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 ***ядерных технологий***
Направление подготовки ***01.04.02 Прикладная математика и информатика***
Отделение школы (НОЦ) ***экспериментальной физики***

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Модель кредитного скоринга

УДК 519.876:336.774

Студент

Группа	ФИО	Подпись	Дата
0ВМ92	Шеров Шерафкан		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ	Семенов М.Е.	к.ф.-м.н., доцент		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОГСН ШБИП	Киселева Е.С.	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Антоневич О.А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ	Семенов М.Е.	к.ф.-м.н., доцент		

Томск – 2021 г.

Планируемые результаты обучения по ООП

Код результата	Результат обучения
ПК(У)-1	Способен проводить научные исследования и получать новые научные и прикладные результаты самостоятельно и в составе научного коллектива
ПК(У)-2	Способен проводить поиск и анализ научной и научно-технической литературы по тематике проводимых исследований
ПК(У)-3	Способен разрабатывать и анализировать показатели качества информационных систем, используемых в производственной деятельности
ПК(У)-4	Способен планировать научно-исследовательскую деятельность, анализировать риски, управлять проектами, управлять командой проекта
ПК(У)-5	Способен к преподаванию математических дисциплин и информатики в образовательных организациях высшего образования
ПК(У)-6	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ОПК(У)-1	Способен решать актуальные задачи фундаментальной и прикладной математики
ОПК(У)-2	Способен совершенствовать и реализовывать новые математические методы решения прикладных задач
ОПК(У)-3	Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности
ОПК(У)-4	Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности с учетом требований информационной безопасности
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке(-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

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 ***ядерных технологий***
Направление подготовки ***01.04.02 Прикладная математика и информатика***
Отделение школы (НОЦ) ***экспериментальной физики***

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ М.Е. Семенов
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

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

Студенту:

Группа	**ФИО**
0ВМ92	Шерову Шерафкану

Тема работы:

Утверждена приказом директора (дата, номер)	15.03.2021 № 74-2/с

Срок сдачи студентом выполненной работы:	31.05.2021

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.).*	*Розничный кредитный портфель банка из 25 906 наблюдений по заемщикам, 1 362 из которых подтвержденные случаи дефолта. Каждый заемщик характеризуется 42 различными параметрами.*

Перечень подлежащих исследованию, проектированию и разработке вопросов *(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. Сравнительный анализ литературы по теме диссертации* *2. Предварительная обработка данных и отбор наиболее значимых признаков для прогнозирования* *3. Построение модели оценки вероятности дефолта заемщика* *4. Составление прогноза вероятности дефолта заемщика для тестовой выборки* *5. Оценка качества работы построенной модели*
Перечень графического материала *(с точным указанием обязательных чертежей)*	*1. Архитектура модели* *2. Графики разведочного анализа модельных данных* *3. Визуализация математической модели*

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Киселева Елена Станиславовна
Социальная ответственность	Антоневич Ольга Алексеевна
Иностранный язык	Утятина Янина Викторовна

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

Литературный обзор

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	15.03.2021

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ	Семенов Михаил Евгеньевич	к. ф.-м. н., доцент		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0ВМ92	Шеров Шерафкан		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
0ВМ92	Шерову Шерафкану

Школа		Отделение школы (НОЦ)	
Уровень образования	Магистратура	Направление/специальность	01.04.02 «Прикладная математика и информатика»

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Стоимость материальных ресурсов определялась в соответствии с рыночными ценами г. Томска. Тарифные ставки исполнителей определены штатным расписанием НИ ТПУ.*
2. *Нормы и нормативы расходования ресурсов*	*Коэффициенты для расчета заработной платы.*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Коэффициент отчислений во внебюджетные фонды – 30,2 %*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*1.Потенциальные потребители результатов исследования; 2.SWOT – анализ; 3.Оценка готовности проекта к коммерциализации*
2. *Разработка устава научно-технического проекта*	*1.Постановка цели, ожидаемых результатов проекта; 2.Определение внутренних и внешних заинтересованных сторон проекта; 3.Определение ограничений/допущений проекта.*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*1.Определение структуры и трудоемкости выполнения работ; 2.Бюджет научно - технического исследования (НТИ); 3. Реестр рисков проекта*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. Сегментирование рынка 2. Матрица SWOT 3. Оценка готовности проекта к коммерциализации 4. Заинтересованные стороны 5. Цели и результат проекта и рабочая группа проект 6. Ограничения/допущения проекта 7. Иерархическая структура работ проекта 8. Комплекс работ по разработке проекта 9. Временные показатели осуществления комплекса работ 10. Календарный план-график выполнения работ (диаграмма Гантта) 11. Расчёт бюджета исследования 12. Реестр рисков	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОГСН ШБИП	Киселева Е.С.	к.э.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0ВМ92	Шеров Шерафкан		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
0ВМ92	Шерову Шерафкану

Школа		Отделение (НОЦ)	
Уровень образования	Магистратура	Направление/специальность	01.04.02 «Прикладная математика и информатика»

Тема ВКР:

Модель кредитного скоринга	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Объект исследования: математическая модель кредитного скоринга. Область применения: банковские системы. Рабочая зона: офисное помещение.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования ГОСТ 12.2.049-80 ССБТ. Оборудование производственное. Общие эргономические требования. ГОСТ 22269-76. Система «человек-машина». Рабочее место оператора. Взаимное расположение элементов рабочего места. Общие эргономические требования.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	Вредные факторы: - повышенный уровень шума; - недостаточная освещенность рабочей зоны; - умственное перенапряжение. - перенапряжение зрительного анализатора Опасные факторы: - воздействие переменных электромагнитных полей - повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека
3. Экологическая безопасность:	Литосфера: утилизация компьютерной техники.
4. Безопасность в чрезвычайных ситуациях:	Возможные ЧС: пожар. Наиболее типичная ЧС: пожар.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД	Антоневич О.А.	к.б.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0ВМ92	Шеров Шерафкан		

РЕФЕРАТ

Выпускная квалификационная работа 113 с., 18 рис., 26 табл., 32 источника.

Ключевые слова: кредитный скоринг, математическая модель скоринга, вероятность дефолта, логистическая регрессия, бинарная классификация, эффективность модели скоринга.

Объектом исследования является розничный кредитный портфель банка по заемщикам.

Цель работы – разработать математическую модель кредитного скоринга для предсказания дефолта заемщиков.

В процессе исследования проводились обзор литературы, предварительная обработка данных и отбор наиболее значимых признаков для прогнозирования, построение модели оценки вероятности дефолта заемщика, составление прогноза вероятности дефолта заемщика для тестовой выборки, оценка качества работы построенной модели.

В результате исследования разработана математическую модель кредитного скоринга для предсказания дефолта заемщиков.

Степень внедрения: на стадии разработки.

Область применения: страховые, кредитные организации, банки.

Экономическая эффективность/значимость работы: разработанная модель кредитного скоринга поможет проводить оценки платежеспособности клиентов и принимать более точные решения о предоставлении банком денежных средств клиентам.

В будущем планируется создание работа над улучшением эффективности предсказаний модели, разработка графического интерфейса для системы кредитного скоринга.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ, НОРМАТИВНЫЕ ССЫЛКИ

В данной выпускной квалификационной работе использовались следующие сокращения:

CHAID – Chi-square Automatic Interaction Detection

CART – Classification and Regression Tree

k-NN – k-nearest Neighbor’s algorithm

ROC – Receiver Operating Characteristic

AUC – Area Under ROC curve

ANOVA – Analysis of Variation

RFE – Recursive Feature Elimination

SWOT – Strengths, Weaknesses, Opportunities, Threats

ИСР – иерархическая структура работ

ВКР – выпускной квалификационная работа

НИР – научно-исследовательская работа

ГОСТ – государственный общесоюзный стандарт

НТИ – научно-техническое исследование

ФСС – фонд социального страхования

ПФ – пенсионный фонд

ПК – персональный компьютер

ПЭВМ – персональная электронно-вычислительная машина

СИЗ – средства индивидуальной защиты

СКЗ – средства коллективной защиты

ЭМП – электромагнитное поле

ФККО – федеральный классификационный каталог отходов

ЧС – чрезвычайная ситуация

Оглавление

1 Литературный обзор

1.1 Понятие кредитного скоринга

Кредитная оценка – один из важнейших процессов при принятии банками решений по управлению кредитными ресурсами. Этот процесс включает сбор, анализ и классификацию различных кредитных элементов и переменных для оценки кредитных решений. Качество банковских кредитов является ключевым фактором, определяющим конкуренцию, выживаемость и прибыльность банков. Одним из наиболее важных наборов для классификации клиентов банка в рамках процесса оценки кредитоспособности с целью снижения текущего и ожидаемого риска плохой кредитной истории клиента является *кредитный скоринг*.

Впервые скоринг как метод оценки заемщика был предложен американским экономистом Д. Дюраном в 40-е гг. XX в.

Авторы во всем мире дают различные определения кредитного скоринга.

Например, Сахабиева Г.А. определяет сущность скоринга в «определении совокупного кредитного балла потенциального клиента путем оценивания его по ряду критериев. Каждый критерий имеет свою весовую характеристику и в дальнейшем агрегируется в интегральный показатель – совокупный кредитный балл» [1].

Побединская Т.Д. дает следующее определение описание кредитного скоринга: «Скоринговая система оценки платежеспособности основана на математико-статистических методах. Главная цель такой системы – оценка на основе некоторых факторах платежеспособности и принятие решения о предоставлении денежных средств. Собирается как можно более полная информация о потенциальном заемщике, обычно используется анкета, заполнения потенциальным заемщиком, после чего скоринговая система обрабатывает имеющиеся данные, начисляя баллы по внутреннему алгоритму» [2].

Что же касается зарубежных авторов, то, например, Трипати Д. определил, что «Кредитный скоринг – это метод, используемый для прогнозирования принадлежности клиента к законной или подозрительной группе клиентов» [3].

Таким образом, обобщив все вышеперечисленные понятия, можно сказать, что кредитный скоринг можно определить как использование статистических моделей для преобразования соответствующих данных в числовые меры, которые определяют кредитные решения. Он представляет собой набор моделей принятия решений и лежащих в их основе методов, которые помогают кредиторам в предоставлении потребительских кредитов. Эти методы определяют, кто получит кредит, какой объем кредита они должны получить и какие операционные стратегии повысят прибыльность заемщиков для кредиторов.

Кредитный скоринг не ограничивается банками. Другие организации, такие как компании мобильной связи, страховые компании, арендодатели и государственные ведомства, используют те же методы. Компании цифрового финансирования, такие как онлайн-кредиторы, также используют альтернативные источники данных для расчета кредитоспособности заемщиков. Кроме того, широкие возможности вычислительной техники делают использование кредитного скоринга намного проще, чем раньше.

1.2 Математические модели скоринга

Выберем для изучения несколько наиболее часто используемых методов для построения математической модели скоринга.

Рассматриваемые статистические методы включают линейный дискриминантный анализ, логистическую регрессию и наивный байесовский анализ, методы машинного обучения включает метод ближайшего соседа, деревья решений, метод опорных векторов, случайные леса, нейронные сети.

Линейный дискриминантный анализ – один из наиболее популярных методов скоринга, и кредитного скоринга, в частности. Он основан на построении одной или нескольких линейных функций, включающих объясняющие переменные. Следовательно, общая модель задается [4]:

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \ , \qquad (1)$$

где Z – оценка дискриминации, α – перехват, β_i – коэффициент, отвечающий за линейный вклад i-ой объясняющей переменной X_i, где $i = 1, 2\ldots, p$.

Этот метод имеет следующие допущения: ковариационные матрицы каждого подмножества классификации равны; каждая классификационная группа подчиняется многомерному нормальному распределению. Другой возможный метод дискриминантного анализа – это не линейный, а квадратичный дискриминантный анализ [4].

Логистическая регрессия – это линейная модель, в которой целевая переменная является категориальной. Модель создается на основе сигмоидной функции, которая используется для оценки вероятности бинарного ответа (в нашем случае, если клиент, запрашивающий кредит, не выплатит свои долги) на основе входных данных. Это также один из наиболее широко используемых методов для кредитного скоринга [5].

Деревья решений – это метод классификации, который использует данные для построения так называемых правил принятия решений, организованных в древовидные архитектуры. В общем, цель этого метода – определить набор логических условий «если-то», которые позволяют прогнозировать или классифицировать случаи. Существует три применяемых алгоритма: детектор автоматического взаимодействия хи-квадрат (*CHAID*), дерево классификации и регрессии (*CART*) и *C*5, которые различаются критерием построения дерева. *CART* использует коэффициент Джини в качестве критерия расщепления, *C*5 использует энтропию, а *CHAID* использует критерий хи-квадрат. Другими возможными и конкретными методами деревьев решений являются алгоритм деревьев решений *C*4.5 и *J*4.8 [4].

Генетический алгоритм основан на аналогии с биологическим процессом естественного отбора. В сфере кредитования это выглядит следующим образом: имеется набор классификационных моделей, которые подвергаются «мутации», «скрещиваются», и в результате отбирается «сильнейший», т. е. модель, дающая наиболее точную классификацию.

k-NN – это непараметрический метод, это означает, что он не делает никаких предположений о базовом распределении данных. Это очень полезно, так как в реальном мире большинство практических данных не подчиняются типичным теоретическим предположениям.

k-NN – один из простейших алгоритмов классификации, который часто используется в качестве эталона для более сложных классификаторов. Фикс и Ходжес [6] представили метод классификации образов, который с тех пор стал известен как правило *k-ближайшего соседа*. Классификатор *k*-ближайших соседей обычно основан на евклидовом расстоянии между тестовой выборкой и заданной обучающей выборкой. Основная идея алгоритма *k-NN* заключается в том, что всякий раз, когда есть новая точка для прогнозирования, ее k ближайших соседей выбираются из обучающих данных. Тогда прогноз новой точки может быть средним из значений ее *k* ближайших соседей.

При использовании *метода ближайших соседей* выбирается единица измерения для определения расстояния между клиентами. Все клиенты в выборке получают определенное пространственное положение. Каждый новый клиент классифицируется исходя из того, каких клиентов – платежеспособных или нет – больше вокруг него.

Наивный алгоритм Байеса – это алгоритм классификации, основанный на правиле Байеса, который предполагает, что все атрибуты X_1 , X_2, … X_n условно независимы друг от друга при заданном Y. Ценность этого предположения состоит в том, что оно значительно упрощает представление $P(X|Y)$, и проблему его оценки по данным обучения. Байесовская сеть представляет собой совместное распределение вероятностей по набору

непрерывных входных данных X_i. Этот алгоритм широко применяется в системах кредитного скоринга [7].

Случайный лес – это совокупность деревьев классификации или регрессии, созданных из выборок начальной загрузки обучающих данных с использованием случайного выбора признаков в процессе индукции дерева. Стратегия случайных лесов состоит в том, чтобы случайным образом выбирать подмножества для выращивания деревьев, при этом каждое дерево выращивается на начальной выборке обучающего набора. Случайные леса обычно демонстрируют значительное улучшение производительности по сравнению с классификатором с одним деревом, таким как *C*4.5. Это дает частоту ошибок обобщения, которая выгодно отличается от *Adaboost*, но более устойчива к шуму. По сравнению с другими классификаторами ансамбля, случайные леса имеют следующие преимущества: уменьшение дисперсии, достигаемое за счет усреднения по обучаемым данным, и рандомизированные стадии, уменьшающие корреляцию между отдельными обучаемыми данными в ансамбле [8].

К числу *гибридных и комбинированных* относятся методы, в которых применяются различные техники кредитного скоринга для повышения эффективности. Наиболее употребительны три метода комбинирования: беггинг, бустинг и стекинг.

В беггинге применяют какое-либо определенное агрегирование набора предикторов, которые в совокупности дадут более совершенный предиктор. Его используют в случаях, когда основной алгоритм обучения неустойчив – сильно зависит от небольших изменений в обучающем множестве.

