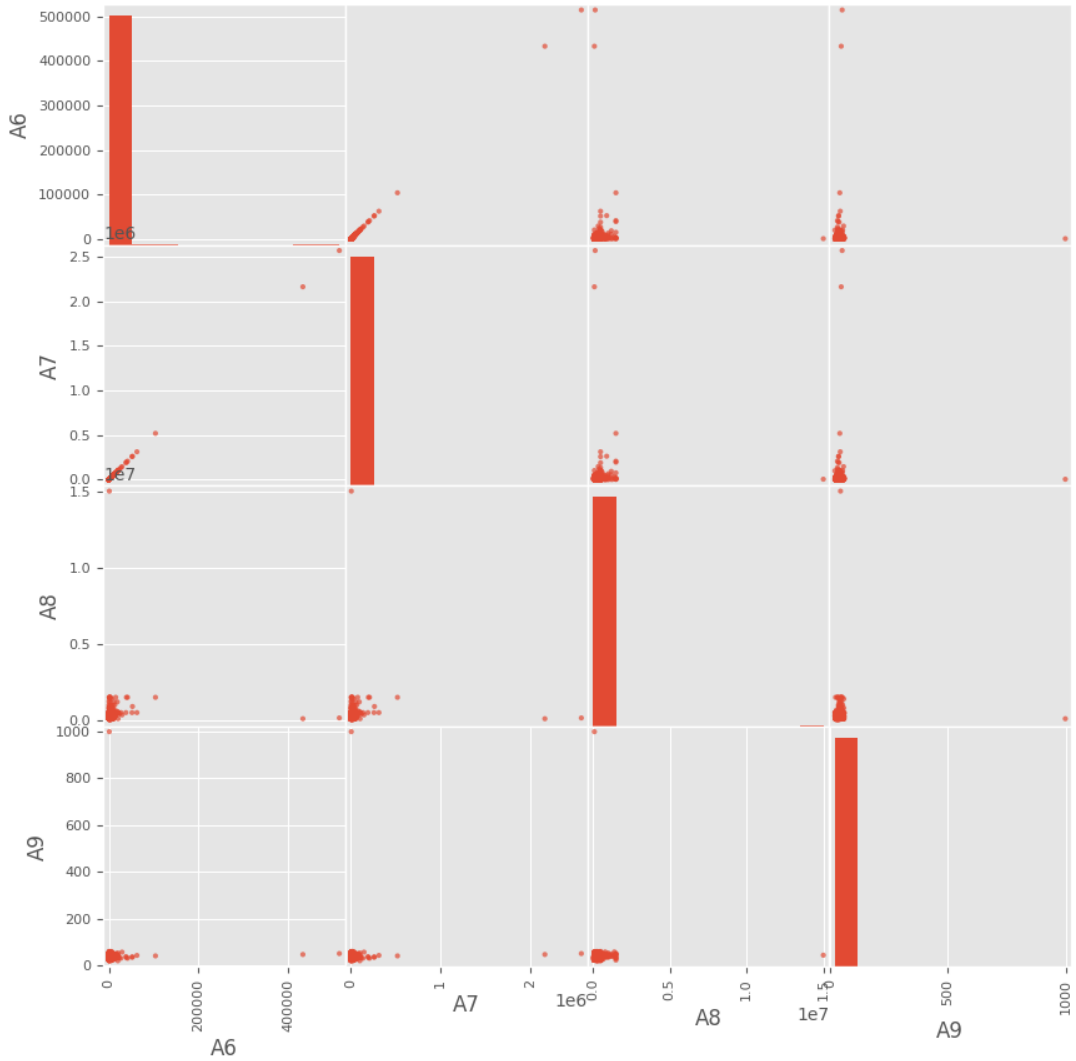



Рисунок 11 – Диаграммы рассеяния.

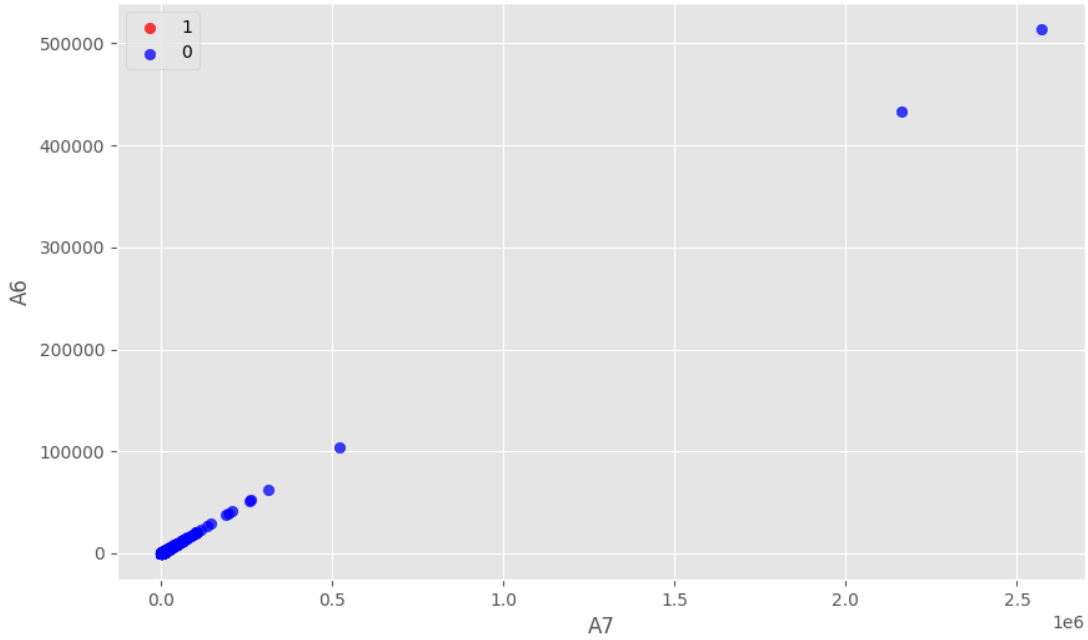



Рисунок 12 – Диаграмма рассеяния коррелирующих признаков.

Наличие мультиколленарности в исходных переменных (атрибутах) приводит к неточности прогнозной модели, и конечный результат будет сильно зависеть от разных выборок.

Поэтому трудно оценить влияние независимых переменных на целевую функцию.

Из построенных диаграмм видно, что между собой сильно коррелируют два признака: A6 – AddIncome (Дополнительный доход) и A7 – Income (Основной доход). Также построим корреляционную матрицу, с помощью которой также можно установить сильную корреляцию признаков (Рисунок 13). Для ее построения можно использовать данный метод:

```
# Корреляционная матрица
print(data.corr())
```

Корреляционная матрица атрибутов используется для оценки мультиколлинеарности. Если значение парных коэффициентов корреляции выше 0,7 - 0,8, это указывает на возможные проблемы с качеством будущей прогнозной модели. В таких случаях необходимо изменить исходные переменные, чтобы атрибуты не были так сильно коррелированы.

| | A6 | A7 | A8 | A9 |
|---|---|---|---|---|
| A6 | 1.000000 | 1.000000 | 0.022975 | 0.018179 |
| A7 | 1.000000 | 1.000000 | 0.022974 | 0.018179 |
| A8 | 0.022975 | 0.022974 | 1.000000 | 0.010302 |
| A9 | 0.018179 | 0.018179 | 0.010302 | 1.000000 |

Рисунок 13 – Корреляционная матрица

Построение данной матрицы также подтвердила корреляцию признаков дополнительного и основного доходов. Чтобы избавиться от этого вышеотмеченные переменные дохода будут объединены в один признак. Все остальные ее недиагональные значения по модулю не превосходят значение экспертной оценки в 0.25.

### 2.4.5 Избавление от коррелирующих признаков

Было выявлено, что между собой сильно коррелируют два признака: A6 – AddIncome (Дополнительный доход) и A7 – Income (Основной доход).

Для того чтобы атрибуты не были так сильно коррелированы, а также для того чтобы избавиться от потенциальных проблем с качеством будущей скоринговой модели, было принято решение изменить исходные переменные.

Вышеотмеченные переменные дополнительного и основного доходов будут объединены в один признак Income путем суммирования значений данных атрибутов для каждого заемщика.

```
# Объединение AddIncome(A6) и Income(A7)
data['A7'] = data['A7'] + data['A6']
data = data.drop(['A6'], axis=1)  # Выбрасываем столбец
'AddIncome'.
print(data['A7'])
numerical_columns = [c for c in data.columns if
data[c].dtype.name != 'object']
```

Результат объединения представлен ниже. Теперь признак доход(A7) содержит в себе сумму всех доходов заемщика, количественные признаки также больше не содержат атрибут дополнительного дохода, их количество сократилось на один.

```
0        19906
1         5474
2         5516
3        24000
4        73020
         ...
1496    175824
1497      9469
1498     51120
1499     11798
1500     25813
Name: A7, Length: 1500, dtype: int64
```

Рисунок 14 – Признак доход.

Также, была построена новая корреляционная матрица. Сильно коррелирующие признаки в ней отсутствуют.

| | A7 | A8 | A9 |
|---|---|---|---|
| A7 | 1.000000 | 0.022974 | 0.018179 |
| A8 | 0.022974 | 1.000000 | 0.010302 |
| A9 | 0.018179 | 0.010302 | 1.000000 |

Рисунок 15 – Корреляционная матрица.

### 2.4.6 Обработка выбросов

Выбросами считаются аномальные значения в данных, которые сильно выделяются из общей выборки. Поскольку алгоритм логистической регрессии, который будет применен при машинном обучении, чувствителен к выбросам, при подготовке данных их следует обработать.

Простейший способ определения выбросов в количественной переменной – считать выбросами наблюдения, не укладывающиеся в заданные квартили. Такой подход графически реализован в виде диаграмм размаха. Данная диаграмма показывает медиану, интерквартильный размах, максимальное и минимальное значение.

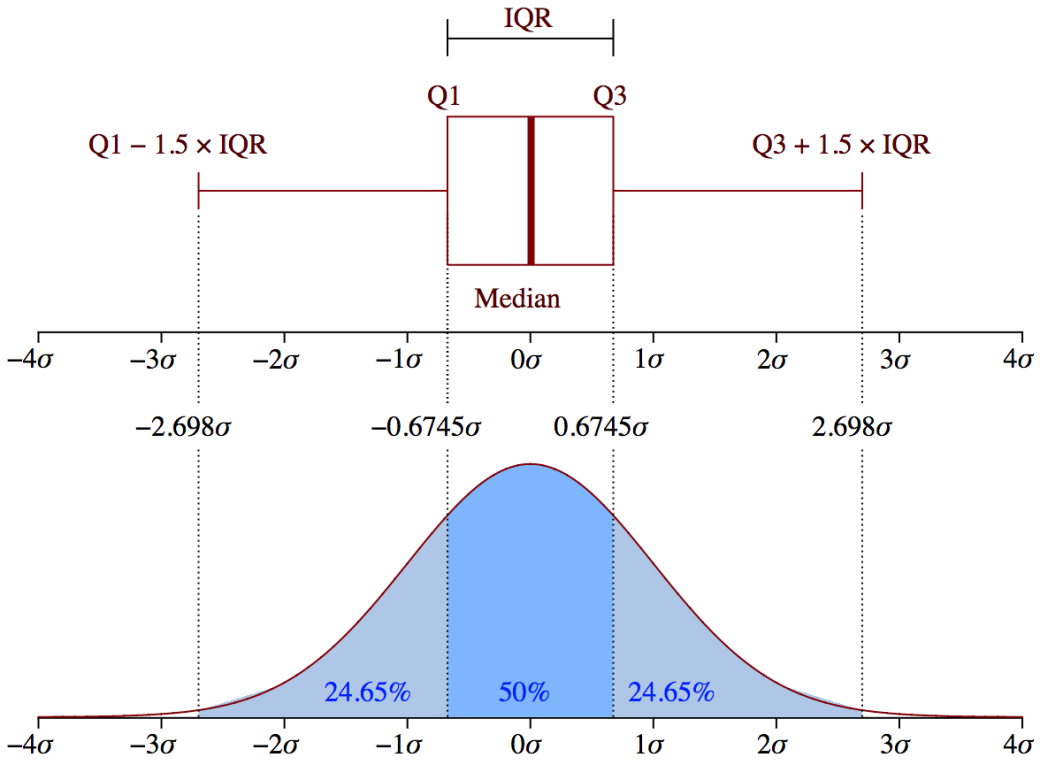



Рисунок 16 – Диаграмма размаха.

Интерквартильный размах (IQR) – это мера разброса данных, которая используется для определения выбросов. Это разница между третьей квартилью и первой квартилью (IQR = Q3 -Q1). значение является выбросом, если оно лежит за пределами отрезка от первой квартили минус 1,5 интерквартильного размаха до третьей квартили плюс 1,5 интерквартильного размаха. Если значение лежит внутри этого интервала — оно нормальное, если вне него — это выброс.

Для визуального обнаружения выбросов в текущем датасете, для каждого количественного признака была построена точечная диаграмма.