Бустинг предполагает формирование на основе менее точно алгоритма более сильный алгоритм классификации. Слабый алгоритм «доучивается» за счет того, что перераспределяются веса примеров из обучающей выборки.

При стекинге происходит комбинирование нескольких алгоритмов с помощью некоторого комбинатора. Как правило, в роли комбинатора

выступает логистическая регрессия. В кредитном скоринге применение комбинированных методов широко распространено [9].

1.3 Логистическая регрессия. Основные понятия и преимущества

Логистическая регрессия – это алгоритм классификации, используемый для назначения наблюдений дискретному набору классов. В отличие от линейной регрессии, которая выводит непрерывные числовые значения, логистическая регрессия преобразует свои выходные данные с использованием функции логистической сигмоиды, чтобы вернуть значение вероятности, которое затем может быть сопоставлено двум или более дискретным классам.

Чтобы определить вероятность предсказанных значений используется сигмовидная функция (2). Функция отображает любое действительное значение в другое значение между 0 и 1 (рисунок 1):

$$S(z) = \frac{1}{1 + e^{-z}}, \tag{2}$$

где $S(z)$ – оценка вероятности от 0 до 1; z – вход в функцию, прогноз модели; e – основание натурального логарифма [7].

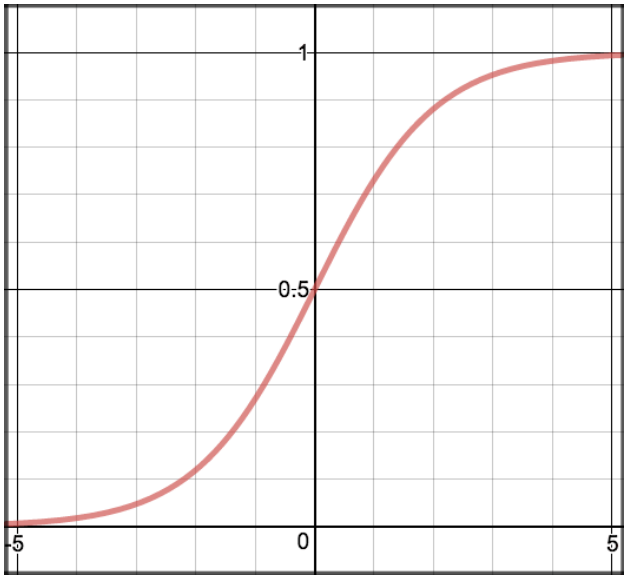


Рисунок 1 – График сигмовидной функции (2)

Входные значения x объединяются линейно с использованием весов или значений коэффициентов $B = (\beta_1, \beta_2, ..., \beta_n)$ для прогнозирования выходного значения y. Ключевым отличием от линейной регрессии является то, что моделируемое выходное значение представляет собой двоичные значения (0 или 1), а не числовое значение.

Ниже приведен пример уравнения логистической регрессии:

$$y = \frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}}, \quad (3)$$

где y – прогнозируемый выходной сигнал, β_0 – смещение или член перехвата, а β_1 – коэффициент для входного значения x. Каждый столбец входных данных имеет ассоциированный коэффициент β (постоянное действительное значение), который необходимо извлечь из тренировочных данных.

Фактическим представлением модели логистической регрессии являются коэффициенты в уравнении (рисунок 2).

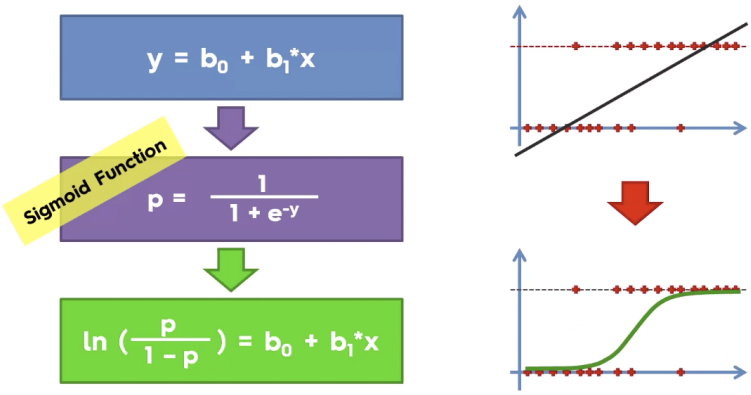


Рисунок 2 – Процесс преобразования логистической функции и вероятностей

Таким образом, функция прогнозирования возвращает оценку вероятности в диапазоне от 0 до 1. Чтобы отобразить ее в дискретный класс, выбирается пороговое значение или точка перелома, выше которой будет классифицировать значения в класс 1 и ниже которого мы классифицируем значения в класс 2, чаще всего это значение $p = 0{,}5$ [10].

Коэффициенты алгоритма логистической регрессии должны оцениваться на основе тренировочных данных. Это делается с использованием оценки максимального правдоподобия. Оценка максимального правдоподобия – это алгоритм обучения, используемый различными алгоритмами машинного обучения, хотя он и делает предположения о распределении данных.

Наилучшие коэффициенты привели бы к модели, которая предсказывала бы значение, очень близкое к 1 для класса по умолчанию, и значение, очень близкое к 0 для другого класса. Прогнозирование для максимального правдоподобия для логистической регрессии состоит в том, что процедура поиска ищет значения для коэффициентов (бета-значений), которые сводят к минимуму ошибку в вероятностях, предсказываемых моделью, и вероятностей в данных [11].

Модели логистической регрессии удобочитаемы, т.е. каждую функцию и ее коэффициент можно анализировать индивидуально, тем самым определяя, насколько важно различать кредитоспособных и некредитоспособных клиентов. Чтобы извлечь вероятности запроса на кредит, к коэффициентам регрессионной модели применяется логит-функция [5].

Основные преимущества метода логистической регрессии были определены с точки зрения менее ограничительных допущений при моделировании. Линейность, условия нормальности, а также независимость между независимыми переменными не предполагаются в подходе с использованием логистической регрессии, что оставляет больше гибкости в работе с реальными данными. Первые сообщенные результаты прогнозирования методом логистической регрессии имели меньшую предсказательную силу, чем те, о которых сообщалось в исследованиях о дискриминантном анализе. Но позже исследования показали, что логистическая регрессия – это надежный и мощный статистический подход для моделирования кредитного риска [12].

1.4 Показатели эффективности модели кредитного скоринга

После построения математической модели и проведения прогнозирования необходимо оценить полученный результат с точки зрения эффективности.

Результат оценки кредитного риска может быть представлен в виде двухмерной матрицы ошибок, которая показана в таблице 1.

Таблица 1 – Матрица ошибок предсказанных результатов

Прогноз	Эмпирические данные	
	+	–
+	*TP*	*FP*
–	*FN*	*TN*

В таблице используются обозначения *TP – True Positive*, *TN – True Negative*, *FP – False Positive*, *FN – False Negative*, которые отображают количество соответственно правильно спрогнозированных позитивных решений о кредитовании, правильно спрогнозированных отказов в кредитовании и неверных ложноотрицательных и ложноположительных решений.

Матрица ошибок является универсальным аналитическим средством для оценки эффективности прогноза. На её основе могут быть вычислены различные метрики качества алгоритма:

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ – доля правильных прогнозов от общего числа.

$Precision = \frac{TP}{TP + FP}$ – точность, отображающая долю истинно-положительных результатов, которые были определены моделью как положительные.

$Recall = \frac{TP}{TP + FN}$ – полнота, которая показывает долю верно предсказанных положительных результатов.

$F-score=2\frac{(Precision\cdot Recll)}{Precision+Recll}$ – *F*-мера, которая позволяет найти гармоническое среднее значение между точностью и полнотой.

$TPR=\frac{TP}{TP+FN}$ – показатель *True Positive Rate* обозначает долю от общего количества носителей признака, верно классифицированных как несущие признак, также этот параметр называется чувствительностью алгоритма классификации.

$FPR=\frac{FP}{FP+TN}$ – показатель *False Positive Rate* обозначает долю объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак.

Также одним из популярных метрик оценки качества бинарной классификации является *ROC*-кривая, которая показывает соотношение между показателями *TPR* и *FPR* и её интерпретация с помощью площади под этим графиком – показателя *AUC*, причем чем его больше значение, тем лучше модель [13].

Статистика Колмогорова – Смирнова (*KS*) указывает максимальное расстояние между кумулятивной функцией распределения вероятностей, полученной клиентами, которые полностью выплатили свои долги, и теми, кто не выполняет свои обязательства. Нулевое значение для статистики KS указывает на то, что кредитный рейтинг не позволяет различать добросовестных плательщиков и неплательщиков, а значение, равное 100, указывает на то, что кредитный скоринг работает с максимальной эффективностью [5].

1.5 Программное обеспечение для реализации алгоритмов скоринга

Для программной реализации алгоритмов скоринга широко используются статистические программные пакеты *SAS*, *SPSS*, *Statistica*. Но

из-за ограниченных возможностей данных коммерческих пакетов в последние годы на их смену пришло использование открытых сред программирования таких как *Python*, *R*. В них возможна реализация большего числа алгоритмов, можно обновлять и совершенствовать уже существующие математические модели, использование множества библиотек существенно расширяет возможности скоринга, а также повышается скорость обработки большого объема данных

Но если в таких сферах, как здравоохранение и промышленность, с распространением *Python* и *R* от применения коммерческих пакетов практически отказались, то в банках к решениям *SAS* пока обращаются чаще, чем к *Python* и *R*.

Язык *R* создавался как специальное средство для статистических вычислений, он стал первой открытой средой, которая начала активно использоваться для анализа данных.

Язык *Python* стал самым популярным средством для анализа данных после выхода документированной библиотеки *scikit-learn*, в которой реализовано большое количество алгоритмов машинного обучения. Кроме *scikit-learn*, популярны также библиотеки *Pandas*, *NumPy*, *TensorFlow*, *Theano* и другие. Основное преимущество *Python* перед *R* – более высокая скорость выполнения скриптов [9].

2 Практическая часть

2.1 Подготовка и интерпретация данных

Для построения модели кредитного скоринга необходимо иметь качественные данные о заемщиках, поскольку точность модели зависит от подобранных данных для её обучения. Для прогнозирования различных областей кредитования необходимы соответствующие данные, например, для потребительских кредитов и бизнес-кредитов модели будут различаться.

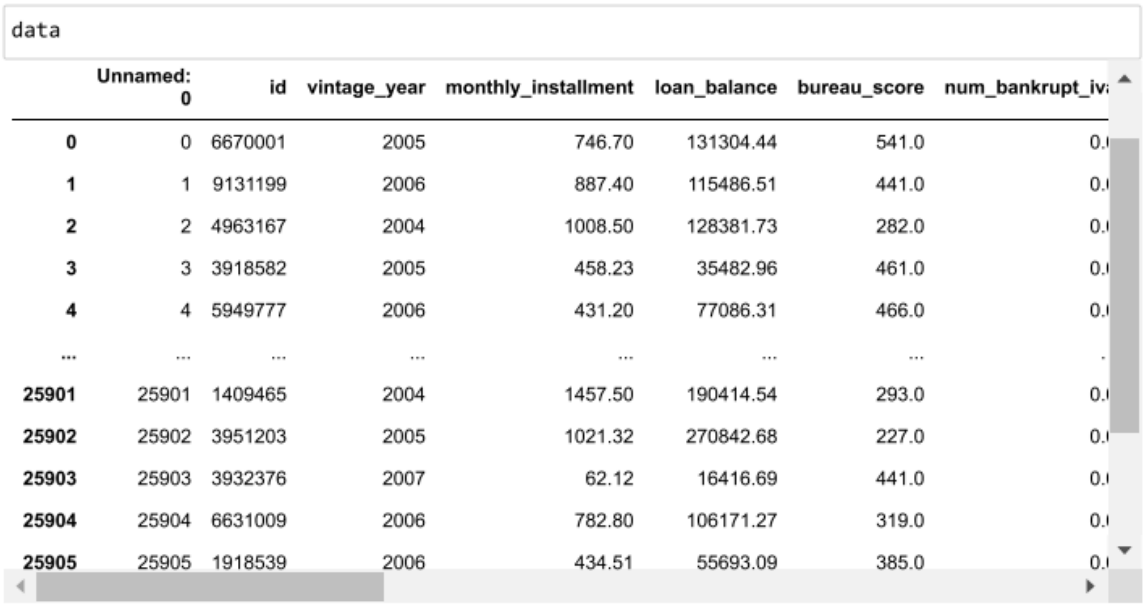
Для последующего построения модели и оценки её качества важно учитывать соотношение количество данных, соответствующих дефолту клиентов и успешной платежеспособности. Для правильной работы модели необходимо иметь в данных примерно одинаковое количество как дефолтных заявок, так и успешных. Для исследования крайне важно определять именно дефолты кредитов, поэтому таких примеров в данных должно быть большое количество.

В данной работе используется набор обезличенных данных, представленных компанией ООО «Эко-Томск». В них имеются данные розничного кредитного портфеля банка из 25 906 наблюдений по заемщикам.

В исследуемых данных из общего количества только 1362 записи имеют значение дефолт. В процентном соотношении имеем 94,7 % добросовестных заемщиков и 5,3 % не выплативших кредит.

Код программы, разработанный в рамках данной работы приведен в Приложении А.

Вид исходных данных, полученный в интерактивной оболочке *Jupyter Notebook*, приведен на рисунке 3.



data

	Unnamed: 0	id	vintage_year	monthly_installment	loan_balance	bureau_score	num_bankrupt_iv
0	0	6670001	2005	746.70	131304.44	541.0	0.
1	1	9131199	2006	887.40	115486.51	441.0	0.
2	2	4963167	2004	1008.50	128381.73	282.0	0.
3	3	3918582	2005	458.23	35482.96	461.0	0.
4	4	5949777	2006	431.20	77086.31	466.0	0.
...	...	...	...	...	...	...	.
25901	25901	1409465	2004	1457.50	190414.54	293.0	0.
25902	25902	3951203	2005	1021.32	270842.68	227.0	0.
25903	25903	3932376	2007	62.12	16416.69	441.0	0.
25904	25904	6631009	2006	782.80	106171.27	319.0	0.
25905	25905	1918539	2006	434.51	55693.09	385.0	0.

Рисунок 3 – Вид исходных данных

Каждый заемщик имеется идентификационный номер и характеризуется различными параметрами, описание доступных параметров приведено в таблице 2. Всего имеется 42 переменных.

Таблица 2 – Описание имеющихся параметров

№	Обозначение в исходных данных	Описание
1	*id*	Идентификатор
2	*vintage_year*	Год открытия счета
3	*monthly_installment*	Ежемесячный взнос
4	*loan_balance*	Остаток кредита
5	*bureau_score*	Оценка заемщика внешним бюро
6	*num_bankrupt_iva*	Количество случаев несостоятельности
7	*time_since_bankrupt*	Время с момента последней просрочки платежа
8	*num_ccj*	Количество обращений в суд
9	*time_since_ccj*	Время с момента обращения в суд
10	*ccj_amount*	-
11	*num_bankrupt*	Количество просрочек платежа
12	*num_iva*	-
13	*min_months_since_bankrupt*	Количество месяцев с момента последней просрочки платежа
14	*pl_flag*	-
15	*region*	Регион
16	*ltv*	Отношение суммы кредита к рыночной стоимости залога
17	*arrears_months*	Просроченные месяцы
18	*origination_date*	Дата выпуска новых обязательств
19	*maturity_date*	Срок погашения

Продолжение таблицы 2

20	*repayment_type*	Тип погашения
21	*arrears_status*	Статус просроченной задолженности
22	*arrears_segment*	Сегмент просроченной задолженности
23	*mob*	Количество месяцев, прошедших с даты выдачи кредита
24	*remaining_mat*	Оставшийся срок погашения
25	*loan_term*	Кредитные условия
26	*live_status*	Состояние кредита
27	*repaid_status*	-
28	*month*	Месяц
29	*worst_arrears_status*	Статус худшей просроченной задолженности
30	*max_arrears_12m*	Максимальное количество месяцев просроченной задолженности за последние 12 месяцев
31	*recent_arrears_date*	Дата последней просроченной задолженности
32	*months_since_2mia*	Месяцев с двух месяцев просрочки
33	*avg_mia_6m*	Ср. месяцев просрочки
34	*max_arrears_bal_6m*	Максимальный остаток по счету просроченной задолженности за 6 последние месяцев
35	*max_mia_6m*	-
36	*avg_bal_6m*	Средний остаток по счету за 6 последние месяцев
37	*avg_bureau_score_6m*	Средняя оценка бюро за 6 последних месяцев
38	*cc_util*	Использование кредитной карты
39	*annual_income*	Годовой доход
40	*emp_length*	Стаж работы
41	*months_since_recent_cc_delinq*	Количество месяцев с момента последней просрочки по кредитной карте
42	*default_flag*	Флаг дефолта заёмщика

Все вышеперечисленные признаки необходимы для построения модели и будут являться предикторами – прогностическими параметрами, средствами прогнозирования. Большая часть описания переменных в таблице 2 было восстановлена в ходе работы из открытых источников, смысловое значение пяти переменных (10, 14, 27, 33 и 35) восстановить не удалось.

2.1.1 Выбор зависимой и независимой переменных модели

Целью кредитного скоринга является определение недобросовестных заёмщиков, поэтому зависимой переменной принимается категориальная переменная величина с двумя категориями «добросовестный» и

«недобросовестный». В нашем наборе данных она обозначена как *default_flag* и принимает значение «0» и «1».

Независимыми (скоринговыми) переменными для физического лица могут выступать личная информация, финансовые показатели, кредитная история и т.д., получаемые из анкет или ранее имеющихся данных. В данной работе будет использоваться предоставленная выборкой информация по 42 предикторам, представленным в таблице 2.

2.1.2 Улучшение качества данных

Одним из важных показателей для дальнейшего анализа является количество пропущенных данных. Анализ и предобработка данных была осуществлена в среде *Jupyter Notebook Python,* с использованием библиотеки Sweetviz.

Анализ показал, что некоторые признаки имеют только одно значение – *live_status*, *repaid_status*, *month* и не могут учитываться в модели, они были удалены. Также были исключены признаки с количеством пропущенных значений более 97% – это *resent_arrears_data*, *months_since_2mia*.

Признаки *origination_date* и *maturity_date* были удалены, так как они имеют формат даты (*datatime*) и являются неинформативными, необрабатываемыми и не могут быть предикторами модели.

Так как в оставшемся наборе данных среди признаков пропущенных значений очень мало – в каждом из предикторов не более 2%, были удалены все строки с пропущенными значениями с помощью функции *dropna*().

При помощи функции *info*() определим тип переменной и количество значений. Предикторы *region* и *repayment_type* имеют тип данных – object, поэтому их необходимо перевести в целочисленных формат, закодировав буквенные значения в цифровые.

Для построения адекватной модели и проверки ее точности исходные данные необходимо подготовить путем удаления дублирующих записей и

обработку выбросов. Данное очищение данных позволит значительно улучшить качество модели. Была проведена проверка на повторяющиеся данные, таковых не оказалось.

Sweetviz – это библиотека Python с открытым исходным кодом, которая генерирует визуализации для исследовательского анализа данных, выходом данной функции является полностью автономная HTML-страница. Система построена на быстрой визуализации целевых значений и сравнении наборов данных. Его цель – помочь в быстром анализе целевых характеристик, данных обучения и тестирования и других подобных задач по характеристике данных.

При помощи пакета *Sweetviz* визуализировали исходные данные, сравнили все предикторы относительно переменной отклика – *default_flag*, визуально посмотрели зависимости между переменными, там самым подготовившись к следующему этапу – выбору предикторов для построения модели.

2.1.3 Выбор признаков

Для выбора значимых признаков, необходимых для построения модели, был проведен однофакторный дисперсионный анализ. Для вычисления *F*-статистики был использован метод *SelectKBest* из пакета *Feature_selection* библиотеки *sklearn*. При установлении в данном методе параметра используемой функции для оценки *f_classif* был получен список предикторов с их значением *F*-статистики и отсортирован в порядке убывания – результат представлен на рисунке 4.

	col	score	p-value
16	arrears_months	4719.892815	0.000000e+00
19	arrears_segment	3466.467190	0.000000e+00
27	max_arrears_12m	3249.738623	0.000000e+00
18	arrears_status	2872.208073	0.000000e+00
33	cc_util	2821.255324	0.000000e+00
26	worst_arrears_status	2254.000417	0.000000e+00
30	max_mia_6m	2196.935770	0.000000e+00
28	avg_mia_6m	1399.307336	3.466409e-298
4	bureau_score	771.029073	3.329029e-167
32	avg_bureau_score_6m	748.188463	2.211788e-162
34	annual_income	567.066064	5.538933e-124
36	months_since_recent_cc_delinq	352.306790	4.489532e-78
7	num_ccj	283.504478	2.847868e-63
35	emp_length	246.084060	3.363755e-55
8	time_since_ccj	124.927187	6.165909e-29
22	loan_term	105.744007	9.380414e-25
21	remaining_mat	94.952586	2.133971e-22
15	ltv	25.847097	3.721765e-07
5	num_bankrupt_iva	22.905085	1.711551e-06
17	repayment_type	22.751773	1.853541e-06
3	loan_balance	20.814733	5.082822e-06
11	num_iva	20.343890	6.498814e-06
31	avg_bal_6m	20.061848	7.530340e-06
6	time_since_bankrupt	17.616717	2.711189e-05
29	max_arrears_bal_6m	17.036399	3.678675e-05
2	monthly_installment	8.960703	2.761135e-03
9	ccj_amount	4.080590	4.338896e-02

Рисунок 4 – Сортировка предикторов по *F*-статистикам

Критическое значение *F* определяется уровнем значимости, поэтому для отбора признаков был определен уровень *p-value* = 0,05 и отобраны только те признаки, которые принимают значение меньше 0,05.

На рисунке 5 представлен график значений *F*-статистики для предикторов.

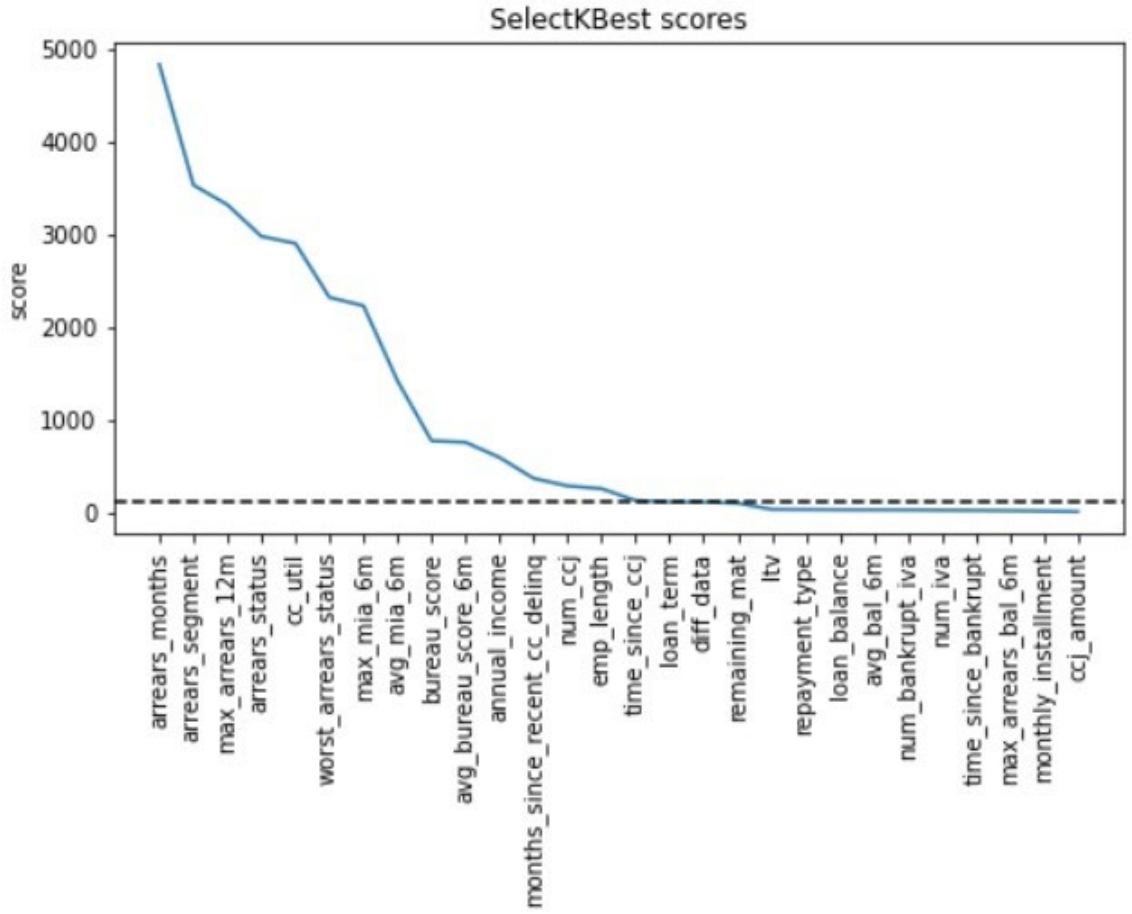


Рисунок 5 – Значения *F*-статистики признаков

Таким образом, в выборку были включены 16 предикторов с *p-value* < 0,05 и значением *F*-статистики больше 100.

Далее для оценки силы связи между признаками проводится корреляционный анализ, по результатам которого будут исключены взаимосвязанные предикторы. Это сделано потому, что существование мультиколлинеарности факторов затрудняет оценку весов каждого параметра модели, что в свою очередь приводит к некорректности самой модели.

Для оценки корреляции признаков была составлена матрица парных корреляций – рисунок 6.

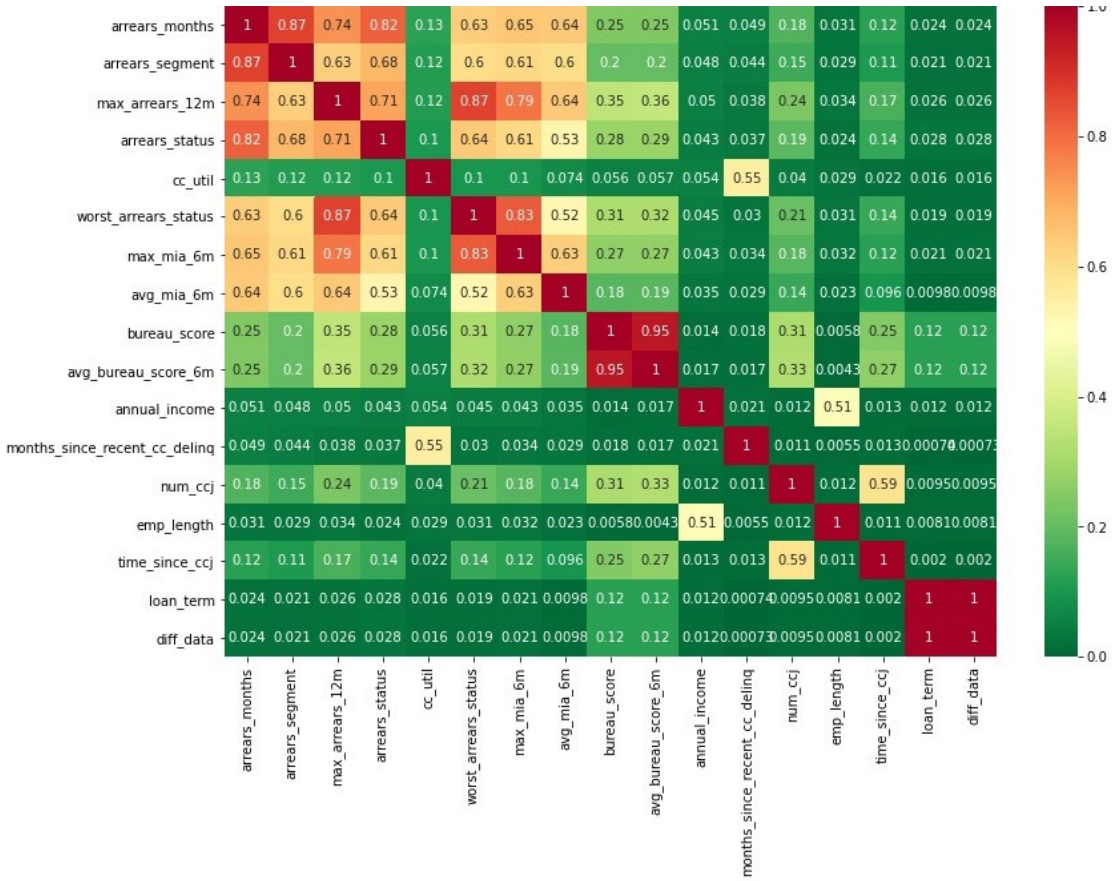


Рисунок 6 – Матрица парных корреляций

Для каждой пары факторов с высокой корреляцией (выбран уровень выше 0,7) был исключен только один предиктор с наименьшим из них значением *F*-статистики. Таким образом, были удалены из списка предикторов следующие признаки: *max_mia_6m*, *avg_bureau_score_6m*, *worst_arrears_status*, *arrears_segment*, *arrears_status*, *max_arrears_12m*.

Повторное построение матрицы корреляций показало, что, как видно из рисунка 7, коррелирующие признаки отсутствуют.

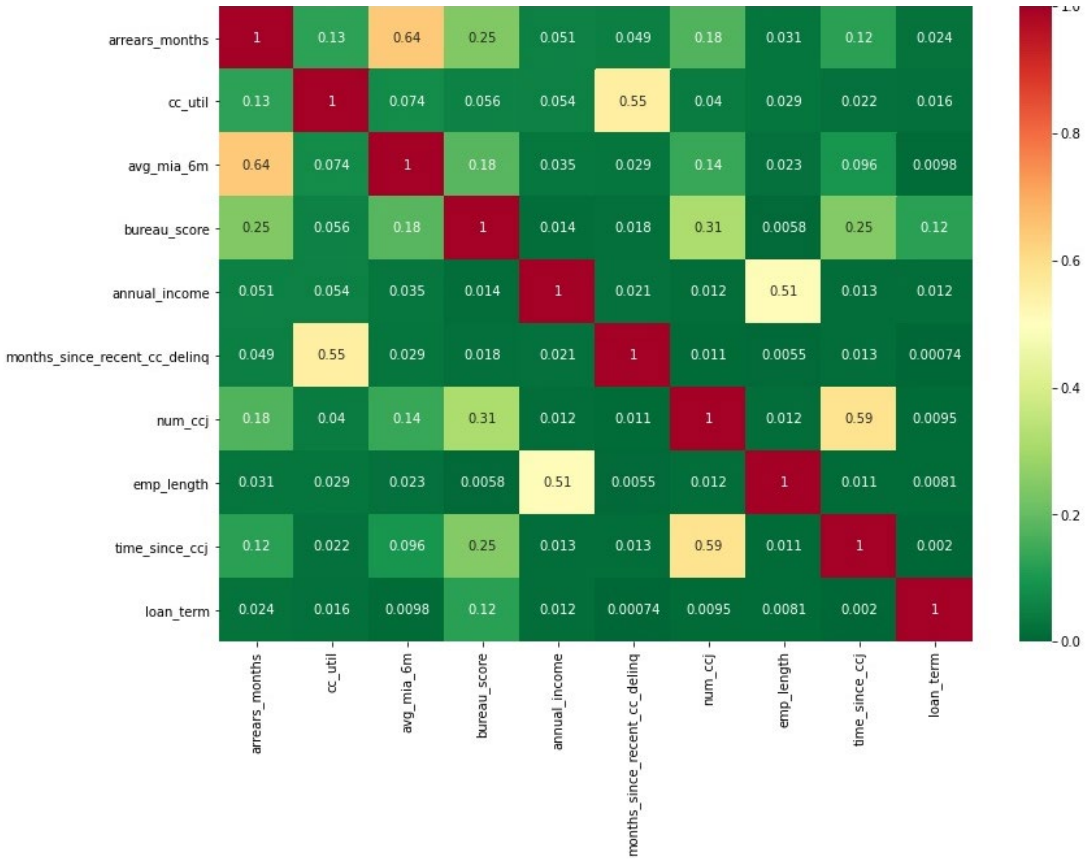


Рисунок 7 – Матрица парных корреляций после исключения признаков

Дальнейшим этапом подготовки окончательного списка признаков являлось использованием метода ранжирования с рекурсивным устранением объектов (*Recursive Feature Elimination*, *RFE*). В данном методе используется рекурсивное рассмотрение всё меньших и меньших наборов данных, которое осуществляется путем обучения какого-либо оценщика (в данном случае была выбрана модель логистической регрессии) и отсечения наименее важных признаков на каждом этапе рекурсии. В итоге может быть получено необходимое количество признаков. Данный метод был программно реализован с помощью одноименного класса из библиотеки *sklearn* пакета *Feature_selection*.