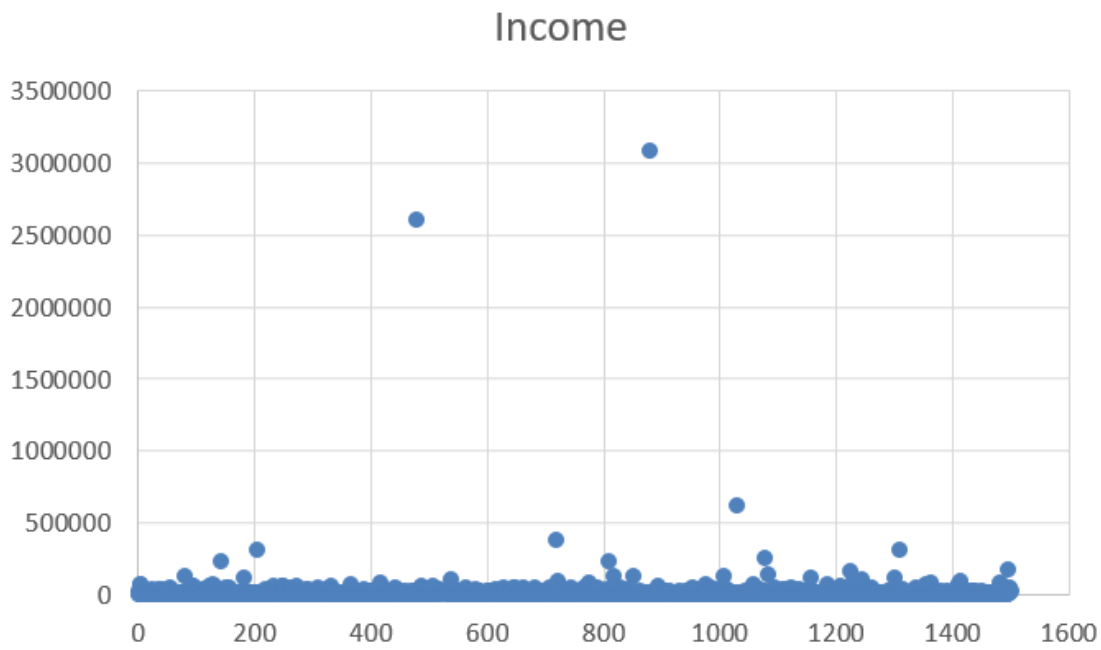



Рисунок 17 – Выбросы для дохода.

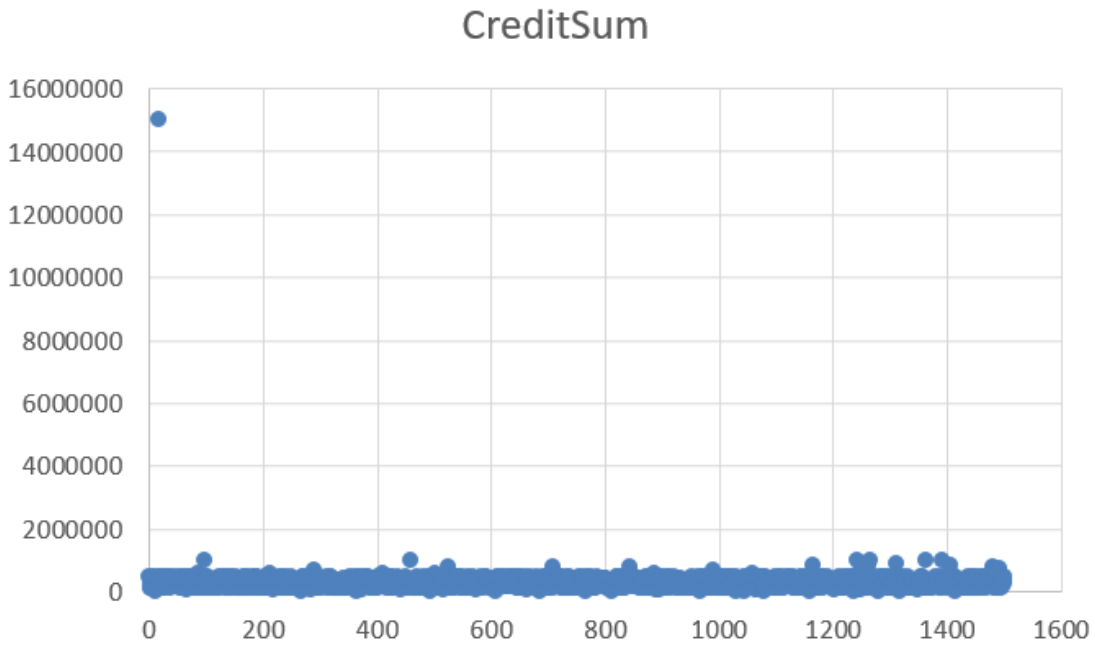



Рисунок 18 – Выбросы для сумм кредита.

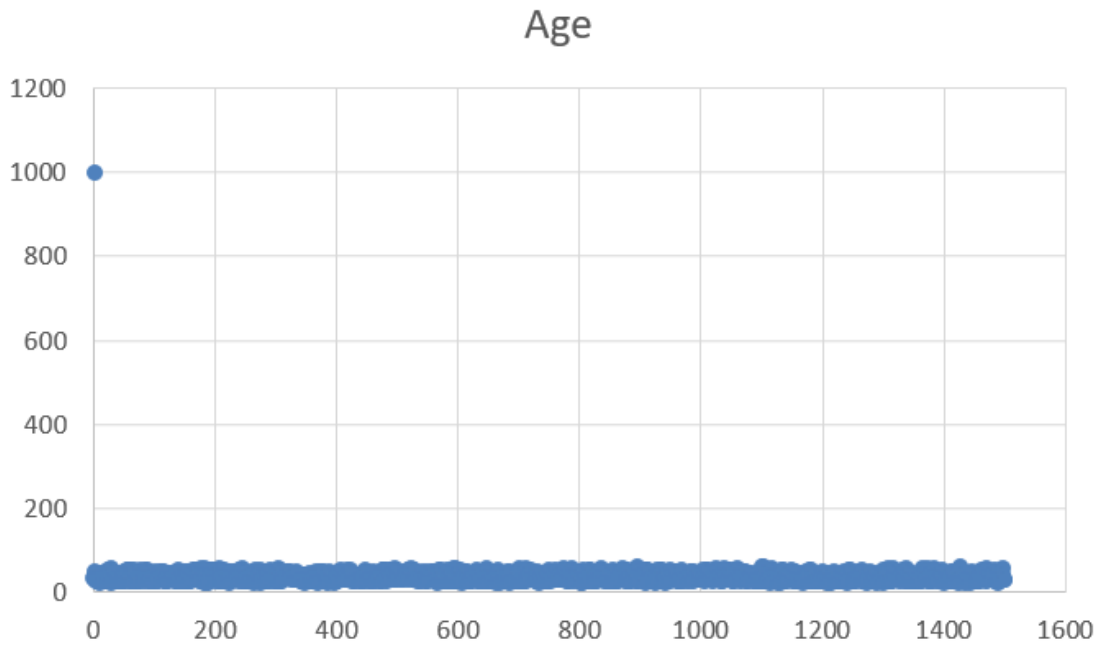



Рисунок 19 – Выбросы для возраста.

Небольшое количество выбросов обычно удаляется или заменяется средним или медианным значением.

Ниже приведен метод для определения выбросов среди количественных переменных и их замену на медианное значение для наименьших потерь данных:

```
# Удаление выбросов
def outlier_detect(df):
    for i in df.describe().columns:
        Q1=df.describe().at['25%',i]
        Q3=df.describe().at['75%',i]
        IQR=Q3 - Q1
        LTV=Q1 - 1.5 * IQR
        UTV=Q3 + 1.5 * IQR
        x=np.array(df[i])
        p=[]
        for j in x:
            if j < LTV or j>UTV:
                p.append(df[i].median())
            else:
                p.append(j)
        df[i]=p
    return df
outlier_detect(data)
print(data)
```

### 2.4.7 Обработка пропусков

Для того чтобы узнать количество заполненных элементов можно воспользоваться методом count.

```
print(data.count(axis=0))
```

В данном методе параметр axis, равный 0, указывает, что мы двигаемся по размерности 0, то есть сверху вниз, другими словами мы хотим вывести количество непропущенных элементов в каждом столбце, а не строке. Данные представлены на рисунке 20.

```
Amount of fillings:
A1    1499
A2    1481
A3    1500
A4    1500
A5    1500
A7    1500
A8    1499
A9    1500
GB    1500
dtype: int64
```

Рисунок 20 – Общее количество заполненных ячеек

Из полученной информации видно, что данные содержат пропущенные значения. Данную проблему можно решить следующими способами:

- удалить столбцы с пропусками
- удалить строки с пропусками

Однако, применив один из таких способов решения, есть риск лишиться ценных данных, которые в будущем могли бы сильно повлиять на итоговую точность предсказательной модели. По этой причине рассмотрим альтернативные пути решения проблемы.

Как видно из рисунка 20, пропуски есть в колонках A1, A2 и A8, что соответствует признакам пола заемщика, типа документа, который он предоставил и сумме запрашиваемого кредита. Таким образом пропущенные значения есть и среди количественных, и среди категориальных признаков.

Заполнить пропущенные значения количественных признаков можно с помощью метода библиотеки Pandas fillna (заполняет значения NA/NaN используя заданные методы). Мною было принято решение использовать медианные значения. Однако, можно заменить пропуск на средний доход по всей выборке. Кажется, что медиана – это более хороший вариант, поскольку он устойчив к выбросам.

```
# Замена пропусков медианой для количественных признаков
data = data.fillna(data.median(axis=0), axis=0)
print(data.describe())
print('\r')
print(data.count(axis=0))
print('\r')
```

| | A7 | A8 | A9 |
|---|---|---|---|
| count | 1500.000000 | 1500.000000 | 1500.000000 |
| mean | 8900.438333 | 264913.621333 | 35.503333 |
| std | 7200.258994 | 138927.276414 | 9.073158 |
| min | 0.000000 | 193.000000 | 19.000000 |
| 25% | 3886.000000 | 150000.000000 | 28.000000 |
| 50% | 7947.750000 | 250000.000000 | 34.000000 |
| 75% | 11776.500000 | 350000.000000 | 42.000000 |
| max | 33223.000000 | 750000.000000 | 59.000000 |

Рисунок 21 – Измененная сводная информация После замены пропусков.

```
A1    1499
A2    1481
A3    1500
A4    1500
A5    1500
A7    1500
A8    1500
A9    1500
GB    1500
dtype: int64
```

Рисунок 22 – Общее количество заполненных ячеек после замены.

Далее рассмотрим пропущенные значения в столбцах, соответствующих категориальным признакам. Стратегией заполнения пропущенных значений в моей работе стала замена пропуска самым популярным в столбце значением. Автоматизированный процесс замены значений и вывода сводной информации выглядит следующим образом:

```
# Замена пропусков наиболее часто встречающимся значением
# для категориальных признаков
data_describe = data.describe(include=[object])
print(data_describe)
for c in categorical_columns:
    data[c] = data[c].fillna(data_describe[c]['top'])

print('\r')
print(data.describe(include=[object]))
print('\r')
print(data.count(axis=0))
print('\r')
```

На рисунке ниже представлена сводная информация о категориальных признаках до и после замены пропущенных значений.

| | A1 | A2 | A3 | A4 | A5 | GB |
|---|---|---|---|---|---|---|
| count | 1499 | 1481 | 1500 | 1500 | 1500 | 1500 |
| unique | 2 | 4 | 6 | 3 | 2 | 2 |
| top | Male | Driving licence | Industry | Secondary education | 0 | 0 |
| freq | 786 | 666 | 516 | 782 | 1011 | 1302 |

| | A1 | A2 | A3 | A4 | A5 | GB |
|---|---|---|---|---|---|---|
| count | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| unique | 2 | 4 | 6 | 3 | 2 | 2 |
| top | Male | Driving licence | Industry | Secondary education | 0 | 0 |
| freq | 787 | 685 | 516 | 782 | 1011 | 1302 |

Рисунок 23 – Сводная информация До и После замены пропусков.

Также еще раз выведем общее количество заполненных ячеек, чтобы убедиться, что на текущий момент в данных пропуски отсутствуют.

```
A1    1500
A2    1500
A3    1500
A4    1500
A5    1500
A7    1500
A8    1500
A9    1500
GB    1500
dtype: int64
```

Рисунок 24 - Общее количество заполненных ячеек после замены.

### 2.4.8 Категоризация количественных признаков

В скоринговой системе в качестве независимых переменных могут быть использованы и категориальные, и количественные предикторы. Однако большинство разработчиков скоринговых систем всегда категорируют количественные переменные.

Исторически сложилось, что чаще для построения скоринговых карт используют категориальные предикторы. Категоризация количественных

переменных позволяет в дальнейшем упростить интерпретацию скоринговой карты. Ниже приведен пример категоризации возраста заемщика:

```
bins9 = [18, 24, 27, 29, 32, 34, 38, 41, 44, 49, 100]
labels9 = [str(i) for i in range(1, 11)]
data['A9_cat'] = pd.cut(data['A9'], bins=bins9, labels=labels9)
print(data['A9_cat'].describe())
print(data.columns)
```

После категоризации всех количественных переменных, стартовые количественные переменные удаляются из набора данных.

### 2.4.9 Векторизация

Используемая для машинного обучения библиотека Scikit-learn не может напрямую обрабатывать категориальные признаки. Поэтому категориальные признаки следует преобразовать в количественные, перед тем, как подавать на вход алгоритмов машинного обучения данные о заемщиках.

Категориальные признаки, которые принимают два значения (бинарные признаки) и те признаки, которые принимают более двух значений, в том числе бывшие количественные признаки, будем обрабатывать по-разному.

Сначала выделим категориальные-количественные, бинарные и не бинарные признаки следующим набором команд:

```
# Выделим бывшие количественные, категориальные бинарные и
категориальные не бинарные признаки
num_cat_columns = [c for c in data.columns if
data[c].dtypes.name == 'category']
binary_columns = [c for c in categorical_columns if
data_describe[c]['unique'] == 2]
nonbinary_columns = [c for c in categorical_columns if
data_describe[c]['unique'] > 2]
print('numerical, binary and non-binary columns respectively')
print(num_cat_columns)
print(binary_columns)
print(nonbinary_columns)
print('\r')
```

Результат работы кода представлен ниже на рисунке 25.

```
numerical, binary and non-binary columns respectively
['A7_cat', 'A8_cat', 'A9_cat']
['A1', 'A5', 'GB']
['A2', 'A3', 'A4']
```

Рисунок 25 – Бинарные и не бинарные признаки.

В результате признаки были поделены на численные категориальные, категориальные бинарные и не бинарные.

['A7_cat', 'A8_cat', 'A9_cat] – численные категориальные признаки

['A1', 'A5', 'GB'] - бинарные признаки

['A2', 'A3', 'A4'] - не бинарные признаки

Сначала рассмотрим бинарные признаки. Для них всех заменим значения на 0 и 1. В данной задаче достаточно было заменить только значения атрибута A1, однако был написан общий алгоритм для замены значений любых бинарных признаков и вывода сводной информации по ним.

```
for c in binary_columns:
    top = data_describe[c]['top']
    top_items = data[c] == top
    data.loc[top_items, c] = 0
   data.loc[np.logical_not(top_items), c] = 1
```

| | A1 | A5 | GB |
|---|---|---|---|
| count | 1500 | 1500 | 1500 |
| unique | 2 | 2 | 2 |
| top | 0 | 0 | 0 |
| freq | 787 | 1011 | 1302 |

Рисунок 26 – Сводная информация по бинарным признакам.