После применения метода был сформирован окончательный список предикторов для модели: *arrears_months*, *cc_util*, *avg_mia_6m*, *num_ccj*, *loan_term*, *emp_length*, то есть значимыми признаками для построения модели кредитного скоринга по исходным данным являются количество обращений в

суд, просроченные месяцы, кредитные условия, использование кредитной карты и стаж работы заемщиков.

- *arrears_months* – просроченные месяцы, №17
- *cc_util* – использование кредитной карты, №38
- *avg_mia_6m* – ср. с 6ого месяца просрочки, №33
- *num_ccj* – количество обращений в суд, №8
- *loan_term* – кредитные условия, №25
- *emp_length* – стаж работы, №40

2.2 Разработка математической модели кредитного скоринга

2.2.1 Построение моделей

Для построения модели логистической регрессии в Python существует различных библиотеки, в частности:

statsmodels – чаще используется для статистических целей, так как обладает широким набором статистических тестов

sklearn – чаще используется для реализации методов машинного обучения. Направлена на решение задач классификации, включает наиболее распространенные метрики, не отличается большим набором статистических тестов.

Обе библиотеки равноправны, дают одинаковый результат и выбор библиотеки производится с учетом целей пользователя.

Первым шагом для построения модели является разбиение выборки на тестовую и тренировочную в случайном порядке. Для этого воспользуемся функцией *train_test_split*(). Данная функция принимает следующие аргументы: *X* – датафрейм независимых переменных, *y* – зависимая переменная, *test_size* – размер тестовой выборки (в %), *random_state* – начальное значение генератора случайных чисел (необходимо для воспроизведения результата).

Результатом работы функции являются: *X_train* – датафрейм независимых переменных тренировочная выборка, *X_test* – датафрейм независимых переменных тестовая выборка, *y_train* – зависимая переменная тренировочная выборка, *y_test* – зависимая переменная тестовая выборка.

Таким образом, данные были разделены с помощью функции *train_test_split*() в соотношении 70% и 30% для тестовой и тренировочной выборок соответственно.

Далее было проверено распределение данных тренировочной выборки – в ней 16 883 значения *default_flag* = 0 и 942 значений *default_flag* = 1, следовательно, доля дефолтов в тренировочной выборке 5,28%.

В тестовой выборке 7 246 значений *default_flag* = 0, *default_flag* = 1 – 394, что в процентном соотношении составляет 5,16%.

Следовательно, дефолты представлены равномерно в тестовой и тренировочной выборке.

Теперь воспользуемся классификаторами из пакета *sklearn* для построения моделей.

Независимо друг от друга были построены шесть моделей:

1. K-ближайших соседей (*K-Nearest Neighbors*)
2. Метод опорных векторов (*Support Vector Machines*)
3. Логистическая регрессия (*Logistic Regression*)
4. Стохастический градиентный спуск (*SGD Classifier*)
5. Наивный байесовский классификатор (*Gaussian Naive Bayes*)
6. Дерево решений (*Decision Tree*)

Для каждой из них был обучен классификатор, проведена оценка параметров модели и спрогнозировано отношение к классу 0 или 1 с помощью функции *predict*(). При помощи функции *predict_prob*() были выведены прогнозные вероятности отношения к классу 0 и 1.

2.2.2 Оценка эффективности моделей

На тестовой выборке была проведена проверка работы полученных моделей, на рисунке 8 показаны матрицы ошибок при выполнении прогнозирования.

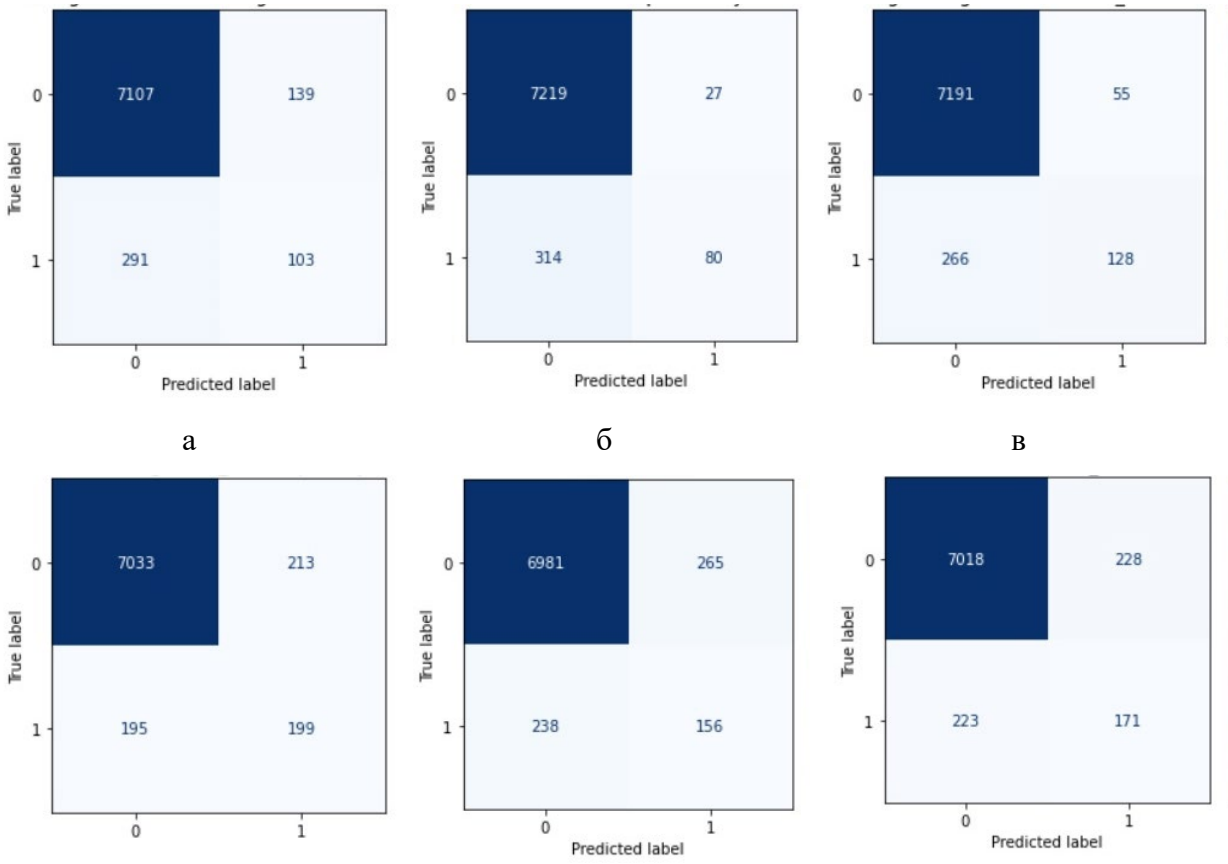


Рисунок 8 – Матрицы ошибок построенных моделей а) *k*-ближайших соседей, б) метод опорных векторов, в) логистическая регрессия, г) стохастический градиентный спуск, д) наивный байесовский классификатор, е) дерево решений

Матрица ошибок показывает нам, что все модели отлично предсказывают добросовестных заемщиков, но в случаях с дефолтами ситуация хуже.

Для понимания эффективности моделей необходимо определить метрики качества классификации, они приведены в таблице 3.

Таблица 3 – Основные метрики качества классификации

Модель	*Precision* 0	*Recall* 0	*F*-мера 0	*Precision* 1	*Recall* 1	*F*-мера 1	*Accuracy*
K-ближайших соседей	0.961	0.981	0.971	0.426	0.261	0.324	0.944
Метод опорных векторов	0.958	0.996	0.977	**0.748**	0.203	0.319	0.955
Логистическая регрессия	0.964	0.992	0.978	**0.699**	0.325	0.444	0.958
Стохастический градиентный спуск	0.973	0.971	0.972	0.483	0.505	0.494	0.947
Наивный байесовский классификатор	0.967	0.963	0.965	0.371	0.396	0.383	0.934
Дерево решений	0.969	0.969	0.969	0.429	0.434	0.431	0.941

Все модели точно предсказывают кредитоспособных заемщиков, ни один из показателей *Precision*, *Recall* и *F*-мера для определения этого класса не опускается ниже 0,95.

Для определения некредитоспособных клиентов по показателю *Precision*, который отвечает за долю объектов, действительно принадлежащих данному классу относительно всех объектов, которые система отнесла к этому классу, наилучший результат показала модель метода опорных векторов – показатель равен 0,748, немного хуже сработала модель логистической регрессии – 0,699 (выделено жирным в таблице 3).

Наибольшую метрику, которая говорит о доле истинно положительных классификаций, *Recall* имеет модель с использованием метода стохастического градиентного спуска, схожие значения показали практически все модели, за исключением *k*-ближайших соседей и метода опорных векторов, у которых данный показатель ниже почти на 20%.

F-мера, которая является гармоническим средним между точностью *Precision* и полнотой *Recall*, показывает оптимальный баланс между этими

двумя метриками. И наилучшее значение этого показателя имеют модели стохастического градиентного спуска и логистической регрессии.

Показатель *Accuracy*, отвечающий за долю правильных ответов алгоритма, нецелесообразно использовать при неравных классах, поэтому его анализ проводиться не будет, все модели имеют примерно одинаковый показатель.

Таким образом, более высокими значениями данных показателей эффективности отличаются модели стохастического градиентного спуска и логистической регрессии.

Следующей метрикой для оценки качества модели в машинном обучении используют *ROC*-кривую, отражающую соотношение долей верно найденных несущих признак исходов и неверно найденных не несущих признак исходов. По-другому *ROC*-кривая называется кривой ошибок, чем ближе кривая к левому верхнему углу графика, тем лучше предсказательная способность модели (рисунок 9).

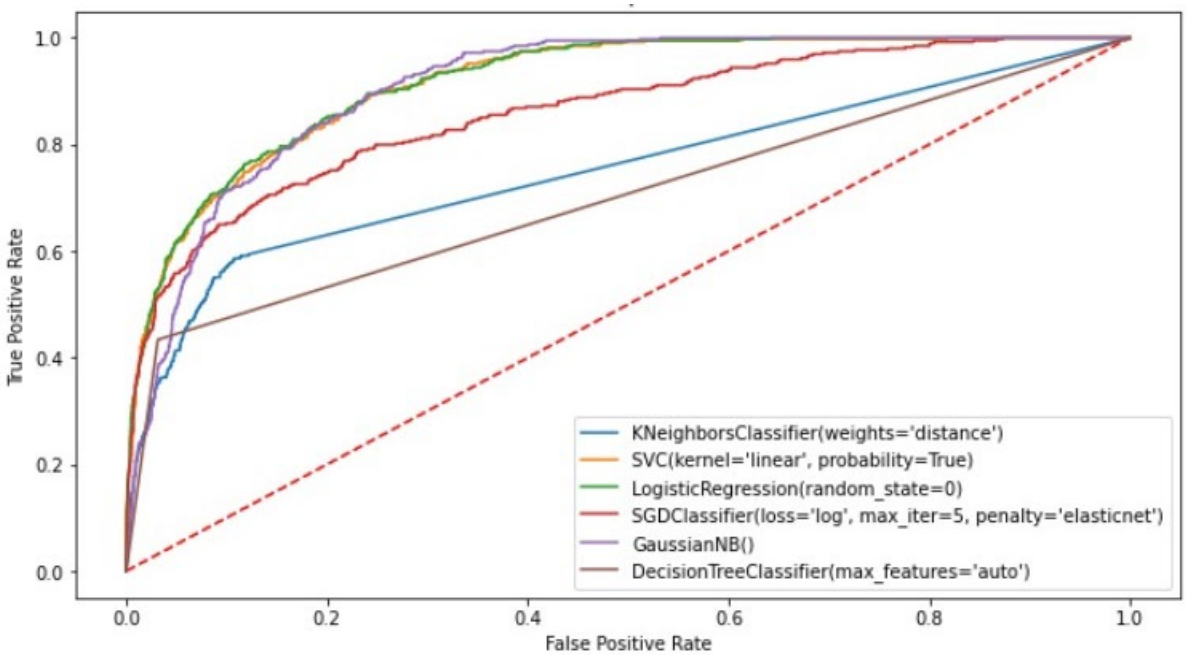


Рисунок 9 – *ROC*-кривая всех моделей

Площадь под кривой ошибок — показатель *AUC* – отражает качество классификации модели. Чем больше значение *AUC*, тем лучше предсказывает модель. Также по значению *AUC* вычисляется такой показатель, как индекс

Джини. Этот показатель переводит значение площади под кривой в диапазон от 0 до 1 и вычисляется по формуле $G = 2(AUC - 0.5)$. В таблице 4 приведены данные рассчитанные показатели для всех построенных моделей.

Таблица 4 – Коэффициенты *AUC* и Джини для построенных моделей

Модель	***AUC***	**Коэффициент Джини**
K-ближайших соседей	0.752	0.504
Метод опорных векторов	0.915	0.831
Логистическая регрессия	0.917	0.835
Стохастический градиентный спуск	0.859	0.718
Наивный байесовский классификатор	0.909	0.818
Дерево решений	0.701	0.402

Наилучшими показателями в соответствии с *ROC*-кривой обладает модель логистической регрессии, метод опорных векторов и наивный байесовский классификатор, но последние имеют более низкие предыдущие метрики, поэтому модель логистической регрессии обладает наиболее стабильно высокими параметрами прогнозирования по сравнению с другими одиночными моделями.

2.2.3 Построение ансамблевых моделей

Ансамблевое моделирование сосредоточено на сборе информации нескольких классификаторов, обученных для решения одной и той же проблемы, с использованием их мнений для принятия эффективного и точного решения [14]. Ошибка и отклонение одного классификатора в моделях ансамблевой классификации компенсируются другими членами ансамбля, поэтому возможности модели ансамбля обычно намного выше, чем у одного классификатора [15].

Существует такой ансамблевый алгоритм машинного обучения, как голосование – *voting*. Для задачи классификации класс, предсказываемый каждой моделью, можно рассматривать как голос, а класс, который получает большинство голосов, является ответом модели ансамбля (это называется

мажоритарным голосованием – *hard voting*). При мягком голосовании (*soft voting*) берутся вероятности каждого класса, предсказываемые всеми моделями, усредняются эти вероятности и определяется класс с самой высокой средней вероятностью, что и послужит ответом модели ансамбля [16].

Bagging (*Bootstrap Aggregating*) – широко используемый алгоритм ансамблевого обучения в машинном обучении. Алгоритм строит несколько моделей из случайно выбранных подмножеств набора тренировочных данных, объединяет обучаемые модели для создания более сильной, итоговые ответы по классам по каждой модели усредняются. *Bagging* уменьшает дисперсию и помогает избежать переобучения. Обычно он используется с методами обучения на основе деревьев решений, но также его можно применять с любым методом.

Случайные леса (*Random forests*) – это метод обучения ансамбля для классификации, регрессии и других задач, который работает путем построения множества деревьев решений во время обучения. Для задач классификации выходом случайного леса является класс, выбранный большинством деревьев. Леса случайных решений корректируют особенность деревьев решений переобучаться в соответствии их обучающей выборке.

AdaBoost, сокращение от *Adaptive Boosting*, представляет собой метаалгоритм статистической классификации, который необходимо использовать вместе со многими другими типами алгоритмов обучения для повышения производительности. Выходные данные других алгоритмов обучения («слабые ученики») объединяются во взвешенную сумму, которая представляет собой окончательный результат усиленного классификатора. *AdaBoost* адаптивен в том смысле, что последующие слабые ученики настраиваются в пользу тех ответов, которые были неправильно классифицированы предыдущими классификаторами. В некоторых задачах он может быть менее подвержен проблеме переобучения, чем другие алгоритмы обучения.

Gradient Boosting – это поэтапная аддитивная модель, в которой ее итерации можно рассматривать как минимизацию наискорейшего спуска с учетом данной функции потерь. *Gradient Boosting* обучает множество моделей постепенно, аддитивно и последовательно. Основное различие между *AdaBoost* и *Gradient Boosting* заключается в том, как эти два алгоритма выявляют недостатки слабых учеников (например, деревья решений). В то время как модель *AdaBoost* выявляет недостатки с помощью точек данных с большим весом, повышение градиента выполняет то же самое, используя градиенты в функции потерь – это мера, показывающая, насколько хорошо коэффициенты модели соответствуют базовым данным [17].

Для сравнения работы моделей были построены перечисленные выше ансамблевые модели и рассчитана их эффективность.

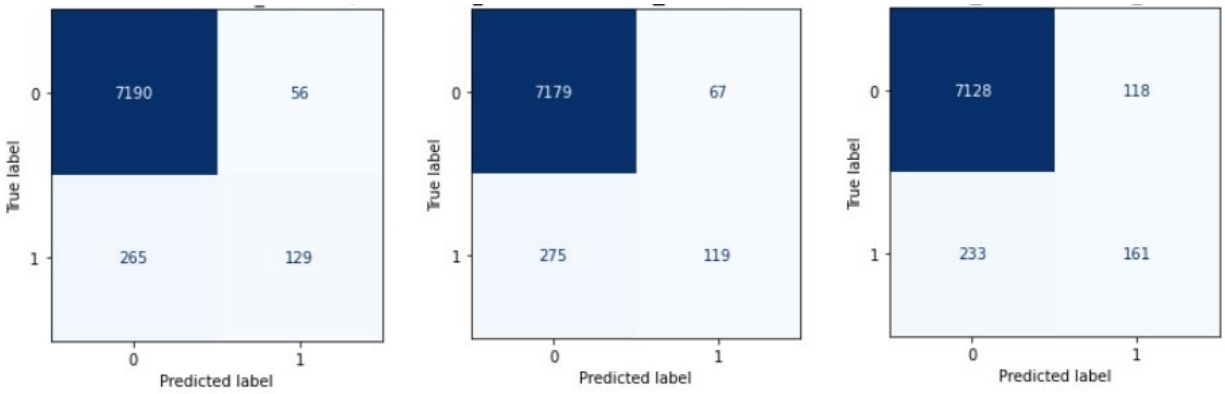
На тех же тренировочной и тестовой выборке были обучены и проверены модели с помощью функций, содержащихся в модуле *ensemble* библиотеки *sklearn*.

Для *Bootstrap Aggregating* была использована функция *BaggingClassifier*, для организации случайных подмножеств набора данных в ней были опробованы методы стохастического градиентного спуска, логистической регрессии и деревьев решений.

Для других моделей применялись соответствующие функции – *RandomForestClassifier*, *AdaBoostClassifier*, *GradientBoostingClassifier*, *VotingClassifier*. С помощью последней были рассмотрены различные комбинации моделей с применением мягкого голосования (*soft voting*).

2.2.4 Оценка эффективности ансамблевых моделей

На рисунке 10 показаны матрицы ошибок при выполнении прогнозирования моделями *BaggingClassifier*.



а б в

Рисунок 10 – Матрицы ошибок построенных моделей *Bagging* а) на базе логистической регрессии, б) на базе стохастического градиентного спуска, в) на базе деревьев решений

По результатам матрицы ошибок были вычислены различные метрики и построена *ROC*-кривая (рисунок 11), по которой были вычислены показатели *AUC* и Джини (таблица 5).

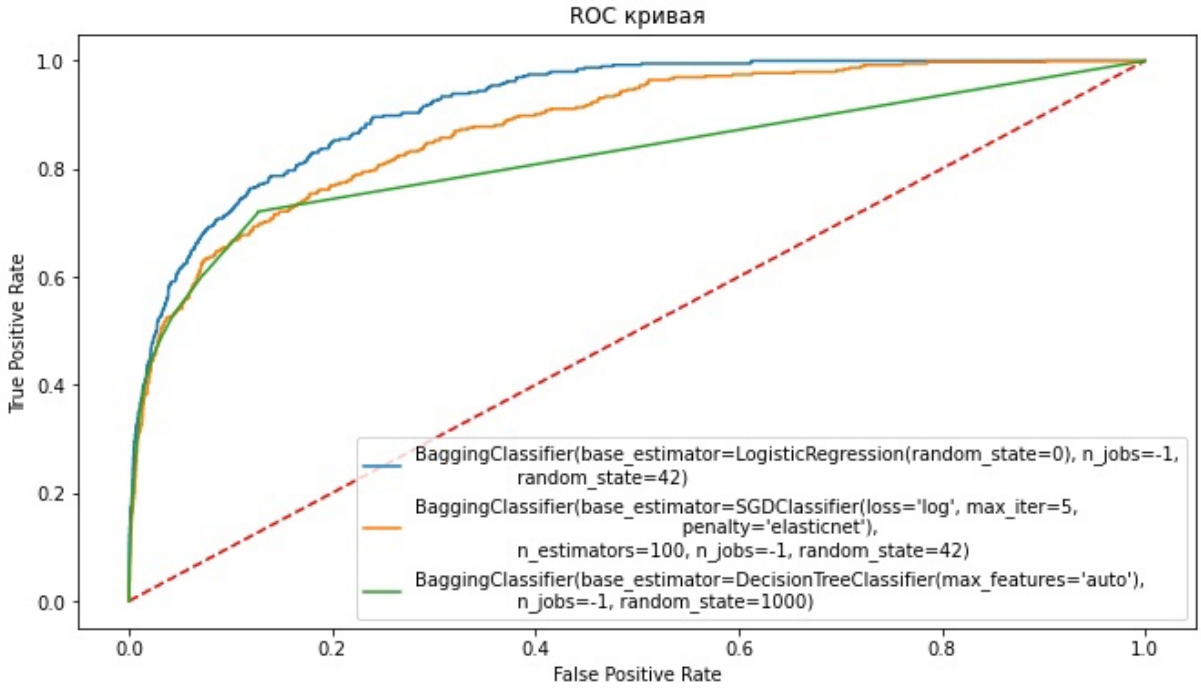


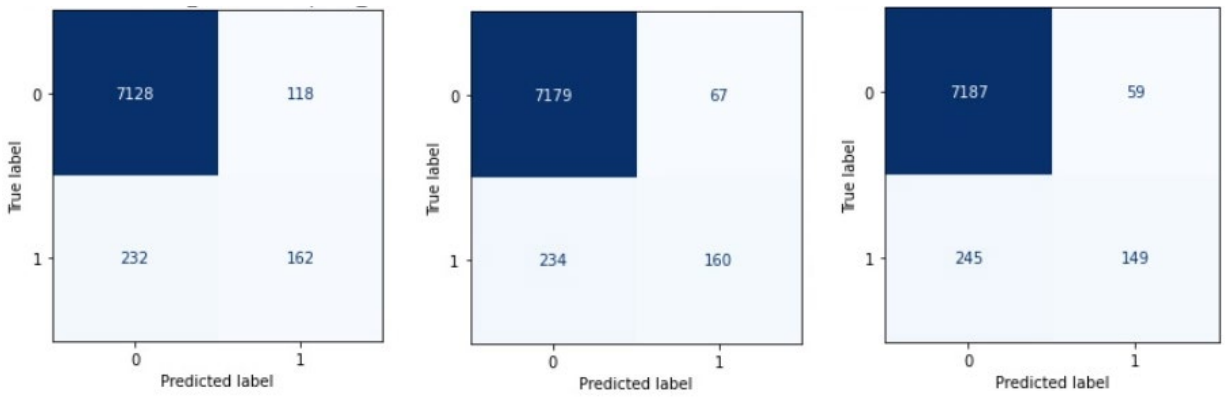
Рисунок 11 – *ROC*-кривая всех моделей

Таблица 5 – Основные метрики качества классификации ансамблевых моделей

Модель	*Precision* 0	*Recall* 0	*F*-мера 0	*Precision* 1	*Recall* 1	*F*-мера 1	*Accuracy*	*AUC*	*Коэфф. Джини*
Bagging на базе стохастического градиентного спуска	0.964	0.992	0.978	0.697	0.327	0.446	0.958	0.878	0.755
Bagging на базе логистической регрессии	0.963	0.991	0.977	0.640	0.302	0.410	0.955	0.918	0.835
Bagging на базе деревьев решений	0.968	0.984	0.976	0.577	0.409	0.478	0.954	0.821	0.642

Ансамблевая модель *Bagging* на базе деревьев решений показала наибольшее количество верно определенных случаев дефолта, она имеет самую высокую среди данных моделей оценку *F*-меры по распознаванию дефолтов – 0,478, тем самым имея наиболее сбалансированную способность определения некредитоспособности. Но стоит отметить, что другие две модели меньше по этому показателю только на три и шесть десятых и остаются на уровне 0,4 и число неверных предсказаний дефолта, от которых будет завесить упущенная выгода банка, принимающего решений, также больше у модели беггинга на базе деревьев решений. А по уровню AUC и Джини, оценивающих качество работы классификатора, модели на основе стохастического градиентного спуска и логистической регрессии превосходят третью модель на одну десятую. Если сравнивать с одиночными моделями, то улучшила свои показатели только модель деревьев решений, другие две показали себя лучше при одиночной работе без применения бэггинга.

Следующими были рассмотрены три самостоятельные ансамблевые модели – случайных лесов, *AdaBoost* и градиентного бустинга. Результаты моделирования приведены на рисунке 12.



а б в

Рисунок 12 – Матрицы ошибок построенных моделей
а) случайных лесов, б) *AdaBoost*, в) градиентного бустинга

По результатам матрицы ошибок были вычислены различные метрики и построена *ROC*-кривая (рисунок 13), по которой были вычислены показатели *AUC* и Джини (таблица 6).

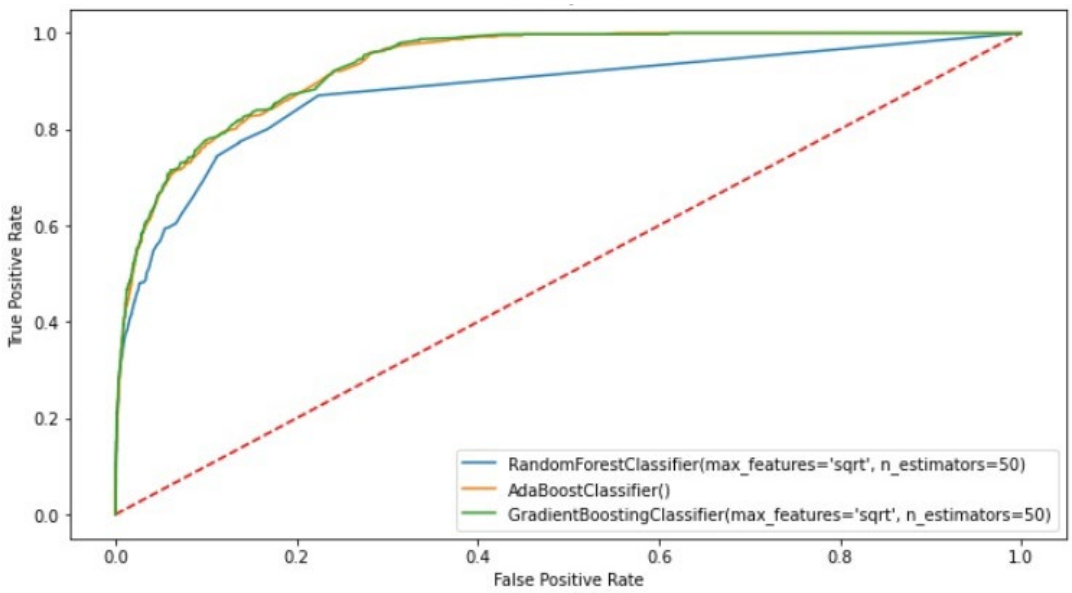


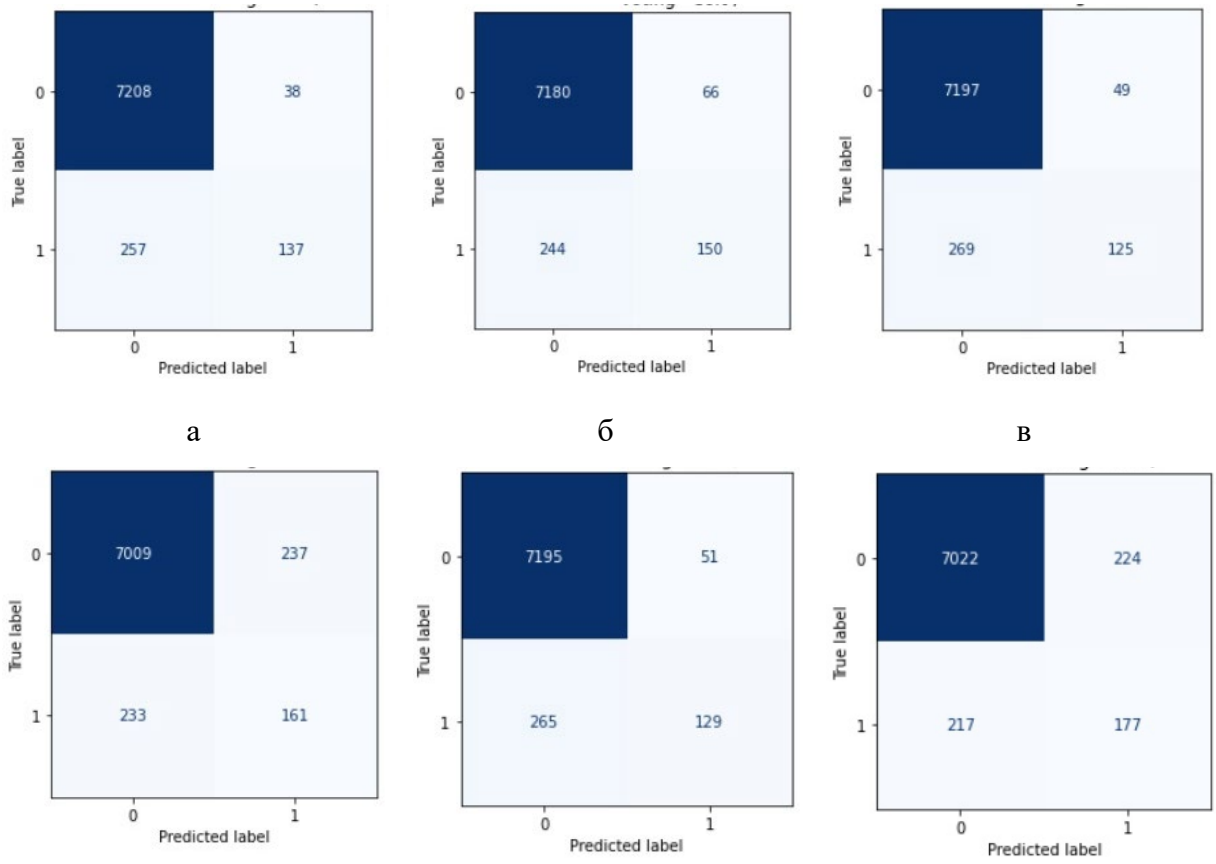
Рисунок 13 – *ROC*-кривая всех моделей

Таблица 6 – Основные метрики качества классификации ансамблевых моделей

Модель	*Precision* 0	*Recall* 0	*F*-мера 0	*Precision* 1	*Recall* 1	*F*-мера 1	*Accuracy*	*AUC*	*Коэфф. Джини*
Случайных лесов	0.968	0.984	0.976	0.579	0.411	0.481	0.954	0.878	0.757
AdaBoost	0.968	0.991	0.979	0.705	0.406	0.515	0.961	0.934	0.868
Градиентный бустинг	0.967	0.992	0.979	0.716	0.378	0.495	0.960	0.936	0.871

Исходя из таблицы 6 можно сделать вывод о том, что все три модели имеют высокие показатели метрик качества. Модель *AdaBoost* показала наивысшее значение *F*-меры среди всех ранее рассмотренных моделей – она единственная получила значение выше 0,5, это касается и показателей, связанные с *ROC*-кривой, они также выше любой другой построенной модели, как одиночной, так и ансамблевой.