К остальным признакам будет применен метод векторизации. Данный метод заключается в том, что признак принимающий s значений, заменяется

на s бинарных признаков, которые принимают значения либо 0, либо 1, в зависимости от того, какое значение принимал исходный признак.

В качестве примера возьмем признак A2. Он принимает четыре уникальных значения: 'Driving licence', 'Foreign passport', 'Military card', 'other'.

Следуя методы векторизации, заменим признак A2 на четыре бинарных признака: A2_Driving licence, A2_Foreign passport, A2_Military card, A2_other.

- Если признак A2 принимает значение 'Driving licence', то признак A2_Driving licence равен 1, A2_Foreign passport равен 0, A2_Military card равен 0, A2_other равен 0.
- Если признак A2 принимает значение 'Foreign passport', то признак A2_Driving licence равен 0, A2_Foreign passport равен 1, A2_Military card равен 0, A2_other равен 0.
- Если признак A2 принимает значение 'Military card', то признак A2_Driving licence равен 0, A2_Foreign passport равен 0, A2_Military card равен 1, A2_other равен 0.
- Если признак A2 принимает значение 'other', то признак A2_Driving licence равен 0, A2_Foreign passport равен 0, A2_Military card равен 0, A2_other равен 1.

Для осуществления векторизацию в библиотеке pandas существует метод get_dummies:

```
# Осуществление векторизации
data_nonbinary = pd.get_dummies(data[nonbinary_columns])
print(data_nonbinary.columns)
print('\r')

data_cat_columns = pd.get_dummies(data[num_cat_columns])
print(data_cat_columns.columns)
print('\r')
```

Результат работы метода и вывод всех новых признаков представлен на рисунке 27.

```
Index(['A2_Driving licence', 'A2_Foreign passport', 'A2_Military card',
       'A2_other', 'A3_Education', 'A3_Finance /Banking/Insurance',
       'A3_IT and Telecoms', 'A3_Industry', 'A3_Retail', 'A3_other',
       'A4_Graduate', 'A4_Higher education', 'A4_Secondary education'],
      dtype='object')

Index(['A7_cat_1', 'A7_cat_2', 'A7_cat_3', 'A7_cat_4', 'A7_cat_5', 'A7_cat_6',
       'A7_cat_7', 'A7_cat_8', 'A7_cat_9', 'A7_cat_10', 'A8_cat_1', 'A8_cat_2',
       'A8_cat_3', 'A8_cat_4', 'A8_cat_5', 'A8_cat_6', 'A8_cat_7', 'A8_cat_8',
       'A8_cat_9', 'A9_cat_1', 'A9_cat_2', 'A9_cat_3', 'A9_cat_4', 'A9_cat_5',
       'A9_cat_6', 'A9_cat_7', 'A9_cat_8', 'A9_cat_9', 'A9_cat_10'],
      dtype='object')
```

Рисунок 27 – Результат векторизации.

Для последующего применения логистической регрессии, соединим всё в одну таблицу. Однако далее столбцы, соответствующие входным признакам (матрица X), и выделенный признак целевой функции GB (вектор y) будем рассматривать отдельно.

```
data = pd.concat((data_numerical, data[binary_columns],
data_nonbinary), axis=1)
data = pd.DataFrame(data, dtype=float)
print(data.shape)
print('\r')

X = data.drop(['GB'], axis=1)  # Выбрасываем столбец 'GB'.
y = data['GB']
feature_names = X.columns
print(feature_names)
print('\r')
print(X.shape)
print(y.shape)
N, d = X.shape
```

Полученный список признаков, а также размер общей матрицы data и размеры матрицы X и вектора y представлены ниже на рисунке 28.

```
(1500, 45)

Index(['A1', 'A5', 'A2_Driving licence', 'A2_Foreign passport',
       'A2_Military card', 'A2_other', 'A3_Education',
       'A3_Finance /Banking/Insurance', 'A3_IT and Telecoms', 'A3_Industry',
       'A3_Retail', 'A3_other', 'A4_Graduate', 'A4_Higher education',
       'A4_Secondary education', 'A7_cat_1', 'A7_cat_2', 'A7_cat_3',
       'A7_cat_4', 'A7_cat_5', 'A7_cat_6', 'A7_cat_7', 'A7_cat_8', 'A7_cat_9',
       'A7_cat_10', 'A8_cat_1', 'A8_cat_2', 'A8_cat_3', 'A8_cat_4', 'A8_cat_5',
       'A8_cat_6', 'A8_cat_7', 'A8_cat_8', 'A8_cat_9', 'A9_cat_1', 'A9_cat_2',
       'A9_cat_3', 'A9_cat_4', 'A9_cat_5', 'A9_cat_6', 'A9_cat_7', 'A9_cat_8',
       'A9_cat_9', 'A9_cat_10'],
      dtype='object')

(1500, 44)
(1500,)
```

Рисунок 28 – Объединенные признаки, размеры массива До и массивов После.

Результатом раздела стало формирование «чистого» набора данных для дальнейшего прогнозного анализа. Для достижения результаты были предприняты следующие действия:

- Определены и устранены выбросы в данных;
- Восстановлены отсутствующие данные;
- Удалены дубликаты строк;
- Удалены описки в данных;
- Проведена проверка на мультиколлинеарность;
- Проведена категоризация количественных признаков
- Категориальные данные закодированы

Теперь после корректного проведения подготовки данных стало возможным дальнейшее проведение регрессионного анализа.

# ГЛАВА 3 РЕАЛИЗАЦИЯ СКОРИНГОВОЙ КАРТЫ

## 3.1 Обучающая и тестовая выборки

Для проверки адекватности и точности предсказания скоринговой модели на этапе ее разработки историческую выборку необходимо разделить на две группы обучающую и тестовую выборки. Обучающая выборка представляет собой наблюдения, по которым будет строится модель. Тестовая (контрольная) выборка – это наблюдения по которым значение зависимой переменной будет известно, однако данные наблюдения не будут участвовать в построении модели, а будут использоваться только для проверки ее предсказательной точности.

Обучающая и тестовая выборки формируются случайным образом и обычно в соотношении 70–80% и 30–20% соответственно от исходного объема исторической выборки. В данной работе было принято решение разбить исходные данные на обучающую и тестовую выборки в соотношение 70% на 30%.

Тестовая выборка используется после построения модели логистической регрессии для проверки ее достоверности. Для кредитного скоринга — это прежде всего способность модели отличать «хороших» заемщиков от «плохих». Проверка достоверности модели заключается в ее применении и сравнении результатов на контрольной и тестовой выборке.

Модель должна давать корректные прогнозы не только на обучающей совокупности, но и на практике при ее применении. Схожие показатели точности, которые были получены на обучающей и тестовой выборке означают, что при практическом применении скоринговая модель будет работать приблизительно также.

В данной работе целевой функцией будет выбран параметр «GB».

### 3.2 Логистическая регрессия. Алгоритм машинного обучения

В библиотеке Scikit-learn реализовано большое количество алгоритмов машинного обучения. Однако для своей работы я решила использовать именно алгоритм логистической регрессии.

Логистическая регрессия - это статистический метод прогнозирования бинарных классов. Результат или целевая переменная имеют двоичный характер (1/0, Да / Нет, Истина / Ложь). Он вычисляет вероятность возникновения события.

Логистическая регрессия представляет собой особый случай линейной регрессии, при котором целевая переменная носит категориальный характер. Логистическая регрессия предсказывает вероятность возникновения бинарного события, используя сигмовидную функцию (также называемую логистической функцией).

Уравнение линейной регрессии:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

Где y - зависимая переменная, а $X_1, X_2 \ldots X_n$ - объясняющие переменные.

Сигмовидная функция:

$$P(y = 1) = 1/(1 + e^{-y}) \quad (2)$$

Применение сигмоиды к линейной регрессии:

$$P(y = 1) = 1/(1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}) \quad (3)$$

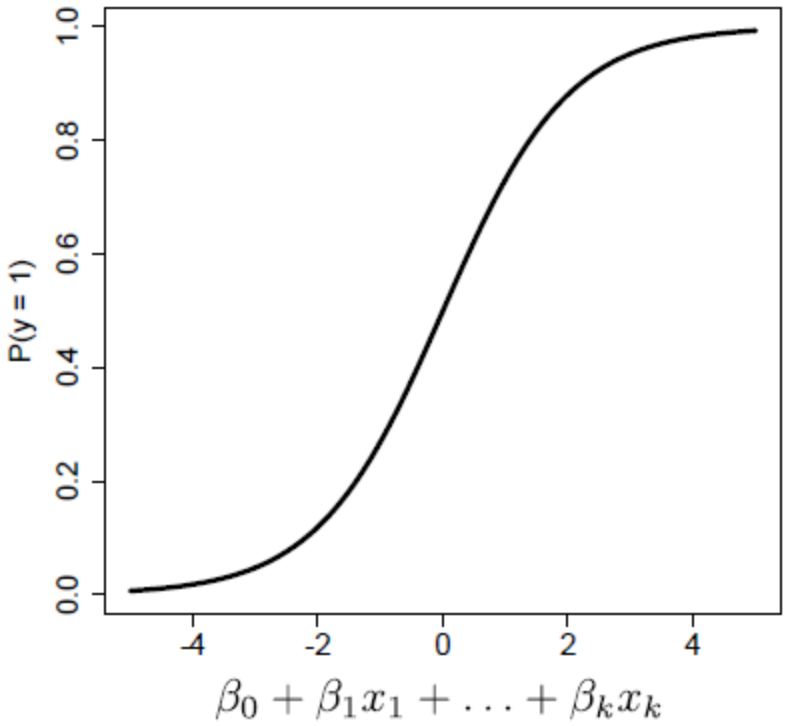



Рисунок 29 – Логистическая функция.

Функция может принимать любое действительное число и отображать его в значение между 0 и 1. Если выходные данные сигмоидальной функции больше 0,5, мы можем классифицировать результат как положительный, и напротив, если он меньше 0,5, он будет классифицирован как отрицательный.

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

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n, \quad (4)$$

где $p$ — вероятность наступления дефолта по кредиту для заемщика; $X_j$ — значение j-ой независимой переменной; $\beta_0$ — независимая константа модели, $\beta_j$ — параметры модели.

Представленное уравнение демонстрирует зависимость вероятности наступления просрочки по кредиту от значений независимых переменных. Константа в модели отражает естественный уровень риска наступления моделируемого события при равенстве всех независимых переменных нулю. Значения коэффициентов при независимых переменных, отражающих степень их влияния на шанс дефолта в логарифмической шкале, используются для построения скоринговой карты.

### 3.3 Обучение

Для начала произведем разбиение на тестовую и обучающую выборку. Как уже было отмечено ранее, выборка будет осуществлена в пропорциях 70 на 30 (1050 и 450 записей). Для этого будет применена встроенная функция train_test_split.

В качестве алгоритма машинного обучения будет использован алгоритм логистической регрессии (sklearn.linear_model.LogisticRegression).

Когда модель обучена, становится возможным предсказание для новых объектов значения целевого признака по входным. Для этого воспользуемся методом predict. Чтобы получить матрицу с предсказанными бинарными данными, выведем предсказнные значения бинарного признака добросовестности заемщика.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=1)
N_train, _ = X_train.shape
N_test,  _ = X_test.shape
print(N_train, N_test)

lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
print(y_pred)
```

```
1050 450
[0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
 0. 0. 1. 1. 0. 0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.
 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
 1. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 1. 0.
 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0.
 1. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0.
 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0.
 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 1. 0.
 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0.
 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0.
 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 0. 1. 0. 1.
 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
```

Рисунок 30 – Размеры выборок и матрица предсказанных значений.

Нас интересует качество построенной модели на обучающей и (что более важно) тестовой выборках:

```
err_train = np.mean(y_train != lr.predict(X_train))
err_test = np.mean(y_test != lr.predict(X_test))
print(err_train, err_test)
err_test_perc = round(err_test, 3)*100
print('Train-test mistake is: ' + str(err_test_perc) + '%')
logmodel_score = lr.score(X_test, y_test)
print('\nThis is how Model Scored: ', logmodel_score)
```

```
0.0 0.0
Train-test mistake is: 0.0%

This is how Model Scored:  1.0
```

Рисунок 31 – Качество модели.

Переменные err_train и err_test – это ошибки на обучающей и тестовой выборках. Как мы видим, и та, и другая ошибки составили 0.0%.

Более важной является ошибка на тестовой выборке, так как мы должны уметь предсказывать наиболее точное, в меру возможностей, значение для новых заемщиков, которые при обучении были недоступны.

Для того чтобы рассчитать коэффициенты уравнения регрессии воспользуемся встроенным методом с использованием метода модели обучения LogisticRegression – coef_.