При различных комбинациях моделей с помощью метода мягкого голосования были обнаружены следующие прогнозы, отраженные в матрицах ошибок, представленные на рисунке 14.



а б в

г д е

Рисунок 14 – Матрицы ошибок построенных моделей

а) градиентного бустинга и логистической регрессии,

б) логистической регрессии, метода опорных векторов и деревьев решений,

в) логистической регрессии и метода опорных векторов,

г) логистической регрессии и деревьев решений,

д) *AdaBoost* и логистической регрессии,

е) деревьев решений и метода опорных векторов

По результатам матрицы ошибок были вычислены различные метрики и построена *ROC*-кривая (рисунок 15), по которой были вычислены показатели *AUC* и Джини (таблица 7). В ней также приведены расчетные параметры по матрице ошибок.

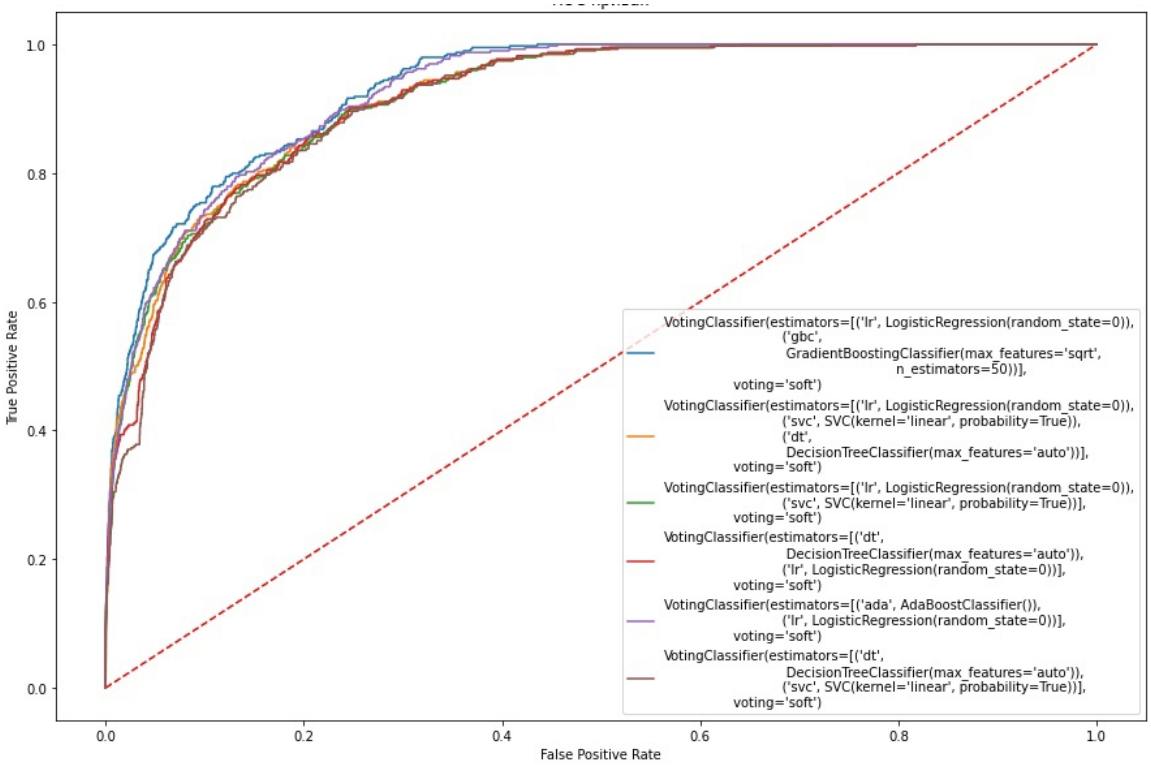


Рисунок 15 – *ROC*-кривая всех моделей

По графикам *ROC*-кривых очевидно, что все рассмотренные модели практически не отличаются по данному параметру – все линии кривых почти совпадают друг с другом и площадь под ними изменяется в небольшом диапазоне.

Таблица 7 – Основные метрики качества классификации ансамблевых моделей

Модель	*Precision* 0	*Recall* 0	*F-*мера 0	*Precision* 1	*Recall* 1	*F-*мера 1	*Accuracy*	*AUC*	*Коэфф. Джини*
Градиентного бустинга и логистической регрессии	0.966	0.995	0.980	0.783	0.348	0.482	0.961	0.932	0.863
Логистической регрессии, метода опорных векторов и деревьев решений	0.967	0.991	0.979	0.694	0.381	0.492	0.959	0.915	0.83

Продолжение таблицы 7

Логистической регрессии и метода опорных векторов	0.964	0.993	0.978	0.718	0.317	0.440	0.958	0.917	0.834
Логистической регрессии и деревьев решений	0.968	0.967	0.968	0.405	0.409	0.407	0.938	0.915	0.829
AdaBoost и логистической регрессии	0.964	0.993	0.979	0.717	0.327	0.449	0.959	0.926	0.852
Деревьев решений и метода опорных векторов	0.970	0.969	0.970	0.441	0.449	0.445	0.942	0.91	0.82

Рассмотренные гибридные модели представляют собой комбинации моделей преимущественно на основе логистической регрессии, так как она показала себя наиболее успешной при построении одиночных моделей. Также было исследовано сочетание методов логистической регрессии и *AdaBoost* – двух лучших моделей, но их совместный результат оказался хуже, чем при применении их по отдельности по всем показателям качества. Наиболее успешной комбинацией для логистической регрессии, которая бы улучшила ее показатели относительно отдельной работы, оказалась комбинация с методом опорных векторов и деревьями решений. Она дала улучшение в пять десятых в метрике F-меры, но другие оценки остались на уровне работы одиночной модели логистической регрессии.

Таким образом, в результате проведенного исследования было выяснено, что для данного набора данных наиболее подходящей с точки зрения прогнозирования дефолтных случаев заема оказались модели логистической регрессии и ансамблевой модели адаптивного бустинга, основанной на классификаторе деревьев решений. Полученные результаты качества классификации на классы в дальнейшем можно улучшить путем

исследования и подбора параметров моделей, так в ансамблях бустинга новых показателей можно добиться, подбирая базовый алгоритм, максимальное количество оценок, после которого процесс обучения прекращается, вклад каждой модели в весовые коэффициенты. Также и в одиночных моделях возможности библиотек и функций на языке Python позволяют изменять их различные параметры.

3. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Целью магистерской диссертации является разработка математической модели кредитного скоринга для банковских систем.

Целью раздела «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» является определения перспективности разработанной математической модели кредитного скоринга, успешности её реализации на рынке.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить потенциальных потребителей результатов исследования.
2. Выявить сильные и слабые стороны научно-исследовательского проекта, а также его возможности и вероятные угрозы при помощи SWOT-анализа.
3. Оценить степень готовности научного проекта к коммерциализации.
4. Определить заинтересованные стороны и ограничения/допущения научно-технического исследования; сформулировать цель и ожидаемые результаты проекта.
5. Определить структуру и трудоемкость выполнения работ, разработать график проведения научного исследования.
6. Рассчитать бюджет научно-технического исследования.
7. Определить риск возникновения неопределённых событий при выполнении НТИ, которые могут повлечь за собой нежелательные эффекты.

3.1 Предпроектный анализ

3.1.1 Потенциальные потребители результатов исследования

Для анализа потребителей результатов исследования необходимо рассмотреть целевой рынок и провести его сегментирование.

Целевой рынок – сегменты рынка, на котором будет продаваться в будущем разработка. В свою очередь, сегмент рынка – это особым образом выделенная часть рынка, группы потребителей, обладающих определенными общими признаками.

Сегментирование – это разделение покупателей на однородные группы, для каждой из которых может потребоваться определенный товар (услуга).

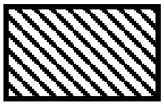
Разработка системы кредитного скоринга является актуальной, поскольку она поможет проводить оценки платежеспособности клиентов и принимать более точные решения о предоставлении банком денежных средств клиентам.

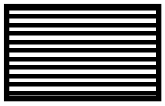
Потенциальными потребителями результатов исследования могут быть частые и государственные банковские и финансовые организации, как крупные, так и мелкие.

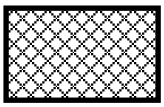
В настоящее время участники рынка используют различные скоринговые системы, основные из которых Basegroup Labs, «Диасофт», в то же время многие банки разрабатывает свои собственные системы (таблица 8).

Таблица 8 – Карта сегментирования рынка услуг по методам прогнозирования

Потребитель	Скоринговые системы		
	Basegroup Labs	«Диасофт»	Собственные системы
Государственные банки			
Частные банки			
Крупные микрофинансовые организации			

 - фирма А

 - фирма В

 - фирма Б

Таким образом, мы определили, какие ниши на рынке услуг по применению модели скоринга не заняты конкурентами или где уровень конкуренции низок, следовательно определили потенциальных потребителей исследования – результаты данной работы будут наиболее привлекательны преимущественно для государственных и частных банков.

3.1.2 SWOT-анализ

Для комплексной оценки научно-исследовательского проекта применяют SWOT-анализ, результатом которого является описание сильных и слабых сторон проекта, выявление возможностей и угроз для его реализации, которые проявились или могут появиться в его внешней и внутренней среде.

Разработанная для данного исследования матрица SWOT представлена в таблице 9.

Таблица 9 – Матрица SWOT

	Сильные стороны научно-исследовательского проекта: С1. Переобучение модели на основе новых данных. С2. Низкая стоимость проекта. С3. Актуальность скоринговой модели с использованием нескольких моделей машинного обучения. С4. Высокая точность прогнозирования по сравнению с конкурентами.	**Слабые стороны научно-исследовательского проекта:** Сл1. Ограниченные данные для обучения модели. Сл2. Отсутствие продвижения на рынке. Сл3. Невозможность контроля ввода неверных данных в модель.
Возможности: В1. Расширение прогнозирования не только решения о выдачи или не выдачи кредита, но и прогнозирования сроков погашение клиентом и т.д. В2. Развитие проекта с привлечением данных крупных банков. В3. Улучшение модели с помощью других методов машинного обучения, например, нейронных сетей.	1. Благодаря точности модели целесообразно будет развивать прогнозирования по многим экономическим параметрам. 2. Низкая стоимость проекта повышает конкурентоспособность по отношению к имеющимся скоринговым системам.	1. Переобучение модели на реальных данных банков позволит создать ещё более точную систему. 2. Дальнейшие исследования и улучшении модели с помощью машинного обучения будут способствовать продвижению на рынке.
Угрозы: У1. Потеря актуальности в связи с устаревшими данными. У2. Развитие конкурентных технологий. У3. Отсутствие доверия у крупных банков к новой модели.	1. Актуальность, эффективность и дешевизна проекта способствуют преодолению недоверия к новой системе. 2. Эффективная переобучаемость модели не позволит данной модели стать менее конкурентоспособной в случае развития конкурентной технологии.	1. Отсутствие доверия к новой технологии может усугубиться ограниченностью имеющихся данных для обучения. 2. Улучшение параметров конкурентных моделей может негативно отразится на продвижении проекта.

Таким образом, в ходе проведения SWOT-анализа были выявлены сильные и слабые стороны научно-исследовательского проекта, а также его возможности и вероятные угрозы. Необходимо сделать упор на такие сильные стороны, как точность, низкая стоимость и переобучаемость, так как именно эти сильные стороны проекта связаны с наибольшим количеством возможностей. Что касается слабых стороны, необходимо обратить внимание на работу с данными для обучения модели, их актуальностью. Работа над

этими недостатками позволить повысить конкурентоспособность, уменьшить влияние внешних угроз на проект.

3.1.3 Оценка готовности проекта к коммерциализации

На какой бы стадии жизненного цикла не находилась научная разработка полезно оценить степень ее готовности к коммерциализации и выяснить уровень собственных знаний для ее проведения (или завершения).

Для этого необходимо заполнить форму (таблица 10), которая содержит показатели о степени проработанности проекта с позиции коммерциализации и компетенциям разработчика научного проекта.

Оценки степени проработанности научного проекта трактуются следующим образом:

1 – не проработано;

2 – проработано слабо;

3 – выполнено, но качество под сомнением;

4 – выполнено качественно;

5 – имеется положительное заключение независимого эксперта.

Оценка уровня имеющихся знаний у разработчика определяется в соответствии со следующей системой баллов:

1 – не знаком или знаком мало;

2 – знаком с теорией;

3 – знаком с теорией и практическими примерами применения;

4 – знаком с теорией и самостоятельно выполняет;

5 – знаком с теорией, выполняет, может консультировать.

Таблица 10 – Бланк оценки степени готовности научного проекта к коммерциализации

п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1	Определен имеющийся научно-технический задел	5	4
2	Определены перспективные направления коммерциализации научно-технического задела	4	3
3	Определены отрасли и технологии (товары, услуги) для предложения на рынке	4	3
4	Определена товарная форма научно-технического задела для представления на рынок	3	3
5	Определены авторы и осуществлена охрана их прав	4	4
6	Проведена оценка стоимости интеллектуальной собственности	3	2
7	Проведены маркетинговые исследования рынков сбыта	3	2
8	Разработан бизнес-план коммерциализации научной разработки	3	2
9	Определены пути продвижения научной разработки на рынок	3	3
10	Разработана стратегия (форма) реализации научной разработки	5	4
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	5	4
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	4	3
13	Проработаны вопросы финансирования коммерциализации научной разработки	4	4
14	Имеется команда для коммерциализации научной разработки	5	5
15	Проработан механизм реализации научного проекта	4	5
	Итого баллов	59	51

Таким образом, по результатам проведенной оценки степени готовности научного проекта к коммерциализации суммарное количество баллов по степени проработанности научного проекта превышает уровень

имеющихся знаний у разработчика. Согласно полученным баллам, можно сказать, что перспективность данной разработки выше среднего.

3.1.4 Методы коммерциализации результатов научно-технического исследования

При коммерциализации научно-технических разработок продавец (владелец соответствующих объектов интеллектуальной собственности) преследует определенную цель, которая во многом зависит от того, куда он намерен направить полученный коммерческий эффект. Это может быть получение средств для продолжения научных исследований, одноразовое получение финансовых ресурсов, обеспечение постоянного притока финансовых средств, а также их различные сочетания. В связи с этим необходимо выбрать наиболее подходящий метод коммерциализации и обосновать его целесообразность.