```
# Расчет коэффициентов с использованием метода coef_
column_label = list(X_train.columns)
lr_Coeff = pd.DataFrame(lr.coef_, columns = column_label)
lr_Coeff['intercept'] = lr.intercept_
print("Coefficient Values Of The Surface Are:\r ", lr_Coeff)
```

```
        A1        A5  A2_Driving licence  A2_Foreign passport  \
-0.406278 -0.829098           -0.045522             0.094774

 A2_Military card  A2_other  A3_Education  A3_Finance /Banking/Insurance  \
         -0.58971  0.540262     -0.511537                      -0.026487

 A3_IT and Telecoms  A3_Industry  A3_Retail  A3_other  A4_Graduate  \
           0.260423     0.336665   0.093437 -0.152696     -0.12753

 A4_Higher education  A4_Secondary education  A7_cat_1  A7_cat_2  A7_cat_3  \
            -0.20466                0.331995  0.425543  0.306144  0.426842

 A7_cat_4  A7_cat_5  A7_cat_6  A7_cat_7  A7_cat_8  A7_cat_9  A7_cat_10  \
 0.033302  0.630499 -0.236585  0.423421 -0.511594 -0.993938  -0.503828

 A8_cat_1  A8_cat_2  A8_cat_3  A8_cat_4  A8_cat_5  A8_cat_6  A8_cat_7  \
-0.303967 -0.633159 -0.099936  0.239665 -0.023341  0.119718  0.179317

 A8_cat_8  A8_cat_9  A9_cat_1  A9_cat_2  A9_cat_3  A9_cat_4  A9_cat_5  \
 0.382577  0.138932  0.542806  0.471707  0.435871  0.521296   0.63465

 A9_cat_6  A9_cat_7  A9_cat_8  A9_cat_9  A9_cat_10  intercept
 0.388309  0.086784 -0.755388 -0.828716  -1.497513  -2.066028
```

Рисунок 32 – Весовые коэффициенты признаков.

### 3.4 Перевод весовых коэффициентов в баллы скоринговой карты

Заключительным этапом разработки скоринговой модели выступает перевод коэффициентов логистической регрессии в скоринговые баллы. Если взять оценки коэффициентов логистической регрессии и умножить их на

значения независимых переменных, получиться итоговый скоринговый балл в шкале натуральных логарифмов:

$$\text{итоговый балл} = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k, \quad (5)$$

где $x_j$ — значение независимых преременных для оцениваемого заемщика, $b_j$ — оценки коэффициентов логистической регрессии.

Чтобы привести скоринговые баллы в линейную шкалу используют прием масштабирования. Оно позволяет избежать изменения прогностической способности скоринговой карты, лишь переводя скоринговые баллы в новую удобную для использования шкалу.

Скоринговый балл в линейной шкале является отношением шансов «хороших» заемщиков к «плохим». Для масштабирования следует задать диапазон числовой шкалы с минимум по максимум. Для разрабатываемой в данной работе скоринговой карты был взят диапазон от 0 до 1000.

Для приведения коэффициента логистической регрессии в скоринговый балл в линейной шкале применяют следующее преобразование:

$$\text{балл} = A \pm R \cdot b_j, \quad (6)$$

где R — множитель; A — смещение.

Для расчета баллов скоринговой карты множитель в равен 117,34, а смещение равно 907,23. Они были вычислены благодаря заданному диапазону значений и разнице коэффициентов. Это является одним из общепринятых стандартов расчета скоринговых баллов.

В таблице 2 приведены рассчитанные скоринговые баллы в линейной шкале. Для получения общего скорингового балла по заемщику, необходимо сложить баллы, относящихся к нему, значений каждой характеристики. Затем можно будет определить скоринговый балл с учетом смещения.

Таблица 2 – Характеристики в разработанной скоринг-системе.

| Характеристика | Значение | Оценка коэффициента лог.регрессии | Скоринговый балл в линейной шкале (знак сохраняется) |
|---|---|---|---|
| Sex (Пол) | Male | 0 | 0 |
| | Female | -0.406278 | 48 |
| DocType (Тип предоставленного документа) | Driving licence | -0.045522 | 5 |
| | Foreign passport | 0.094774 | 11 |
| | Military card | -0.58971 | 69 |
| | other | 0.540262 | 63 |
| Business (Сфера деятельности) | Education | -0.511537 | 60 |
| | Finance /Banking/Insurance | -0.026487 | 3 |
| | IT and Telecoms | 0.260423 | 31 |
| | Industry | 0.336665 | 40 |
| | Retail | 0.093437 | 11 |
| | other | -0.152696 | 18 |
| Education (Образование) | Graduate | -0.12753 | 15 |
| | Higher education | -0.20466 | 24 |
| | Secondary education | 0.331995 | 39 |
| CarStatus (Есть ли машина) | No | 0 | 0 |
| | Yes | -0.829098 | 97 |
| Income (Сумма всех доходов) | До 850 | 0.425543 | 50 |
| | От 850 до 2900 | 0.306144 | 36 |

| | | | |
|---|---|---|---|
| | От 2900 до 3500 | 0.426842 | 50 |
| | От 3500 до 4000 | 0.033302 | 4 |
| | От 4000 до 4500 | 0.630499 | 74 |
| | От 4500 до 5500 | -0.236585 | 28 |
| | От 5500 до 7500 | 0.423421 | 50 |
| | От 7500 до 10000 | -0.511594 | 60 |
| | От 10000 до 13000 | -0.993938 | 117 |
| | От 13000 | -0.503828 | 59 |
| CreditSum (Сумма кредита) | До 100000 | -0.303967 | 36 |
| | От 100000 до 130000 | -0.633159 | 74 |
| | От 130000 до 150000 | -0.099936 | 12 |
| | От 150000 до 200000 | 0.239665 | 28 |
| | От 200000 до 250000 | -0.023341 | 3 |
| | От 250000 до 300000 | 0.119718 | 14 |
| | От 300000 до 400000 | 0.179317 | 21 |
| | От 400000 до 500000 | 0.382577 | 45 |
| | От 500000 до 750000 | 0.138932 | 16 |
| Age (Возраст) | От 18 до 24 | 0.542806 | 64 |
| | От 24 до 27 | 0.471707 | 55 |

| | | | |
|---|---|---|---|
| | От 27 до 29 | 0.435871 | 51 |
| | От 29 до 32 | 0.521296 | 61 |
| | От 32 до 34 | 0.63465 | 74 |
| | От 34 до 38 | 0.388309 | 46 |
| | От 38 до 41 | 0.086784 | 10 |
| | От 41 до 44 | -0.755388 | 89 |
| | От 44 до 49 | -0.828716 | 97 |
| | От 49 | -1.497513 | 176 |
| Независимая константа модели | | -2.066028 | 242 |

Для построенной скоринговой карты, был посчитан минимальный общий балл, при котором заемщику можно выдавать кредит.

При общем скоринговом балле, равном 907,23 или более (т.е. выше смещения), заемщик будет считаться «хорошим». Напротив, с результатом ниже данного общего балла, потенциальный заемщик скорее всего окажется «плохим», так как для него высок риск не справиться со своими кредитными обязательствами, кредитной организации следует отказать ему в выдаче кредита.

### 3.5 Проверка результатов обучения. Матрицы путаницы

Проверенным методом поведения итогов работы алгоритма классификации является построение матрицы путаницы.

Только лишь точность классификации может ввести в заблуждение, если у вас разное количество наблюдений в каждом из классов или если классов в наборе данных более двух.

Матрица путаницы может дать лучшее представление того, что классификационная модель делает правильно и какие ошибки она допускает.

Проверка модели также будет проводится встроенными функциями библиотек «Python». Запускаем проверку на обучающей выборке:

```
y_pred = lr.predict(X_test)
cnf_matrix = metrics.confusion_matrix(y_train, y_pred)
classes = ["positive", "negative"]