Выделяют следующие методы коммерциализации научных разработок: торговля патентными лицензиями, передача ноу-хау, инжиниринг, франчайзинг и пр.

Перспективность данного научного исследования выше среднего, однако еще не все аспекты глубоко изучены и проработаны.

Таким образом, проанализировав перечисленные методы коммерциализации, успешному продвижению проекта на данной стадии, на которой находится научный проект, соответствует торговля патентными лицензиями, поскольку степень проработанности проекта и уровня знаний разработчика будет достаточно для реализации данного метода.

3.2 Инициация проекта

Группа процессов инициации состоит из процессов, которые выполняются для определения нового проекта или новой фазы существующего. В рамках процессов инициации определяются изначальные

цели и содержание, фиксируются изначальные финансовые ресурсы. Определяются внутренние и внешние заинтересованные стороны проекта, которые будут взаимодействовать и влиять на общий результат. Данная информация закрепляется в уставе проекта.

Устав проекта документирует бизнес-потребности, текущее понимание потребностей заказчика проекта, а также новый продукт, услугу или результат, который планируется создать.

Заинтересованные стороны проекта представлены в таблице 11, цели и результат проекта – в таблице 12.

Таблица 11 – Заинтересованные стороны проекта

Заинтересованные стороны проекта	Ожидания заинтересованных сторон
Банки Финансовые организации	Иметь возможность спрогнозировать кредитоспособность клиентов. Получать высокую точность результатов. Использовать большие объёмы данных.

Таблица 12 – Цели и результат проекта

Цели проекта	Разработка математической модели для системы кредитного скоринга.
Ожидаемые результаты проекта	Осуществление прогнозирования данной моделью кредитоспособности клиента.
Критерии приемки результата проекта	Высокая точность результата прогнозирования полученной моделью.
Требования к результату проекта	– Адекватность полученных результатов – Точность системы на уровне % – Возможность обучения модели на новых данных

В данном подразделе мы определили заинтересованные стороны проекта и сформулировали цели и ожидаемый результат НТИ. Целью проекта является разработка математической модели для системы кредитного скоринга, а результатом – осуществление прогнозирования данной моделью кредитоспособности клиента.

3.2.1 Организационная структура проекта

На данном этапе необходимо сформировать рабочую группу (таблица 13), определить роль каждого участника, прописать функции, выполняемые каждым из участников и их трудозатраты в проекте.

Таблица 13 – Рабочая группа проекта

п/п	ФИО, основное место работы	Роль в проекте	Функции	Трудозатр. час.
1	Семенов М.Е., доцент ОЭФ	Руководитель	Составление и утверждение научного задания, календарное планирование работ по теме, оценка эффективности полученных результатов	88
2	Шеров Ш. Магистрант	Исполнитель	Выполнение поставленной задачи, составление и оформление пояснительной записки к ВКР	728
			Итого:	816

Таким образом, мы сформировали рабочую группу проекта, а именно – определили роль каждого участника, прописали функции, выполняемые каждым из участников и их трудозатраты в проекте, основная часть трудозатрат ложится на исполнителя-магистранта.

3.2.2 Ограничения и допущения проекта

Ограничения проекта (таблица 14) – это все факторы, которые могут послужить ограничением степени свободы участников команды проекта, а

также «границы проекта» – параметры проекта или его продукта, которые не будут реализованы в рамках данного проекта.

Таблица 14 – Ограничения проекта

Фактор	Ограничения/допущения
Источник финансирования	НИ ТПУ
Сроки проекта	01.03.2021 – 06.06.2021
Дата утверждения плана управления проектом	01.03.2021
Дата завершения проекта	06.06.2021
Прочие ограничения и допущения	Отсутствуют

Таким образом, мы определили ограничения и допущения проекта. Ограничениями научно-технического исследования являются сроки выполнения.

3.3 Планирование научно-исследовательских работ

Планирование комплекса предполагаемых работ осуществляется в следующем порядке:

– Формирование иерархической структуры работ проекта

– Определение ключевых (контрольных) событий проекта

– Построение календарного графика проекта

– Планирование бюджета научного исследования.

3.3.1 Иерархическая структура работ проекта

Иерархическая структура работ (ИСР) – это детализация укрупненной структуры работ. В процессе создания ИСР структурируется и определяется содержание всего проекта. На рисунке 16 представлена ИСР для выполнения магистерской диссертации.

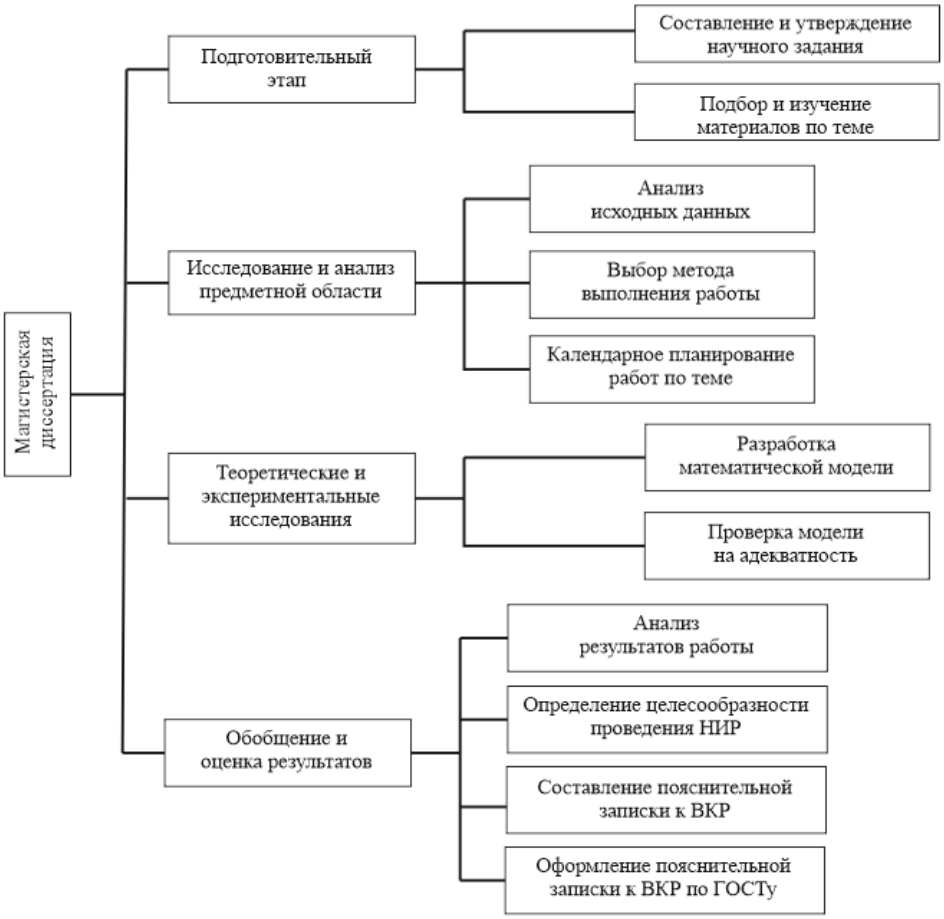


Рисунок 16 – Иерархическая структура работ проекта

Таким образом, мы сформировали иерархическую структуру работ проекта для выполнения данной магистерской диссертации, она поможет в последовательном исполнении поставленных задач.

3.3.2 Структура работ в рамках научного исследования

Трудоемкость выполнения ВКР оценивается экспертным путем в человеко-днях и носит вероятностный характер, т.к. зависит от множества трудно учитываемых факторов.

Для выполнения научно-исследовательской работы формируется рабочая группа, в состав которой могут входить:

– Руководитель проекта (Р);

– Инженер (магистрант) (И).

На следующем этапе составляется перечень работ в рамках проведения научного исследования, а также проводится распределение исполнителей по

видам работ. Примерный порядок составления этапов и работ, распределение исполнителей по данным видам работ приведен в таблице 15.

Таблица 15 – Комплекс работ по разработке проекта

Основные этапы	№	Содержание работ	Должность исполнителя
Подготовительный	1	Составление и утверждение научного задания	Руководитель
	2	Подбор и изучение материалов по теме	Инженер
Исследование и анализ предметной области	3	Анализ исходных данных	Инженер
	4	Выбор метода выполнения работы	Руководитель Инженер
	5	Календарное планирование работ по теме	Инженер
Теоретические и экспериментальные исследования	6	Разработка математической модели	Инженер
	7	Проверка модели на адекватность	Инженер
Обобщение и оценка результатов	8	Анализ результатов работы	Руководитель Инженер
	9	Определение целесообразности проведения НИР	Руководитель Инженер
	10	Составление пояснительной записки к ВКР	Инженер
	11	Оформление пояснительной записки к ВКР по ГОСТу	Инженер

На данном этапе мы составили перечень работ в рамках проведения научного исследования, а также провели распределение исполнителей по видам работ, также основные работы выполняет инженер, руководитель осуществляет контролирующую и вспомогательную роль.

3.3.3 Определение трудоемкости выполнения работ и разработка графика проведения научного исследования

Трудовые затраты в большинстве случаев образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, т.к. зависит от множества трудно учитываемых факторов. Для определения ожидаемого (среднего) значения трудоемкости используется следующая формула:

$$t_{ожi} = \frac{3t_{\min i} + 2t_{\max i}}{5}, \quad (1)$$

где $t_{ож\ i}$ – ожидаемая трудоемкость выполнения i-й работы, человеко-дни;

$t_{min\ i}$ – минимально возможная трудоемкость выполнения заданной i-й работы, человеко-дни;

$t_{max\ i}$ – максимально возможная трудоемкость выполнения заданной i-й работы, человеко-дни;

Рассчитаем значение ожидаемой трудоемкости работы.

Установление длительности работ в рабочих днях осуществляется по формуле:

$$T_{p_i} = \frac{t_{ожi}}{Ч_i}, \quad (2)$$

где T_{pi} – продолжительность одной работы, раб. дн.;

$Ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

При выполнении дипломных работ студенты в основном становятся участниками сравнительно небольших по объему научных тем. Поэтому наиболее удобным и наглядным является построение ленточного графика проведения научных работ в форме диаграммы Гантта.

Диаграмма Гантта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Для удобства построения графика, длительность каждого из этапов работ из рабочих дней следует перевести в календарные дни. Для этого необходимо воспользоваться формулой:

$$T_{ki} = T_{pi} \cdot k_{кал}, \quad (3)$$

где $T_{кi}$ – продолжительность выполнения i-й работы в календарных днях;

T_{pi} – продолжительность выполнения i-й работы в рабочих днях;

$k_{кал}$ – коэффициент календарности, предназначен для перевода рабочего времени в календарное.

Коэффициент календарности определяется по формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}}, \quad (4)$$

где $T_{кал}$ – количество календарных дней в году;

$T_{вых}$ – количество выходных дней в году;

$T_{пр}$ – количество праздничных дней в году.

Рассчитанные значения в календарных днях по каждой работе необходимо округлить до целого числа.

Согласно производственному календарю (для 6-дневной рабочей недели) в 2021 году 365 календарных дней, 122 выходных/праздничных дней.

Вычислим коэффициент календарности:

$$k_{кал} = \frac{365}{365 - 122} = 1,5$$

Рассчитанные временные показатели были сведены в таблицу 16, представленную ниже.

Таблица 16 – Временные показатели осуществления комплекса работ

№ работы	Продолжительность работ			Исполнители	T_{pi}, человеко-дни	T_{ki}, человеко-дни
	$t_{min\ i}$, человеко-дни	$t_{max\ i}$, человеко - дни	$t_{ожi}$, человеко - дни			
1	1	5	3	Р	2	3
2	11	17	13	И	13	20
3	2	9	5	И	5	7
4	5	7	6	Р, И	3	4
5	1	3	2	И	2	3
6	11	17	13	И	13	20
7	1	1	1	И	1	1
8	4	6	5	Р, И	5	7
9	5	7	6	Р, И	3	4
10	5	10	8	И	8	12
11	2	5	3	И	3	5
				Итого:	57	86

На основании таблицы 16 составлен календарный план-график, показывающий продолжительность выполнения работ ВКР. В результате планирования графика, продолжительность работ равна трём месяцам (рисунок 17).

Красным цветом на рисунке обозначены работы, выполненные руководителем, синим – руководителем и инженером, зеленым – инженером.

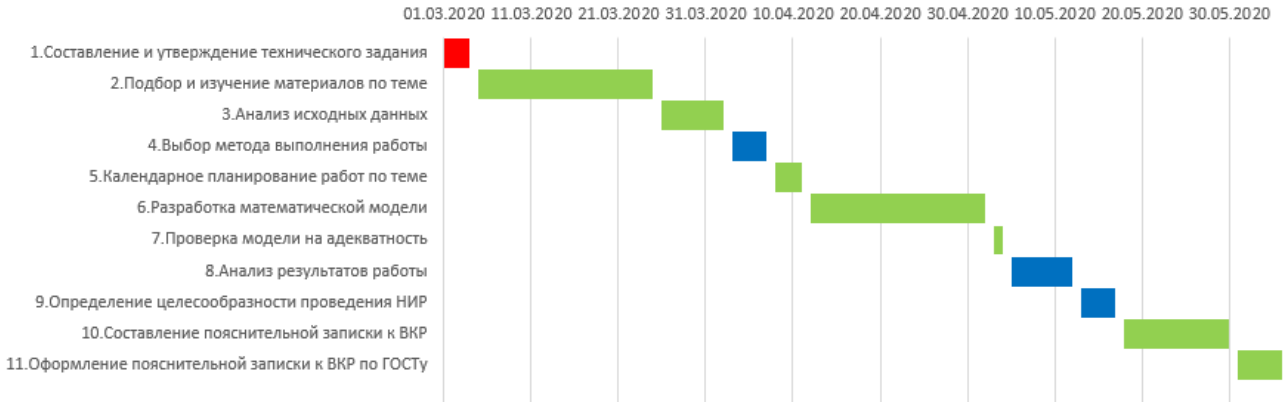


Рисунок 17 – Календарный план-график выполнения работ (диаграмма Гантта)

Таким образом, мы определили трудоемкость выполнения работ и разработали календарный план-график проведения магистерской диссертации по теме. По нему можно увидеть, что самые продолжительные по времени работы – это подбор и изучение материалов по теме (20 дней) и разработка математической модели (20 дней).

3.3.4 Бюджет научно-технического исследования

При планировании бюджета НТИ должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В процессе формирования бюджета НТИ используется следующая группировка затрат по статьям:

– Сырье, материалы (за вычетом возвратных отходов), покупные изделия и полуфабрикаты;

– Основная заработная плата;

– Отчисления на социальные нужды;

– Накладные расходы.

3.3.4.1 Затраты на материалы

Данная статья отражает стоимость всех материалов, используемых при разработке проекта, включая расходы на их приобретение и доставку. В материальные затраты, помимо вышеуказанных, включаются дополнительно затраты на канцелярские принадлежности, диски, картриджи и т.п. Однако их учет ведется в данной статье только в том случае, если в научной организации их не включают в расходы на использование оборудования или накладные расходы.

Материальные затраты, необходимые для данной разработки, заносятся в таблицу 17.

Таблица 17 – Материальные затраты

Наименование	Единица измерения	Количество	Цена за ед., руб	Затраты на материалы ($З_м$), руб.
Бумага, формат А4	Пачка	1	310	310
Канцелярские принадлежности	Шт	1	250	250
Ноутбук	Шт	1	35000	35000
			Итого:	35560

На данном этапе мы рассчитали материальные затраты, необходимые для проведения научно-технического исследования. Они составляют 35560 руб.

3.3.4.2 Основная заработная плата

Величина расходов по заработной плате определяется исходя из трудоемкости выполняемых работ и действующей системы окладов и тарифных ставок. В состав основной заработной платы включается премия, выплачиваемая ежемесячно из фонда заработной платы в размере 20 – 30 % от тарифа или оклада.