df_cfm = pd.DataFrame(cnf_matrix, index = classes, columns =
classes)
plt.figure(figsize = (10,7))
cfm_plot = sn.heatmap(df_cfm, annot=True)
cfm_plot.figure.savefig("cfm.png")
```

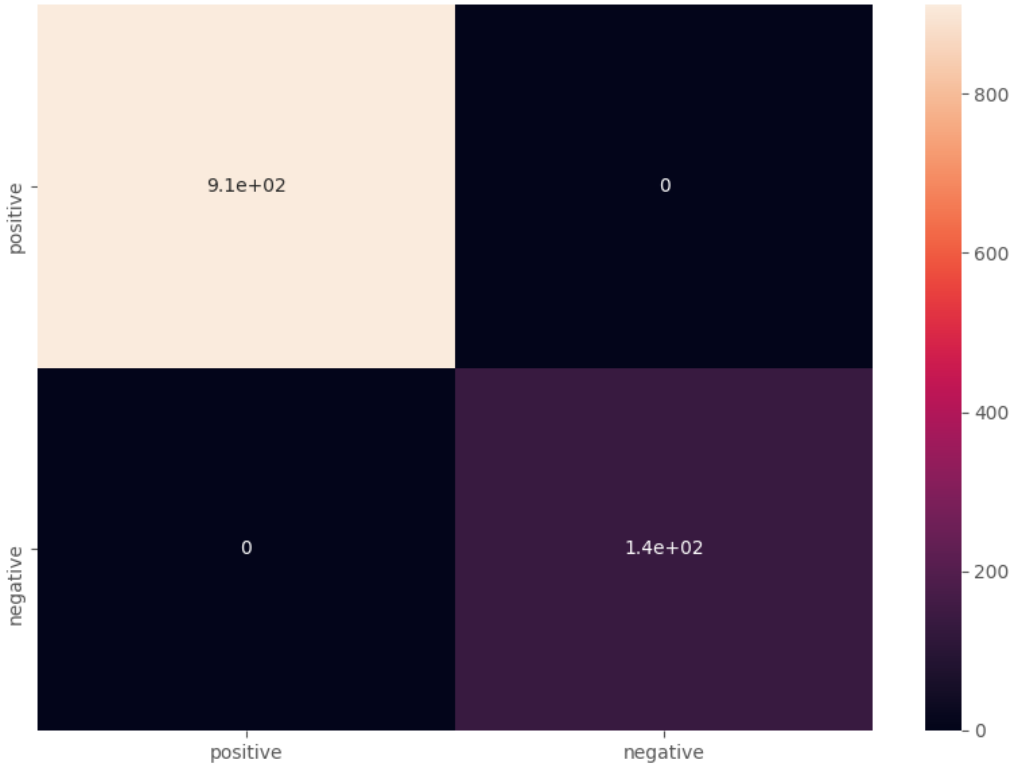



Рисунок 33 – Матрица путаницы (обучающиеся данные).

На обучающей выборке модель выдала результат с погрешностью в 0%, как и было получено ранее.

На построенной матрице путаницы наглядно видно, что для 910 людей, с отсутствием задолженности по кредиту более 90 дней, модель бы выдала все 910 положительных результатов, а из 140 заемщиков с задолженностью, модель выдала бы 140 отрицательных результатов.

Данный результат свидетельствует о том, что модель хорошо обучилась. Проверим правильность результатов на тестовой выборке.

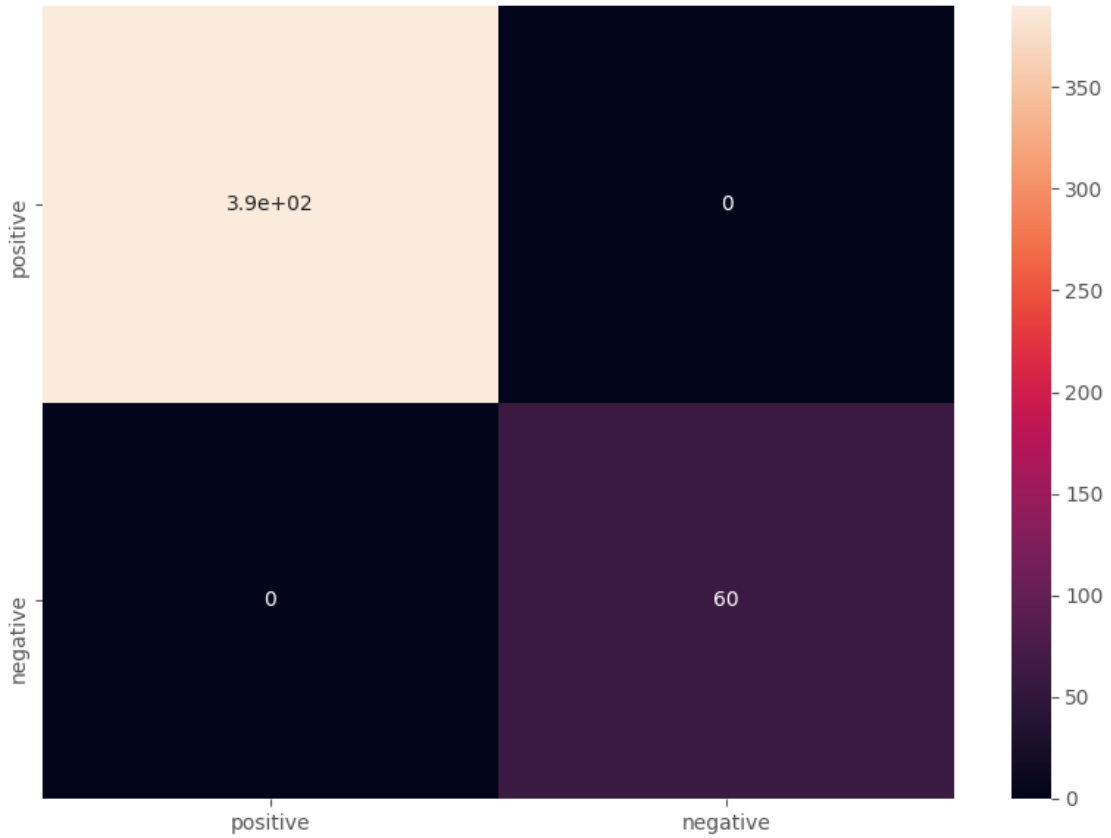



Рисунок 34 – Матрица путаницы (тестовые данные).

На тестовой выборке модель также показала хороший результат, из 450 тестов все вероятности были предсказаны верно. В данном случае кредит был бы выдан 390 людям, а 60 получили бы отказ.

# ГЛАВА 4 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

**Введение**

Целью данного раздела является определение оценки коммерческого потенциала, перспективности и альтернатив проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения, а также планирование и формирование бюджета научных исследований, определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования. Работа заключалась в разработке скоринговой модели для кредитных организаций.

## 4.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

### 4.1.1 Потенциальные потребители результатов исследования

Целевой аудиторией могут являться:

1) эмиссионные банки

2) неэмиссионные банки

3) специализированные кредитно-финансовые организации.

Целевым рынком для данной разработки является рынок кредитования, осуществляющих выдачу кредитов физическим лицам.

Исходя из вышеизложенного сегментацию рынка можно произвести по виду потребителей:

1) организации, входящие в банковскую систему

2) парабанковские организации

### 4.1.2 Анализ конкурентных технических решений

В качестве конкурентов разработки целесообразно рассмотреть конкурентные технические решения, представляющие собой банковские системы для обработки данных по кредитованию физических лиц. В качестве конкурентных продуктов выбраны следующие:

«Неофлекс» предлагает банкам комплекс продуктов для управления рисками, в том числе для принятия решения по кредитной заявке, требующего всестороннего анализа потенциального заемщика: социально-демографические факторы, финансовое состояние, история взаимоотношений с кредитором, качество кредитной истории, профиль в социальных сетях и многое другое. Сегодня при необходимости можно получить почти любую информацию, но не все данные необходимы для оценки кредитоспособности, большинство источников являются платными, да и время на принятие решения ограничено. «Неофлекс» обладает обширным опытом внедрения систем принятия решений по кредитным заявкам, экспертизой по взаимодействию с большинством известных источников информации и обеспечению бесперебойной работы высоконагруженных кредитных конвейеров на многие тысячи заявок в день.

Кредитный конвейер Brainysoft для банков и МФО предназначен для полной автоматизации процесса кредитования физических и юридических лиц, в том числе в режиме онлайн, включая индивидуальные доработки под бизнес-процессы клиента. Программа предлагает разнообразие каналов поступления заявок, идентификацию личности, обработку и хранение электронных документов, три режима принятия решения по кредиту (ручной, полуавтоматический, автоматический), осуществление денежных операций и клиентское сопровождение. Благодаря API реализуется установление связи с АБС. Кроме того, интеграционный шлюз открывает возможности для подключения внешних сервисов и систем.

Sputnik Data Connector – агрегатор интеграций с сервисами, предоставляющими информацию по физическим и юридическим лицам.

Сервис позволяет получить полную картину по заемщику для МФО, кредитно-потребительских кооперативов, банков, ломбардов и других финансовых организаций. На сегодняшний день компании не имеют возможности кредитовать граждан других стран из-за отсутствия интеграций с иностранными источниками информации. Сервис подключен к ряду иностранных БКИ, государственных и негосударственных сервисов, что позволяет компаниям работать с иностранными заемщиками.

### 4.1.3 SWOT-анализ

SWOT – Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы) – представляет собой комплексный анализ научно-исследовательского проекта. Применяется для исследования внешней и внутренней среды проекта. Сильные и слабые стороны являются факторами внутренней среды объекта анализа, возможности и угрозы – факторами внешней среды. Результат проведения первого этапа SWOT-анализа приведен в таблице 3.

Таблица 3 – Матрица SWOT

| | **Сильные стороны научно-исследовательского проекта:**<br>С1. Использование математической точности при прогнозе надежности заемщика при кредитовании.<br>С2. Снижение коррупционной составляющей.<br>С3. Востребованность подобных систем на рынке кредитования. | **Слабые стороны научно-исследовательского проекта:**<br>Сл1. Сложность поддержки продукта.<br>Сл2. Отсутствие финансирования.<br>Сл3. Низкий уровень проникновения на рынок и популярность. |
|---|---|---|
| **Возможности:** В1. Возможность продажи | | |

| | | |
|---|---|---|
| кредитным организациям. В2. Использовать возможности разработки ПО предоставляемого ТПУ. | | |
| **Угрозы:** У1. Появление аналогов для той же целевой аудитории. | | |

Чтобы более детально рассмотреть степень соответствия возможностей и угроз сильным и слабым сторонам проекта, построим две интерактивные матрицы (таблица 4 – 5).

Таблица 4 – Интерактивная матрица проекта (сильные стороны)

| **Сильные стороны проекта** | | | | |
|---|---|---|---|---|
| | | С1 | С2 | С3 |
| **Возможности проекта** | В1 | - | + | + |
| | В2 | + | - | - |
| **Угрозы проекта** | У1 | - | - | + |

Таблица 5 – Интерактивная матрица проекта (слабые стороны)

| **Слабые стороны проекта** | | | | |
|---|---|---|---|---|
| | | Сл1 | Сл2 | Сл3 |
| **Возможности проекта** | В1 | - | - | + |
| | В2 | - | - | - |
| **Угрозы проекта** | У1 | - | + | + |

В результате работ над данным подпунктом была составлена итоговая матрица SWAT-анализа (таблица 6), выявлены сильные и слабые стороны проекта и возможные направления дальнейших улучшений системы.

Таблица 6 – Итоговая матрица SWOT

| | **Сильные стороны научно-исследовательского проекта:**<br>С1. Использование математической точности при прогнозе надежности заемщика при кредитовании.<br>С2. Снижение коррупционной составляющей.<br>С3. Востребованность подобных систем на рынке кредитования. | **Слабые стороны научно-исследовательского проекта:**<br>Сл1. Сложность поддержки продукта.<br>Сл2. Отсутствие финансирования.<br>Сл3. Низкий уровень проникновения на рынок и популярность. |
|---|---|---|
| **Возможности:** В1. Возможность продажи кредитным организациям. В2. Использовать возможности разработки ПО предоставляемого ТПУ. | Использование кредитными организациями разрабатываемой скоринговой модели приведет к снижению коррупционной составляющей при выдаче кредитов, что также несомненно повышает востребованность подобного программного продукта на рынке кредитования. ПО, предоставляемое ТПУ обеспечивает разработке мощный математический аппарат для составления скоринг-карты. | В связи с тем, что у разрабатываемой системы скоринга существуют конкурентные технические решения, возможен низкий уровень проникновения на рынок и популярность. |
| **Угрозы:** У1. Появление аналогов для той же целевой аудитории. | Появление аналогов скоринговых моделей, отвечающих интересам той же целевой аудитории, напрямую влияет на востребованность данной модели на рынке кредитования. С ростом количества аналогов спрос на конкретную систему становится меньше. | Отсутствие финансирования замедляет процесс разработки скоринговой системы, вследствие чего уровень проникновения на рынок низкий. Программный продукт имеет низкую популярность у потенциальных потребителей. |

## 4.2 Планирование научно-исследовательских работ

### 4.2.1 Структура работ в рамках научного исследования

При организации работ в рамках научно-исследовательской работы необходимо планировать занятость каждого участника проекта в работе. На данном этапе определяется полный перечень работ, распределение времени работ между всеми участниками. В качестве структуры, показывающей необходимые данные, используется линейный график работ, представленный в таблице 7 (НР – научный руководитель; С – студент).

Таблица 7 – Перечень этапов работ и распределение исполнителей

| Основные этапы | № работ | Содержание работ | Исполнители |
|---|---|---|---|
| Разработка технического задания | 1 | Постановка целей и задач, определение исходных данных | С |
| | 2 | Составление и утверждение ТЗ | НР, С |
| | 3 | Составление и утверждение календарного плана работ | С |
| Определение направлений исследований и проектирование | 4 | Подбор и изучение материалов по теме | С |
| | 5 | Уточнение и корректировка методов решения | С, НР |
| Реализация | 6 | Проектирование | С |
| | 7 | Разработка | С |
| | 8 | Обсуждение проблем и отладка | С, НР |
| Тестирование | 9 | Тесты, оптимизация работы | С |
| Анализ и оформление результатов | 10 | Анализ результатов исследований | С |
| | 11 | Оформление результатов | С |
| | 12 | Проверка работы | НР |

### 4.2.2 Определение трудоемкости выполнения работ и разработка графика проведения научного исследования

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях. Она носит вероятностный характер, вследствие ее зависимости от множества трудно учитываемых факторов.

Для расчета длительности работ в календарных днях использовался соответствующий 2021 календарному году коэффициент календарности, равный 1,49.

Рассчитываемое время для каждого из этапов проекта представлено в приложении Б.

На основе временных показателей для научного исследовательского проекта был построен календарный план-график. В графике использованы максимальные по длительности исполнения работы из работ студента и научного руководителя. Календарный план-график в виде диаграммы Ганта представлен в Приложении В.

### 4.3 Бюджет научно-технического исследования (НТИ)

Для подробного планирования бюджета НТИ, необходимо отразить все расходы, связанные с его выполнением. Далее каждая из статей расходов будет рассмотрена подробно.

#### 4.3.1 Расчет материальных затрат НТИ

В материальных затратах были учтены только расходы на канцелярские принадлежности, так как все остальные необходимые для работы над проектом материалы имелись в распоряжении исполнителей. Материалы, необходимые для выполнения данной работы, и расчет материальных затрат представлены в таблице 8.

Таблица 8 – Материальные затраты

| **Наименование** | **Единица измерения** | **Количество** | **Цена за ед., руб.** |
|---|---|---|---|
| Бумага для принтера, А4 | уп. | 1 | 243,00 |
| Шариковая ручка | шт. | 2 | 126,00 |
| **Итого:** | | | 369,00 |

### 4.3.2 Расчет амортизационных затрат

Затраты на специальное оборудование приведены в таблице 9. В специальное оборудование входят оборудование для рабочего места и ПО, необходимое для реализации скоринговой модели.

Таблица 9 – Затраты на специальное оборудование

| **Наименование оборудования** | **Количество единиц оборудования** | **Цена за 1 ед. оборудования** | **Затраты, руб.** |
|---|---|---|---|
| Персональный компьютер (ноутбук) | 1 | 40000 | 40000 |
| Компьютерная мышь | 1 | 299 | 299 |
| Windows 10 Home | 1 | 9990 | 9990 |
| Годовая подписка PyCharm (IDE для профессиональной разработки на Python) | 1 | 14764 (199 USD) | 14764 |
| Ежегодное продление ТПУ лицензионного соглашения SAS | 1 | 5000 | 5000 |
| **Итого:** | | | 70053 |

Амортизационные отчисления для рассматриваемого проекта включают в себя амортизацию используемого оборудования за время выполнения работы. Амортизационные отчисления рассчитываются по времени использования компьютера по формуле:

$$С_{АМ} = \frac{Н_А \cdot Ц_{ОБ}}{F_д} \cdot t_{рф} \cdot n, \quad (7)$$

где $Н_А$ – годовая норма амортизации;

$Ц_{ОБ}$ – цена оборудования;

$F_д$ – действительный годовой фонд рабочего времени;

$t_{рф}$ – время работы вычислительной техники;

$n$ – число задействованных единиц оборудования, n = 1.

Годовая амортизация $Н_А$ определяется как величина, обратная сроку амортизации оборудования $С_А$, который определяется согласно постановлению правительства РФ «О классификации основных средств, включенных в амортизационные группы». Для компьютера и периферийного оборудования, использующегося с ним, примем $С_А$ = 3 года, тогда $Н_А$ = 0,33.

Расчет затрат на амортизационные отчисления представлен в таблице 10.

Таблица 10 – Затраты на амортизационные отчисления

| **Наименование оборудования** | **Норма аморт. Оборуд., $Н_А$** | **Стоим. Оборуд., $Ц_{об}$, руб.** | **Факт. р/вр. Оборуд., $t_{рф}$, ч** | **Действ. Год. Фонд р/вр., $F_Д$, ч.** | **Аморт. Отчисл., $С_{АМ}$, руб.** |
|---|---|---|---|---|---|
| Персональный компьютер (ноутбук) | 0,33 | 40000 | 240 | 1720 | 1841,86 |
| **Итого:** | | | | | 1841,86 |

### 4.3.3 Основная заработная плата исполнителей темы

Оклад научного руководителя от ТПУ (доцента, к.т.н) в среднем составляет 34615 рубля (без учета районного коэффициента).

Оклад студента (стипендия студента) в среднем составляет 3750 руб. (без учета районного коэффициента).

Планирование основной заработной платы приведено в приложении Г.

Таблица 11 – Баланс рабочего времени

| **Показатели рабочего времени** | **Студент** | **Научный руководитель** |
|---|---|---|
| Календарное число дней | 365 | 365 |
| Количество нерабочих дней<br>− выходные дни<br>− праздничные дни | 120 | 120 |

| | | |
|---|---|---|
| Потери рабочего времени<br>− отпуск<br>− невыходы по болезни | 30 | 30 |
| Действительный годовой фонд рабочего времени | 215 | 215 |

Таблица 12 – Расчет основной заработной платы

| **Исполнитель** | **Разряд** | **$З_{тс}$, руб.** | **$k_{пр}$** | **$k_д$** | **$k_р$** | **$З_м$, руб** | **$З_{дн}$, руб** | **$Т_р$, раб. Дн** | **$З_{осн}$, руб** |
|---|---|---|---|---|---|---|---|---|---|
| **Студент** | - | 3750 | - | - | 1,3 | 4875 | 163 | 30 | 4875 |
| **Научный руководитель** | **Ведущий программист** | 34615 | - | - | 1,3 | 45000 | 1500 | 5 | 7500 |
| **Итого:** | | | | | | | | | 12375 |

### 4.3.4 Дополнительная заработная плата исполнителей темы

Расчет дополнительной заработной платы ведется по формуле:

$$З_{доп} = k_{доп} \cdot З_{осн}, \quad (8)$$

где $З_{доп}$ – затраты по дополнительной заработной плате исполнителей темы;

$k_{доп}$ – коэффициент дополнительной заработной платы (на стадии проектирования принимается равным 0,12 – 0,15);

$З_{осн}$ – затраты по основной заработной плате исполнителей темы.

В данном случае коэффициент дополнительной заработной платы будет взят равным 0,13. Таким образом, затраты на дополнительную заработную плату можно считать равными 1608,75 рублей.

### 4.3.5 Отчисления 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 формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}), \quad (9)$$

где $З_{внеб}$ – затраты на отчисления во внебюджетные фонды;

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

$З_{осн}$ – затраты по основной заработной плате исполнителей темы;

$З_{доп}$ – затраты по дополнительной заработной плате исполнителей темы.

Результаты расчета приведены в таблице 13.

Таблица 13 – Отчисления во внебюджетные фонды

| **Исполнитель** | **Основная заработная плата, руб** | **Дополнительная заработная плата, руб** |
|---|---|---|
| Студент | 4875 | 633,75 |
| Научный руководитель | 7500 | 975 |
| Коэффициент отчислений во внебюджетные фонды | 0,302 | |
| **Итого** | | |
| Студент | 1663,64 | |
| Научный руководитель | 2559,45 | |

### 4.3.6 Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов: печать и ксерокопирование материалов исследования, оплата услуг связи, электроэнергии, почтовые и телеграфные расходы, размножение материалов и т.д. Их величина определяется по следующей формуле:

$$З_{накл} = (\text{сумма статей } 1 \div 5) \cdot k_{нр}, \quad (10)$$

где $З_{накл}$ – затраты на накладные раходы;

$k_{нр}$ – коэффициент, учитывающий накладные расходы.

Величину коэффициента накладных расходов можно взять в размере 16%.

Таким образом, накладные расходы для максимального по длительности исполнения работ можно считать равными 14217,91 рублей.

### 4.3.7 Формирование бюджета затрат научно-исследовательского проекта

По итогам всех рассчитанных статей приведён расчёт бюджета затрат НТИ (таблица 14).

Таблица 14 – Расчет бюджета затрат НТИ

| **Наименование статьи** | **Сумма, руб** | | **Примечание** |
|---|---|---|---|
| | **Студент** | **Научный руководитель** | |
| Материальные затраты | 369 | 0 | 3.1 |
| Затраты на специальное оборудование для научных работ | 65053 | 5000 | 3.2 |
| Амортизационные затраты | 1841,86 | 0 | 3.2 |
| Основная заработная плата исполнителей | 4875 | 7500 | 3.3 |
| Дополнительная заработная плата исполнителей | 633,75 | 975 | 3.4 |
| Отчисления во внебюджетные фонды | 1663,64 | 2559,45 | 3.5 |

| | | | |
|---|---|---|---|
| Накладные расходы | 5600,99 | 8616,92 | 3.6 |
| Бюджет затрат НТИ | 104688,61 | | 3.7 |

### 4.4 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальное и экономической эффективности исследования

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

Интегральный финансовый показатель разработки определяется как:

$$I_{\text{финр}}^{\text{исп.}i} = \frac{\Phi_{\text{p}i}}{\Phi_{max}}, \quad (11)$$

где $I_{\text{финр}}^{\text{исп.}i}$ – интегральный финансовый показатель разработки;

$\Phi_{\text{p}i}$ – стоимость i-го варианта исполнения;

$\Phi_{max}$ – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное увеличение бюджета затрат разработки в разах (значение больше единицы), либо соответствующее численное удешевление стоимости разработки в разах (значение меньше единицы, но больше нуля).

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить следующим образом:

$$I_{pi} = \sum a_i \cdot b_i, \quad (12)$$

где $I_{pi}$ – интегральный показатель ресурсоэффективности для i-го варианта исполнения разработки;

$a_i$ – весовой коэффициент i-го варианта исполнения разработки;

$b_i$ – бальная оценка i-го варианта исполнения разработки, устанавливается экспертным путем по выбранной шкале оценивания;

$n$ – число параметров сравнения.

Результаты расчета интегрального показателя ресурсоэффективности приведены в таблице 15.

Таблица 15 – Расчет интегрального показателя ресурсоэффективности

| Объект исследования/Критерии | Весовой коэффициент параметра | Исп.1 |
|---|---|---|
| 1.Потребность в ресурсах памяти | 0,1 | 4 |
| 2.Функциональность | 0,1 | 3 |
| 3.Простота эксплуатации | 0,1 | 4 |
| 4.Скорость работы | 0,15 | 5 |
| 5.Надежность | 0,15 | 4 |
| 6.Удобство эксплуатации | 0,2 | 3 |
| 7.Точность | 0,2 | 4 |
| Iр | | 3,85 |

Результаты расчета интегрального показателя эффективности разработки приведены в таблице 16.

Таблица 16 – Результаты расчета интегрального показателя эффективности разработки

| № п/п | Показатели | Исп.1 |
|---|---|---|
| 1 | Интегральный финансовый показатель разработки | 1 |
| 2 | Интегральный показатель ресурсоэффективности разработки | 3,85 |
| 3 | Интегральный показатель эффективности | 3,85 |

Таблица показала, что разрабатываемая в данной работе скоринговая модель, имеет показатель эффективности выше среднего.

**Вывод по разделу**

В результате проделанной по разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» работы, был исследован проект, выполняемый в рамках научно-исследовательской работы, определены слабые и сильные стороны проекта, его конкуренты, потенциальные потребители и эффективность. Для проекта был посчитан бюджет затрат, он составил 104688,61 руб.

# ГЛАВА 5 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

## Введение

В наши дни, объемы кредитования в России растут и в связи с этим, возникают проблемы увеличения дефолтов, когда заемщики не выполняют свои кредитные обязательства. Так как, прибыль банки получают только при возврате денег заемщиками, чем меньше у них ненадежных клиентов, тем выше их доход.

Одним из примеров решения вышеприведенных проблем является применение скоринг-систем коммерческими банками или кредитными организациями. Данные системы позволяют улучшить качества кредитного портфеля, и, следовательно, на его финансовое состояние. Скоринг-системы – достаточно сложные информационно-технологические системы, которые постоянно требуют обновлений и улучшений, в связи с быстроменяющейся экономической ситуацией и устаревающими данными о заемщиках.

Научно-исследовательская работа заключается в создании скоринговой модели с собственными коэффициентами и параметрами оценки, на основе исторических данных заемщиков при использовании современные технологические решений. Использование скоринговой модели позволит повысить качество оценки потенциального заемщика, снизить риски кредитных организаций, и будет использоваться сотрудниками банковских организаций для упрощения процесса оценки.

Скоринговая карта разрабатывалась во время обучения в ТПУ, ИШИТР, ОИТ в аудитории кибернетического центра ТПУ.

## 5.1 Правовые и организационные вопросы обеспечения безопасности

Трудовой Кодекс РФ [14] устанавливает права и обязанности работника и работодателя, регулирует вопросы охраны труда, трудоустройства, правила оплаты и нормирования труда, порядок разрешения трудовых споров и другое.

В статье 108 Трудового Кодекса РФ «Перерывы для отдыха и питания» сказано, что в течение рабочего дня работнику должен быть предоставлен перерыв продолжительностью не более двух часов и не менее 30 минут, который в рабочее время не включается.

В соответствии со статьей 162 ТК РФ «Введение, замена и пересмотр норм труда» о введении новых норм труда работники должны быть извещены не позднее чем за два месяца.

Согласно статье 163 ТК РФ «Обеспечение нормальных условий работы для выполнения норм выработки», работодатель обязан обеспечить:

- исправное состояние помещений и оборудования;
- надлежащее качество материалов и инструментов, необходимых для выполнения работы, их своевременное предоставление работнику;
- условия труда, соответствующие требованиям охраны труда и безопасности производства.

Согласно статье 212 ТК РФ «Обязанности работодателя по обеспечению безопасных условий и охраны труда», работодатель обязан обеспечить:

- безопасность работников при эксплуатации зданий, оборудования, осуществлении технологических процессов, применяемых материалов;
- создание и функционирование системы управления охраной труда;

Следуя статье 219 «Право работника на труд в условиях, отвечающих требованиям охраны труда» ТК РФ, каждый работник имеет право на:

- соответствующее требованиям охраны труда рабочее место;

• обязательное социальное страхование от несчастных случаев на производстве и профессиональных заболеваний;

• получение достоверной информации от работодателя об условиях и охране труда на рабочем месте, о существующем риске повреждения здоровья, мерах защиты от воздействия вредных и опасных факторов производства;

• отказ от выполнения работ в случае возникновения опасности для его жизни и здоровья вследствие нарушения требований охраны труда до устранения такой опасности.

Организация рабочего места при выполнении работы должна производиться в соответствии с требованиями ГОСТ 12.2.032-78 «Система стандартов безопасности труда. Рабочее место при выполнении работ сидя» [13] и соблюдением трудовых норм, регулирующихся Трудовым кодексом РФ.

С учетом требований ГОСТ 12.2.032-78:

• конструкция рабочего места и взаимное расположение всех его элементов (сиденье, органы управления, средства отображения информации и т.д.) должны соответствовать физиологическим и психологическим требованиям, а также характеру работы.

• высота рабочего стола с клавиатурой должна составлять 680 - 800 мм над уровнем стола;

• высота экрана над полом – 900-1280 мм, монитор должен находиться в 600-700 мм от работника на 20 градусов ниже уровня глаз;

• конструкцией оборудования и рабочего места должно быть обеспечено оптимальное положение работающего, которое достигается регулированием: высоты рабочей поверхности, сиденья, пространства для ног.

• при работе двумя руками органы управления размещают с таким расчетом, чтобы не было перекрещивания рук.

• очень часто используемые средства отображения информации, требующие точного и быстрого считывания показаний, следует располагать в

вертикальной плоскости под углом ±15° от нормальной линии взгляда и в горизонтальной плоскости под углом ±15° от сагиттальной плоскости.

На рабочем месте, предоставленном для работы со скоринговой системой, были учтены и соблюдены все требования по организации труда.

### 5.2 Производственная безопасность

На рабочих местах разработчики программного обеспечения или сотрудники банка, работающие со скоринговой системой, подвергаются воздействию различных вредных и опасных факторов. Все выявленные факторы, этапы работ, во время которых работники могут столкнуться с их влиянием, а также нормативные документы, относящиеся к ним, представлены в таблице 17.

Таблица 17 – Возможные опасные и вредные факторы

| Факторы (ГОСТ 12.0.003-2015) | Этапы работ | | | Нормативные документы |
|---|---|---|---|---|
| | Разработка | Внедрение | Эксплуатация | |
| Недостаток освещенности рабочей зоны | + | + | + | СП 52.13330.2016 «Естественное и искусственное освещение». Актуализированная редакция СНиП 23-05-95 [15] |
| Отклонение показателей микроклимата рабочей зоны | + | + | + | СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных помещений» [16] |
| Повышенный уровень шума на рабочем месте | + | + | + | СН 2.2.4/2.1.8.562-96 «Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки» [17] |
| Повышенное значение напряжения в | + | + | + | ГОСТ 12.1.038-82 ССБТ. «Электробезопасность. |

| | | | | |
|---|---|---|---|---|
| электрической цепи, замыкание которой может произойти через тело человека | | | | Предельно допустимые уровни напряжений прикосновения и токов» [18] |
| Психические перегрузки работника | | + | + | Трудовой кодекс Российской Федерации от 30.12.2001 №197-ФЗ (от 20.04.2021) [14] |

Исходя из данной таблицы можно сделать вывод, что на работников в ходе их деятельности оказывают влияние физические и психологические факторы, в то время как влияние химических и биологических факторов отсутствует.

### 5.2.1 Недостаточная освещенность рабочей зоны

Такой вредный фактор как недостаточная освещенность рабочей зоны возникает вследствие отсутствия должного количества источников освещения в рабочей зоне. Недостаточная освещенность снижает работоспособность, значительно влияет на здоровье работников, а именно на их качество зрения.

В СП 52.13330.2016 зрительная работа сотрудника, работающего с ПК охарактеризована как работу разряда Б – высокой точности (наименьший эквивалентный размер объекта различения - 0,3-0,5 мм), подразряда 1 (относительная продолжительность зрительной работы при направлении зрения на рабочую поверхность не менее 70%). В таблице 18 представлены требования к освещению рабочего помещения для вышеуказанного разряда.

Таблица 18 – Требования к освещению рабочего помещения для разряда Б1

| **Искусственное освещение** | | | | **Естественное освещение** |
|---|---|---|---|---|
| Освещенность на рабочей поверхности от системы | Цилиндри-ческая освещен-ность, лк | Объединен-ный показатель диском- | Коэффициент пульсации освещенности, | Коэффициент естественной освещенности, %, при |

| общего освещения, лк | | форта, не более | Кп, %, не более | Верхн-ем или комби-ниро-ванном | Боко-вом |
|---|---|---|---|---|---|
| 300 | 100 | 21 | 15 | 3 | 1 |

Для снижения влияния фактора недостаточной освещенности на рабочем месте необходимо, чтобы уровень естественного освещения и яркость экрана персонального компьютера были приблизительно одинаковыми, так как яркий свет в зоне периферийного зрения заметно увеличивает глазное напряжение и приводит к быстрой утомляемости. Путем решения проблемы недостаточной освещенности помещения может стать расширение оконного проема или установка качественных источников искусственного освещения.

### 5.2.2 Отклонение показателей микроклимата

Использование персональных компьютеров может привести к повышению температуры и снижению относительной влажности в рабочем помещении. Это ведет к изменению микроклимата.

Отклонение показателей микроклимата от комфортных может повлиять на здоровье работников. Понижение температуры и повышение скорости движения воздуха способствуют усилению конвективного теплообмена и процесса теплоотдачи при испарении пота, что может привести к переохлаждению организма. Недостаточная влажность приводит к интенсивному испарению влаги со слизистых оболочек, что приводит к их пересыханию, растрескиванию, а затем и заражению болезнетворными микробами.

Нормативные показатели микроклимата регламентируются СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных помещений». Санитарные нормы устанавливают оптимальные и допустимые значения показателей в рабочей зоне, что позволяет создавать благоприятные

условия работы, соответствующие физиологическим потребностям организма человека.

Работа, выполняемая разработчиком или работником банка, относится к категории Iа, она является мало подвижной и мало интенсивной, выполняется в положении сидя с минимальными физическими напряжениями. В таблице ниже представлены оптимальные значения показателей микроклимата на рабочих местах для данной категории.

Таблица 19 – Оптимальные величины показателей микроклимата

| **Период года** | **Температура воздуха, °С** | **Температура поверхностей, °С** | **Относительная влажность воздуха, %** | **Скорость движения воздуха, м/с** |
|---|---|---|---|---|
| **Холодный** | 22-24 | 21-25 | 60-40 | 0,1 |
| **Теплый** | 23-25 | 22-26 | 60-40 | 0,1 |

Для минимизации воздействия отклонений показателей микроклимата в рабочем помещении необходимо использовать кондиционеры, обогреватели и увлажнители воздуха, которые помогают привести эти показатели к оптимальным значениям.

### 5.2.3 Повышенный уровень шума на рабочем месте

Повышенный уровень шума на рабочем месте обусловлен использованием персональных компьютеров, наличием центральной системы вентиляции и кондиционирования воздуха. Становится сложнее разбирать речь, работоспособность снижается и повышается утомляемость сотрудников.

Ниже представлены предельно допустимые уровни звукового давления, уровни звука и эквивалентные уровни звука для разработчиков программного обеспечения и людей, работающих с программным обеспечением, описанные в СН 2.2.4/2.1.8.562-96.

Таблица 20 – Предельно допустимые уровни звукового давления, уровни звука и эквивалентные уровни звука

| Вид трудовой деятельности, рабочее место | Уровни звукового давления, дБ в октавных полосах со среднегеометрическими частотами, Гц | | | | | | |
|---|---|---|---|---|---|---|---|
| | 31,5 | 63 | 125 | 250 | 500 | 1000 | 2000 |
| Конструирование и проектирование, программирование. Рабочие места в помещениях дирекции, проектно-конструкторских бюро, расчетчиков, программистов вычислительных машин, в лабораториях для теоретических работ и обработки данных. | 86 | 71 | 61 | 54 | 49 | 45 | 42 |

Существуют следующие пути уменьшения воздействий шума: экранирование рабочих мест (установка перегородок между рабочими местами); установка менее шумного оборудования; чистка оборудования от пыли, замена смазывающих веществ, т.к. любое оборудование при загрязнении увеличивает уровень шума.

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

При работе с персональным компьютером, при разработке скоринг-систем или их использовании широко используется электричество для питания компьютерной техники, которое может являться источником опасности.

Правила электробезопасности описаны в ГОСТ 12.1.038-82 ССБТ. «Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов». Несоблюдение данных правил может привести к опасным последствиям, таким как поражение электрическим током, которое может произойти при прикосновении к токоведущим частям, находящимся под напряжением. Электрический ток оказывает на человека термическое, электролитическое, биологическое и механическое воздействия, что может привести к травмам или гибели.

Допустимое значение напряжения прикосновения для переменного тока частотой 50 Гц составляет 2 В, а силы тока – 0,3 мА. Для тока частотой 400 Гц, соответственно – 2 В и 0,4 мА. Для постоянного тока – 8 В и 1 мА.

Защитить себя от воздействия электрического тока можно используя оградительные устройства, устройства автоматического контроля и сигнализации, изолирующие устройства и покрытия, устройства защитного заземления, устройства автоматического отключения, предохранительные устройства.

### 5.2.5 Психические перегрузки работника

При долгой монотонной работе с ПК и программным обеспечением у работника могут возникнуть состояние утомления и психические перегрузки, то есть совокупность таких сдвигов в психофизиологическом состоянии организма человека, которые развиваются после совершения работы и приводят к временному снижению эффективности труда.

Можно выделить следующие нервно-психические перегрузки:

- умственное перенапряжение;
- перенапряжение анализаторов;
- монотонность труда;
- эмоциональные перегрузки.

Для того чтобы предотвратить подобные перегрузки при постоянной работе и избежать психического перенапряжения необходимо дать нервной системе расслабиться, путем чередования периодов отдыха и работы. Установить режим, начать заниматься спортом, ложиться спать и просыпаться в одно и то же время. В случаях, когда справиться самостоятельно не удается - обратиться к врачу.

## 5.3 Экологическая безопасность

Разрабатываемая скоринг-система не оказывает влияния на окружающую среду, так как разрабатывается и используется внутри

персональных компьютеров. Тем не менее можно выделить некоторые источники загрязнения для атмосферы, гидросферы и литосферы планеты.

*Атмосфера.* Источником загрязнения являются захоронения отходов, которые выбрасывают в атмосферу токсические вещества. Свалки бытовых отходов, несанкционированные свалки, полигоны ТБО за чертой города – источники биогазов, изменяющих состав воздуха.

Должны соблюдаться требования нормативных актов, регулирующих отношения в области охраны атмосферного воздуха и быть предприняты следующие меры по предотвращению загрязнения атмосферы: сокращение отходов в ходе промышленного производства; вторичная переработка отходов; использование альтернативных источников энергии; озеленение планеты, частично восстанавливающее начальное соотношение газов в атмосфере.

*Гидросфера.* Источником загрязнения гидросферы при работе в офисе являются детергенты. Это вещества, которые добавляются в моющие средства. Они снижают поверхностное натяжение воды. Это приводит к усилению вспенивания и лучшему очищению поверхностей от загрязненности.

К детергентам относятся: очистители; красящие вещества и пигменты; пластиковые и поливинилхлоридные компоненты; средства для мытья посуды и поверхностей; порошкообразные и гелеобразные стиральные порошки и т.п.

Должны соблюдаться требования нормативных актов, регулирующих отношения в области охраны водных ресурсов. Охрана гидросферы должна включать в себя фильтрацию сточных вод. Для обеспечения безопасного пользования гидросферой следует оборудовать раздельные системы хозяйственно-бытовой и ливневой канализации.

*Литосфера.* Загрязнение литосферы происходит от утилизации техники, бумажных и иных отходов.

Утилизация компьютерной и организационной техники ограничена законодательно, так как при ее производстве используются материалы, способные нанести большой вред окружающей среде. Утилизация

компьютерного оборудования происходит через обязательное извлечение компонент, их сортировку и последующую отправку для повторного использования. Такая утилизация обязательно производится на оборудованных полигонах с привлечением квалифицированного персонала.

Другие мусорные отходы (бумажная макулатура, отходы от канцелярских принадлежностей, продуктов питания, продуктов личной гигиены) проходят через обязательную сортировку и далее утилизируются. Некоторые отходы, которые могут быть использованы повторно, после сортировки отправляют на переработку через занимающиеся сбором таких отходов компании.

Утилизируя отходы от работы таким образом, можно сократить вредное воздействие на окружающую среду и здоровье человека, исключая отравление опасными веществами и попадание тяжелых металлов в организм.

### 5.4 Безопасность в чрезвычайных ситуациях

К возможным чрезвычайным ситуациям на рабочем месте разработчика программного обеспечения или использующего его на ПК работника, выделяют пожар, землетрясения, экстремальные погодные условия (очень низкая или высокая температура воздуха, снежная буря, ураган) и т.п.

Однако, учитывая наличие большого количества вычислительной техники в помещении и постоянное использование электрического тока, при несоблюдении правил электробезопасности наиболее вероятно возникновение пожара. Под ним понимается неконтролируемый процесс горения, обусловленный возгоранием вычислительной техники и угрожающий жизни и здоровью. При использовании электрооборудования причиной пожара является искра, источником которой является либо короткое замыкание, обусловленное использованием неисправного оборудования, либо нагрев участка электросети вследствие больших переходных сопротивлений или перегрузок.

Согласно ГОСТ 12.1.004-91 «Пожарная безопасность. Общие требования», при работе с компьютером необходимо соблюдать следующие нормы пожарной безопасности:

Чтобы предотвратить возникновение данной чрезвычайной ситуации, должна проводиться периодическая, своевременная диагностика по обнаружению неисправностей, а также соблюдение персоналом норм пожарной безопасности. Запрещено одновременное подключение к сети количества потребителей, превышающего допустимую нагрузку. Также для предотвращения потенциального пожара рабочие помещения должны быть оснащены средствами пожаротушения, исправным пожарным оборудованием. Также необходимо проводить инструктаж по плану действий в случае возникновения чрезвычайной ситуации у сотрудников.

При обнаружении пожара, любой, увидевший пожар должен: соблюдая покой, немедленно заявить о данном происшествии в пожарную службу по телефонному номеру 01 или 112. Людям должна быть обеспечена возможность беспрепятственного движения по эвакуационным путям.

**Вывод по разделу**

В результате работы по разделу «Социальная ответственность» были выявлены основные нормативные акты для обеспечения безопасности жизнедеятельности на рабочем месте, рассмотрены наиболее значимые опасные и вредные факторы, возникающие при разработке и работе со скоринг-моделями при оценке заемщика, описано влияние процесса работы на окружающую среду и меры, необходимые для уменьшения влияния вредных и опасных факторов на организм человека и для сокращения негативного влияния процесса разработки скоринг-системы на окружающую среду.

# ЗАКЛЮЧЕНИЕ

Использование автоматизированных систем скоринга на сегодняшний день является основным инструментом снижения рисков. Скоринговая система решает выдавать или не выдавать кредит клиенту банка, опираясь на его баллы, которые рассчитываются автоматически на основе характеристик заемщика, параметров запрашиваемого им кредита, иногда от его кредитной истории.

В данной работе была изложена методика построения скоринговой карты на основе модели логистической регрессии, коэффициенты полученные из ее уравнения были масштабированы в скоринговые баллы. Также были рассмотрены методические подходы к формированию и исследованию характеристик потенциального заемщика для построения модели, преобразования данных для успешного построения модели.

Результатом работы являются:

a. Разработанный алгоритм для анализа и предобработки данных, куда входили;

- Проверка и обработка пустых значений
- Удаление дубликатов строк
- Проверка на ошибки и выбросы в данных
- Проверка мультиколлинеарности
- Устранение сильной корреляции признаков
- Нормализация

b. Проанализирован метод логистической регрессии.

c. Реализованы предобработки данных и обучение модели на языке Python. (Приложение А).

d. Выделены весовые коэффициенты для скоринговой карты, определена точность работы предсказательного алгоритма, для наглядности результата построены матрицы путаницы.

# СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Пошаговое построение логистической регрессии в Python [Электронный ресурс]. URL: https://medium.com/nuances-of-programming/пошаговое-построение-логистической-регрессии-в-python-a7c650ae77c2 (дата обращения: 23.05.2021)

2. Библиотеки Python, необходимые для машинного обучения [Электронный ресурс]. URL: https://techrocks.ru/2018/10/05/python-libraries-for-machine-learning/#:~:text=Scikit-learn%20это%20одна%20из%20самых,кластеризацию%2C%20k-means%20и%20т.%20д (дата обращения: 23.05.2021)

3. Логистическая регрессия с использованием Python (scikit-learn) [Электронный ресурс]. URL: https://www.machinelearningmastery.ru/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a/ (дата обращения: 23.05.2021)

4. ML Boot Camp / Руководство для начинающих [Электронный ресурс]. URL: https://mlbootcamp.ru/ru/article/tutorial/ (дата обращения: 23.05.2021)

5. Что есть в новом JupyterLab для пользователей? [Электронный ресурс]. URL: https://habr.com/ru/company/otus/blog/351820/ (дата обращения: 23.05.2021)

6. Логистическая регрессия в Python — Краткое руководство [Электронный ресурс]. URL: https://coderlessons.com/tutorials/python-technologies/logisticheskaia-regressiia-v-python/logisticheskaia-regressiia-v-python-kratkoe-rukovodstvo (дата обращения: 23.05.2021)

7. Logistic Regression / Google Colaboratory [Электронный ресурс]. URL: https://colab.research.google.com/github/GokuMohandas/MadeWithML/blob/main/notebooks/07_Logistic_Regression.ipynb (дата обращения: 23.05.2021)

8. Реализация логистической регрессии в реальном мире [Электронный ресурс]. URL: https://www.machinelearningmastery.ru/real-world-implementation-of-logistic-regression-5136cefb8125/ (дата обращения: 23.05.2021)

9. 5 способов обнаружить выбросы / аномалии, которые должен знать каждый специалист по данным (код Python) [Электронный ресурс]. URL: https://www.machinelearningmastery.ru/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623/ (дата обращения: 23.05.2021)
10. Графики в Pandas: Визуализация данных [Электронный ресурс]. URL: https://python-scripts.com/plot-with-pandas (дата обращения: 23.05.2021)
11. Что такое Матрица путаницы в машинном обучении [Электронный ресурс]. URL: https://www.machinelearningmastery.ru/confusion-matrix-machine-learning/ (дата обращения: 23.05.2021)
12. 10 современных сервисов, сопровождающих выдачу кредитов [Электронный ресурс]. URL: http://futurebanking.ru/post/3433 (дата обращения: 23.05.2021)
13. ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования // Электронный фонд правовой и нормативно-технической документации [Электронный ресурс]. 2021. URL: http://docs.cntd.ru/document/1200003913 (дата обращения: 16.05.2021)
14. Трудовой кодекс Российской Федерации // Электронный фонд правовой и нормативно-технической документации [Электронный ресурс]. 2021. URL: https://docs.cntd.ru/document/901807664 (дата обращения: 16.05.2021)
15. СП 52.13330.2016 «Естественное и искусственное освещение» // Электронный фонд правовой и нормативно-технической документации [Электронный ресурс]. 2021. URL: http://docs.cntd.ru/document/456054197 (дата обращения: 16.05.2021)
16. СанПиН 2.2.4.548-96 Гигиенические требования к микроклимату производственных помещений // Электронный фонд правовой и нормативнотехнической документации [Электронный ресурс]. 2021. URL: http://docs.cntd.ru/document/901704046 (дата обращения: 06.05.2021)
17. СН 2.2.4/2.1.8.562-96 «Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки» // Электронный

фонд правовой и нормативно-технической документации [Электронный ресурс]. 2021. URL: http://docs.cntd.ru/document/901703278 (дата обращения: 16.05.2021)
18. ГОСТ 12.1.038-82 ССБТ «Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов» // Электронный фонд правовой и нормативно-технической документации [Электронный ресурс]. 2021. URL: https://docs.cntd.ru/document/5200313 (дата обращения: 16.05.2021)

## ПРИЛОЖЕНИЕ А

```
import os
import numpy as np
import pandas as pd
import seaborn as sn
from pandas.plotting import scatter_matrix
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt

plt.style.use('ggplot')

# Сброс ограничений на количество выводимых рядов
pd.set_option('display.max_rows', None)
# Сброс ограничений на число столбцов
pd.set_option('display.max_columns', None)
# Сброс ограничений на количество символов в записи
pd.set_option('display.max_colwidth', None)

data = pd.read_csv(os.path.dirname(os.path.abspath('main.py')) +
'\source_filee.csv',
                   header=None, delimiter=';', engine='python')

data.shape
data.tail()
data.head()

data.columns = ['A' + str(i) for i in range(1, 11)] + ['GB']

print(data)
print('\r')

data = data.drop(['A10'], axis=1)  # Выбрасываем столбец
'DayOff' или A10, т.к. целевая функция должна быть одна.

# Разделение признаков на категориальные и количественные
categorical_columns = [c for c in data.columns if
data[c].dtype.name == 'object']
numerical_columns = [c for c in data.columns if
data[c].dtype.name != 'object']
print('categorical and numerical columns respectively')
print(categorical_columns)
print(numerical_columns)

print('\r')

# Перечислим уникальные значения категориальных признаков
for c in categorical_columns:
    print(data[c].unique())
print('\r')

# Теперь, когда мы знаем значения категориальных признаков,
```

```
видно, что
# Некоторые колонки имеют опечатки или слишком много категорий
# Сократим их для оптимизации моделирования.
data['A2']=np.where(data['A2'] =='Drivine licence', 'Driving
licence', data['A2'])

data['A5']=np.where(data['A5'] =='n', '0', data['A5'])

data['GB']=np.where(data['GB'] =='n', '0', data['GB'])

# Исправленные уникальные значения категориальных признаков
print('Ordered categorical values')
for c in categorical_columns:
    print(data[c].unique())
print('\r')

# Удаление дубликатов
print('Deleting duplicates')
data = data.drop_duplicates()
print(data.shape)
print('\r')


# Сводная информация о количественных признаках
print(data.describe())
print('\r')

# Сводная информация о категориальных признаках
print(data[categorical_columns].describe())
print('\r')


# Функция  scatter_matrix позволяет построить для каждой
количественной переменной гистограмму,
# а для каждой пары таких переменных – диаграмму рассеяния
scatter_matrix(data, alpha=0.7, figsize=(10, 10))

# Диаграмма рассеяния для коррелирующих признаков
col1 = 'A7'
col2 = 'A6'
plt.figure(figsize=(10, 6))
plt.scatter(data[col1][data['GB'] == '1'],
            data[col2][data['GB'] == '1'],
            alpha=0.75,
            color='red',
            label='1')

plt.scatter(data[col1][data['GB'] == '0'],
            data[col2][data['GB'] == '0'],
            alpha=0.75,
            color='blue',
            label='0')
```

```
plt.xlabel(col1)
plt.ylabel(col2)
plt.legend(loc='best');

# Корреляционная матрица
print(data.corr())
print('\r')

# Объединение AddIncome(A6) и Income(A7)
data['A7'] = data['A7'] + data['A6']
data = data.drop(['A6'], axis=1)  # Выбрасываем столбец
'AddIncome'.
print(data['A7'])
print('\r')
numerical_columns = [c for c in data.columns if
data[c].dtype.name != 'object']
# Корреляционная матрица
print(data.corr())
print('\r')

# Удаление выбросов
def outlier_detect(df):
    for i in df.describe().columns:
        Q1=df.describe().at['25%',i]
        Q3=df.describe().at['75%',i]
        IQR=Q3 - Q1
        LTV=Q1 - 1.5 * IQR
        UTV=Q3 + 1.5 * IQR
        x=np.array(df[i])
        p=[]
        for j in x:
            if j < LTV or j>UTV:
                p.append(df[i].median())
            else:
                p.append(j)
        df[i]=p
    return df

outlier_detect(data)
print(data)


# Общее количество заполненных ячеек
print('Amount of fillings:')
print(data.count(axis=0))
print('\r')

print('A1 data:')
print(data['A1'].describe())
print('\r')

# Замена пропусков медианой для количественных признаков
data = data.fillna(data.median(axis=0), axis=0)
```

```
print(data.describe())
print('\r')
print(data.count(axis=0))
print('\r')

# Замена пропусков наиболее часто встречающимся значением
# для категориальных признаков
data_describe = data.describe(include=[object])
print(data_describe)
for c in categorical_columns:
    data[c] = data[c].fillna(data_describe[c]['top'])

print('\r')
print(data.describe(include=[object]))
print('\r')
print(data.count(axis=0))
print('\r')


# Категоризация количественных признаков
bins7 = [-1, 850, 2900, 3500, 4000, 4500, 5500, 7500, 10000,
13000, 35000]
labels7 = [str(i) for i in range(1, 11)]
# bins7 = [-1, 3400, 10200, 17000, 23800, 34000]
# labels7 = [str(i) for i in range(1, 6)]
data['A7_cat'] = pd.cut(data['A7'], bins=bins7, labels=labels7)
print(data['A7_cat'].describe())

bins8 = [1, 100000, 130000, 150000, 200000, 250000, 300000,
400000, 500000, 750000]
labels8 = [str(i) for i in range(1, 10)]
data['A8_cat'] = pd.cut(data['A8'], bins=bins8, labels=labels8)
print(data['A8_cat'].describe())

bins9 = [18, 24, 27, 29, 32, 34, 38, 41, 44, 49, 100]
labels9 = [str(i) for i in range(1, 11)]
data['A9_cat'] = pd.cut(data['A9'], bins=bins9, labels=labels9)
print(data['A9_cat'].describe())
print(data.columns)

data = data.drop(['A7'], axis=1)
data = data.drop(['A8'], axis=1)
data = data.drop(['A9'], axis=1)

print(data.columns)

# print (data['A7_cat'].apply(type))

# Рассмотрим пример небинарного признака
print(data['A7_cat'].unique())
print('\r')
```

```
# Векторизация
# data.columns
# Выделим бывшие количественные, категориальные бинарные и
категориальные не бинарные признаки
num_cat_columns = [c for c in data.columns if
data[c].dtypes.name == 'category']
binary_columns = [c for c in categorical_columns if
data_describe[c]['unique'] == 2]
nonbinary_columns = [c for c in categorical_columns if
data_describe[c]['unique'] > 2]
print('numerical, binary and non-binary columns respectively')
print(num_cat_columns)
print(binary_columns)
print(nonbinary_columns)
print('\r')

# Значения бинарных признаков заменим на 0 и 1
for c in binary_columns:
    top = data_describe[c]['top']
    top_items = data[c] == top
    data.loc[top_items, c] = 0
    data.loc[np.logical_not(top_items), c] = 1

# Результат замены значений бинарных признаков
print(data[binary_columns].describe())
print('\r')

# Рассмотрим пример небинарного признака
print(data['A2'].unique())
print('\r')

# Осуществление векторизации
data_nonbinary = pd.get_dummies(data[nonbinary_columns])
print(data_nonbinary.columns)
print('\r')

data_cat_columns = pd.get_dummies(data[num_cat_columns])
print(data_cat_columns.columns)
print('\r')


# Соединим все столбцы в одну таблицу
data = pd.concat((data[binary_columns], data_nonbinary,
data_cat_columns), axis=1)
data = pd.DataFrame(data, dtype=float)
print(data.shape)
print('\r')


X = data.drop(['GB'], axis=1)  # Выбрасываем столбец 'GB'.
y = data['GB']
feature_names = X.columns
print(feature_names)
```

```
print('\r')
print(X.shape)
print(y.shape)
N, d = X.shape

print('\r')
print(data)

# Моделирование алгоритма с помощью машинного обучения
print('\rОбучение')
from sklearn.model_selection import train_test_split

# X_train, y_train – это обучающая выборка, X_test, y_test –
тестовая.
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=1)
N_train, _ = X_train.shape
N_test,  _ = X_test.shape
print(N_train, N_test)

# Обучение модели
# Логистическая регрессия
lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
print(y_pred)

cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix


err_train = np.mean(y_train != lr.predict(X_train))
err_test = np.mean(y_test != lr.predict(X_test))
print(err_train, err_test)
err_test_perc = round(err_test, 3)*100
print('Train-test mistake is: ' + str(err_test_perc) + '%')

#Пройдем тестовые данные и посмотрим, как модель соотносится с
ними (ее точность)
logmodel_score = lr.score(X_test, y_test)
print('\nThis is how Model Scored: ', logmodel_score)
print('\r')

# Расчет коэффициентов с использованием метода coef_
column_label = list(X_train.columns)
lr_Coeff = pd.DataFrame(lr.coef_, columns = column_label)
lr_Coeff['intercept'] = lr.intercept_
print("Coefficient Values Of The Surface Are:\r ", lr_Coeff)
print('\r')


print(metrics.confusion_matrix(y_test, y_pred))
```

```
# Матрица путаницы на обучающейся выборке
y_pred = lr.predict(X_train)
cnf_matrix = metrics.confusion_matrix(y_train, y_pred)
classes = ["positive", "negative"]

df_cfm = pd.DataFrame(cnf_matrix, index = classes, columns =
classes)
plt.figure(figsize = (10, 7))
cfm_plot = sn.heatmap(df_cfm, annot=True)
cfm_plot.figure.savefig("cfm_train.png")

# Матрица путаницы на тестовой выборке
y_pred = lr.predict(X_test)
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
classes = ["positive", "negative"]

df_cfm = pd.DataFrame(cnf_matrix, index = classes, columns =
classes)
plt.figure(figsize = (10, 7))
cfm_plot = sn.heatmap(df_cfm, annot=True)
cfm_plot.figure.savefig("cfm_test.png")
```

## ПРИЛОЖЕНИЕ Б

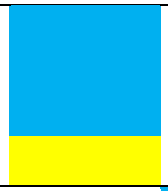
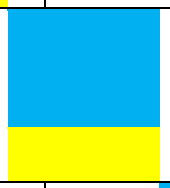
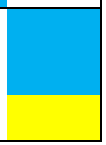
Таблица Б.1 – Временные показатели проведения научного исследования

| **Название работы** | **Трудоёмкость работ** | | | **Исполнители** | **Длительность работ в рабочих днях Трі** | | **Длительность работ в календарных днях Ткі** | |
|---|---|---|---|---|---|---|---|---|
| | **$t_{min}$, чел-дни** | **$t_{max}$, чел-дни** | **$t_{ожі}$, чел-дни** | | С | НР | С | НР |
| Постановка целей и задач, определение исходных данных | 1 | 1 | 1 | С | 1 | 0 | 1,49 | 0 |
| Составление и утверждение ТЗ | 1 | 2 | 1,5 | С, НР | 1 | 0,5 | 1,49 | 0,745 |
| Составление и утверждение календарного плана работ | 1 | 2 | 1,5 | С, НР | 1 | 0,5 | 1,49 | 0,745 |
| Подбор и изучение материалов по теме | 4 | 7 | 5 | С | 5 | 0 | 7,45 | 0 |
| Уточнение и корректировка методов решения | 1 | 2,5 | 2 | С, НР | 1 | 1 | 1,49 | 1,49 |
| Проектирование | 3 | 6 | 4 | С, НР | 3 | 1 | 4,47 | 1,49 |
| Разработка | 3 | 6 | 4 | С, НР | 3 | 1 | 4,47 | 1,49 |
| Обсуждение проблем и отладка | 3 | 8 | 6 | С | 6 | 0 | 8,94 | 0 |
| Тесты, оптимизация работы | 1 | 3 | 2 | С | 2 | 0 | 2,98 | 0 |
| Анализ результатов исследований | 1,5 | 3 | 2 | С | 2 | 0 | 2,98 | 0 |
| Оформление результатов | 2 | 5 | 3 | С | 3 | 0 | 4,47 | 0 |
| Проверка работы | 2 | 4 | 3 | С, НР | 2 | 1 | 2,98 | 1,49 |
| **ИТОГО:** | Студент | | | | 30 | | 44,7 | |
| | Научный руководитель | | | | 5 | | 7,45 | |
| | Всего | | | | 35 | | 52,15 | |

**ПРИЛОЖЕНИЕ В**

Таблица В.1 – Диаграмма Ганта

| № ра-бот | Вид работ | Исполнители | Ткi, кал. дн. | Продолжительность выполнения работ | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Апрель | | Май | | | Июнь |
| | | | | 1 | 3 | 1 | 2 | 3 | 1 |
| 1 | Постановка целей и задач, определение исходных данных | Студент | 1,49 | | | | | | |
| 2 | Составление и утверждение ТЗ | Студент, Научный руководитель | 2,235 | | | | | | |
| 3 | Составление и утверждение календарного плана работ | Студент, Научный руководитель | 2,235 | | | | | | |
| 4 | Подбор и изучение материалов по теме | Студент | 7,45 | | | | | | |
| 5 | Уточнение и корректировка методов решения | Студент, Научный руководитель | 2,98 | | | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | Проектирование | Студент, Научный руководитель | 5,96 | | |  | | | | |
| 7 | Разработка | Студент, Научный руководитель | 5,96 | | |  | | | | |
| 8 | Обсуждение проблем и отладка | Студент | 8,94 | | |  | | | | |
| 9 | Тесты, оптимизация работы | Студент | 2,98 | | |  | | | | |
| 10 | Анализ результатов исследований | Студент | 2,98 | | |  | | | | |
| 11 | Оформление результатов | Студент | 4,47 | | |  | | | | |
| 12 | Проверка работы | Студент, Научный руководитель | 4,47 | | |  | | | | |

Студент – голубой цвет; Научный руководитель – желтый цвет

## ПРИЛОЖЕНИЕ Г

Таблица Г.1 – Планирование основной заработной платы

| **Название работы** | **Трудоём-кость работы** | | **Заработная плата, приходящаяся на один чел.день, руб.** | | **Всего заработная плата по тарифу (окладам)** | |
|---|---|---|---|---|---|---|
| | С | НР | С | НР | С | НР |
| Постановка целей и задач, определение исходных данных | 1 | 0 | 163 | 1500 | 163 | 0 |
| Составление и утверждение ТЗ | 1 | 0,5 | 163 | 1500 | 163 | 750 |
| Составление и утверждение календарного плана работ | 1 | 0,5 | 163 | 1500 | 163 | 750 |
| Подбор и изучение материалов по теме | 5 | 0 | 163 | 1500 | 815 | 0 |
| Уточнение и корректировка методов решения | 1 | 1 | 163 | 1500 | 163 | 1500 |
| Проектирование | 3 | 1 | 163 | 1500 | 489 | 1500 |
| Разработка | 3 | 1 | 163 | 1500 | 489 | 1500 |
| Обсуждение проблем и отладка | 6 | 0 | 163 | 1500 | 978 | 0 |
| Тесты, оптимизация работы | 2 | 0 | 163 | 1500 | 326 | 0 |
| Анализ результатов исследований | 2 | 0 | 163 | 1500 | 326 | 0 |
| Оформление результатов | 3 | 0 | 163 | 1500 | 489 | 0 |
| Проверка работы | 2 | 1 | 163 | 1500 | 326 | 1500 |
| ИТОГО: | | | | | 4875 | 7500 |