Статья включает основную заработную плату работников, непосредственно занятых выполнением НТИ, (включая премии, доплаты) и дополнительную заработную плату:

$$C_{зп} = З_{осн} + З_{доп}, \quad (5)$$

где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата.

Основная заработная плата ($З_{осн}$) руководителя (лаборанта, инженера) от предприятия (при наличии руководителя от предприятия) рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot T_{раб}, \quad (6)$$

где $З_{осн}$ – основная заработная плата одного работника;

T_p – продолжительность работ, выполняемых научно-техническим работником, раб. дн.;

$З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_{м} \cdot М}{F_{д}}, \quad (7)$$

где $З_{м}$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года;

$F_{д}$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн.;

М – количество месяцев работы без отпуска в течение года:

при отпуске в 24 раб. дня М = 11,2 месяца, 5 – дневная неделя;

при отпуске в 48 раб. дней М = 10,4 месяца, 6-дневная неделя.

Таблица 18 – Баланс рабочего времени

Показатели рабочего времени	Руководитель	Инженер
Календарное число дней	365	365
Количество нерабочих дней – выходные дни – праздничные дни	122	122

Потери рабочего времени – отпуск – невыходы по болезни	48 –	48 –
Действительный годовой фонд рабочего времени	195	195

Месячный должностной оклад работника:

$$З_{\text{М}} = З_{\text{б}} \cdot (k_{\text{пр}} + k_{\text{д}}) \cdot k_{\text{р}}, \quad (8)$$

где $З_{\text{б}}$ – базовый оклад, руб.;

$k_{\text{пр}}$ – премиальный коэффициент, определяется Положением об оплате труда;

$k_{\text{д}}$ – коэффициент доплат и надбавок составляет примерно 0,2 – 0,5;

$k_{\text{р}}$ – районный коэффициент, равный 1,3 (для г. Томска).

Результат расчетов заработных плат представлен в таблице 19.

Таблица 19 – Расчёт основной заработной платы

Исполнители	$З_{\text{б}}$, руб	$k_{\text{р}}$	$З_{\text{м}}$, руб	$З_{\text{дн}}$, руб	$Т_{\text{р}}$, дни	$З_{\text{осн}}$, руб
Руководитель	35 120	1,3	45 656	2 374,11	18	42733,98
Инженер	12 000	1,3	15 600	811,2	83	67329,6
					Итого:	110063,58

Таким образом, мы рассчитали основную заработную плату исполнителей данного научно-технического исследования. Общая сумма заработной платы участников проекта составляет 110063,58 руб.

3.3.4.3 Отчисления во внебюджетные фонды

Отчисления во внебюджетные фонды являются обязательными по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников.

Величина отчислений во внебюджетные фонды определяется исходя из следующей формулы:

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

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

Коэффициент отчислений на уплату во внебюджетные фонды равен 30,2%.

Отчисления во внебюджетные фонды представлены в таблице 20.

Таблица 21 – Отчисления во внебюджетные фонды

Исполнители	Основная ЗП, руб
Руководитель	42733,98
Инженер	67329,6
Коэффициент отчислений во внебюджетные фонды	0,302
Итого:	33240

В данном подразделе мы рассчитали отчисления во внебюджетные фонды, которые являются обязательными по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников. Сумма отчислений во внебюджетные фонды составляет 33240 руб.

3.3.4.4 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 и т.д.

Так как работа производилась только с использованием персонального компьютера, все накладные расходы составляет плата за электроэнергию и интернет. В расчётах будем учитывать, что мощность компьютера руководителя равна $P_{рук} = 0.1$ кВт, мощность компьютера исполнителя – $P_{исп}$

= 0,1 кВт. Также учитываем одинаковую плату за интернет $S_и$ = 350 руб/мес. Тогда при 8-часовом рабочем дне накладные расходы составляют:

$$C_{накл} = 8 \cdot (T_{рук} \cdot P_{рук} + T_{исп} \cdot P_{исп}) \cdot S_{эл} + T_р / 30 \cdot S_и, \quad (10)$$

где $S_{эл}$ = 5.8 руб / кВт · ч — удельная плата за электроэнергию.

Следовательно, $C_{накл}$ составляют 3533,8 руб.

Таким образом, мы рассчитали плату за электроэнергию и интернет, которые и составляют накладные расходы. Итоговая сумма накладных расходов равна 3533,8 руб.

3.3.4.5 Формирование бюджета затрат НТИ

Рассчитанная величина затрат научно-исследовательской работы является основой для формирования бюджета затрат проекта, который при формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции. Определение бюджета затрат на научно-исследовательский проект приведен в таблице 22.

Таблица 22 – Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.
1. Материальные затраты НТИ	35560
2. Затраты по основной заработной плате исполнителей темы	110063,58
3. Отчисления во внебюджетные фонды	33240
4. Накладные расходы	3533,8
5. Бюджет затрат НТИ	182398,38

Подводя итог, мы можем сделать вывод, что бюджет затрат научно-технического исследования равен 182398,38 руб.

3.4 Реестр рисков проекта

Во время проекта существует риск возникновения неопределённых событий, которые могут повлечь за собой нежелательные эффекты. Для таких событий составлен реестр рисков, содержащий в себе общую информацию о них (таблица 23). Вероятность наступления и влияние определённого риска оцениваются по пятибалльной шкале. Уровень риска может быть высокий, средний или низкий в зависимости от вероятности наступления и степени влияния риска.

Таблица 23 – Реестр рисков

Риск	Потенциальное воздействие	Вероятность наступления	Влияние	Уровень	Способы смягчения	Условия наступления
Управление проектом	Некорректный сбор информации	3	5	Высокий	Распределение обязанностей	Несогласованность действие
Технический	Некорректные результаты	3	5	Высокий	Чёткое планирование	Несогласованность действий
Внешний	Несоответствие плану	2	2	Низкий	Резервное время	Отсутствие данных

По результатам данного подраздела можно сделать вывод, что риск возникновения неопределённых событий, которые могут повлечь за собой нежелательные эффекты, существует, но вероятность наступления его маловероятна.

Выводы по главе «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

1. Определены потенциальные потребители результатов исследования – результаты данной работы будут наиболее привлекательны преимущественно для государственных и частных банков.

2. Выявлены сильные и слабые стороны научно-исследовательского проекта, а также его возможности и вероятные угрозы при помощи SWOT-анализа. Необходимо сделать упор на такие сильные стороны, как точность,

низкая стоимость и переобучаемость модели, так как именно эти сильные стороны проекта связаны с наибольшим количеством возможностей – это расширение показателей прогнозирования, развитие проекта с привлечением данных крупных банков, улучшение модели с помощью других методов машинного обучения. Отсюда следует, что разработка модели является перспективным проведением научного исследования.

3. Определена степень готовности научного проекта к коммерциализации: согласно полученным результатам, можно сказать, что перспективность данной разработки выше среднего. Выбран метод коммерциализации результатов НТИ: торговля патентными лицензиями.

4. Определены заинтересованные стороны проекта: банки и финансовые организации. Ограничениями научно-технического исследования являются сроки выполнения. Целью проекта является разработка математической модели для системы кредитного скоринга, а результатом – осуществление прогнозирования данной моделью кредитоспособности клиента.

5. В ходе планирования научно-исследовательских работ определены структура и перечень работ, выполняемых рабочей группой. В данном случае рабочая группа состоит из руководителя и инженера, длительность работ для руководителя составляет 18 дней, а для инженера – 68 дней. Был построен календарный план-график на основе диаграммы Гантта, по которому можно увидеть, что самые продолжительные по времени работы – это подбор и изучение материалов по теме (20 дней) и разработка математической модели (20 дней).

6. Бюджет научно-технического исследования составил 182398,38 руб. Основную часть бюджета составила зарплата работников (110063,58 руб).

7. Определен риск возникновения неопределённых событий при выполнении НТИ: риск возникновения неопределённых событий, которые могут повлечь за собой нежелательные эффекты, существует, но вероятность наступления его маловероятна.

8. Таким образом, капиталовложения в размере 182398,38 рубля позволят реализовать разработку математической модели кредитного скоринга. Система кредитного скоринга позволяет оценить риски выдачи кредита и составить кредитный рейтинг клиента для принятия соответствующего решения банком.

4 Социальная ответственность

Тема магистерской диссертации направлена на разработку системы кредитного скоринга с использованием логистической модели.

Рассматриваемое рабочее место находится в 427А аудитории 10 корпуса. Аудитория оснащена персональным компьютером, с помощью программного обеспечения на котором происходит разработка математической модели. Также рабочее место составляет рабочий стол, стул. В помещении имеются окна, через которые осуществляется вентиляция помещения. В зимнее время аудитория отапливается, что обеспечивает достаточное, постоянное и равномерное нагревание воздуха. Также используется комбинированное освещение – искусственное и естественное.

4.1 Правовые и организационные вопросы обеспечения безопасности

4.1.1 Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства

В соответствии с государственными стандартами и правовыми нормами обеспечения безопасности предусмотрена рациональная организация труда в течение смены, которая предусматривает:

1) длительность рабочей смены не более 8 часов;

2) установление двух регламентируемых перерывов (не менее 20 минут после 1-2 часов работы, не менее 30 минут после 2 часов работы);

3) обеденный перерыв не менее 40 минут.

Рабочее место должно быть организовано в соответствии с требованиями стандартов, технических условий и (или) методических указаний по безопасности труда. Оно должно удовлетворять следующим требованиям:

1) обеспечивать возможность удобного выполнения работ;

2) учитывать физическую тяжесть работ;

3) учитывать размеры рабочей зоны и необходимость передвижения в ней работающего;

4) учитывать технологические особенности процесса выполнения работ.

Невыполнение требований к расположению и компоновке рабочего места может привести к получению работником производственной травмы или развития у него профессионального заболевания.

4.1.2 Организационные мероприятия при компоновке рабочей зоны

При организации рабочего места основной целью является обеспечение качественного и эффективного выполнения работы при полном использовании оборудования в соответствии с установленными сроками [17]. В связи с этим требования к рабочему месту носят следующий характер:

1) Конструкция рабочей мебели (рабочий стол, кресло, подставка для ног) должна обеспечивать возможность индивидуальной регулировки соответственно росту пользователя и создавать удобную позу для работы. Вокруг ЭВМ должно быть обеспечено свободное пространство не менее 60-120 см [18];

2) Рекомендованная высота рабочей поверхности стола в пределах 680 – 800 мм. Высота рабочей поверхности, на которую устанавливается клавиатура, должна быть 650 мм. Рабочий стол должен быть шириной не менее 700 мм и длиной не менее 1400 мм. Должно иметься пространство для ног высотой не менее 600 мм, шириной — не менее 500 мм, глубиной на уровне колен — не менее 450 мм и на уровне вытянутых ног — не менее 650 мм [19].

3) Монитор должен быть расположен на уровне глаз оператора на расстоянии 500–600 мм. Согласно нормам, угол наблюдения в горизонтальной

плоскости должен быть не более 45º к нормали экрана. Лучше если угол обзора будет составлять 30º. Кроме того должна быть возможность выбирать уровень контрастности и яркости изображения на экране. Должна предусматриваться возможность регулирования экрана [20].

Место для работы на компьютере и взаиморасположение всех его элементов должно соответствовать антропометрическим, физическим и психологическим требованиям. При устройстве рабочего места человека, работающего за ПК необходимо соблюсти следующие основные условия: наилучшее местоположение оборудования и свободное рабочее пространство [21].

4.2 Производственная безопасность

4.2.1. Анализ вредных и опасных факторов, которые могут возникнуть в лаборатории при проведении исследований

Для выбора факторов использовался ГОСТ 12.0.003-2015 «Опасные и вредные производственные факторы. Классификация» [22]. Перечень опасных и вредных факторов, характерных для проектируемой производственной среды представлен в виде таблицы 24.

Таблица 24 – Опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разработка	Изготовление	Эксплуатация	
Повышенный уровень шума	+	+		СН 2.2.4/2.1.8.562-96 Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки
Недостаточная освещенность рабочей зоны	+	+		СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*

Продолжение таблицы 24

Умственное перенапряжение	+	+		Р 2.2.2006-05. Гигиена труда. Руководство по гигиенической оценке факторов рабочей среды и трудового процесса. Критерии и классификация условий труда
Перенапряжение зрительного анализатора	+	+		Р 2.2.2006-05. Гигиена труда. Руководство по гигиенической оценке факторов рабочей среды и трудового процесса. Критерии и классификация условий труда
Воздействие переменных электромагнитных полей	+	+	+	СанПиН 2.2.4.3359-16 Санитарно-эпидемиологические требования к физическим факторам на рабочих местах
Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	+	+	+	ГОСТ 12.1.019-2017 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты

4.2.2 Обоснование мероприятий по защите исследователя от действия опасных и вредных факторов

Повышенный уровень шума

Шумовое загрязнение среды на рабочем месте приводит к снижению внимания исследователя, замедлению скорости психических реакций. Шумовой фон в помещении возникает из-за работы компьютеров, принтеров, телефонов и систем вентиляции.

Для избегания вышеуказанных последствий воздействия описываемого фактора, необходимо соблюдать следующие требования, обозначенные в СН 2.2.42.1.8.562-96 [23]. Уровень шума на рабочем месте математиков-программистов не должен превышать 50 дБА, а в залах обработки информации на вычислительных машинах – 65 дБА. Уровень звука и звукового давления измеряется на расстоянии 50 см от поверхности оборудования и на высоте

источников звука. В таблице 25 приведены допустимые значения уровней звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ.

Таблица 25 – Допустимые значения уровней звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ

Уровни звукового давления в октавных полосах со среднегеометрическими частотами (Гц)								
31,5	63	125	250	500	1000	2000	4000	8000
Уровни звука (дБА)								
86 дБ	71 дБ	61 дБ	54 дБ	49 дБ	45 дБ	42 дБ	40 дБ	38 дБ

При значениях выше допустимого уровня необходимо предусмотреть средства индивидуальной защиты (СИЗ) и средства коллективной защиты (СКЗ) от шума.

Средства коллективной защиты: устранение причин шума или существенное его ослабление в источнике образования; изоляция источников шума от окружающей среды (применение глушителей, экранов, звукопоглощающих строительных материалов); применение средств, снижающих шум и вибрацию на пути их распространения.

Одним из важных профилактических средств предупреждения усталости при действии шума является чередование периодов работы и отдыха.

Недостаточная освещенность рабочей зоны

Неудовлетворительное освещение приводит к напряжению зрения, ослаблению внимания и наступлению преждевременной утомленности. Ослепление, резь в глазах и раздражение могут быть вызваны чрезмерно ярким освещением. Свет на рабочем месте может создать сильные тени или отблески, а также дезориентировать работающего. Основным документом по требованиям к освещенности является СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95* [24].

В аудитории имеется естественное и искусственное освещение. Естественное освещение односторонние боковое. Общие освещение складывается из естественного источника света и люминесцентных ламп. В лаборатории осуществляется работа с ПК, обработка результатов.

При работе с ПЭВМ рабочие столы следует размещать таким образом, чтобы видеодисплейные терминалы были ориентированы боковой стороной к световым проемам, чтобы естественный свет падал преимущественно слева. Освещенность рабочей поверхности, создаваемая светильниками общего освещения в системе комбинированного, должна составлять не менее 10 % нормируемой для комбинированного освещения. При этом освещенность должна быть не менее 200 лк. [20]. Освещение не должно создавать бликов на поверхности экрана.

Для местного освещения рабочих мест следует использовать светильники с непросвечивающими отражателями. Светильники должны располагаться таким образом, чтобы их светящие элементы не попадали в поле зрения работающих на освещаемом рабочем месте и на других рабочих местах.

Расчёт общего равномерного искусственного освещения горизонтальной рабочей поверхности выполняется методом коэффициента светового потока, учитывающим световой поток, отражённый от потолка и стен. Длина помещения A = 6 м, ширина B = 4 м, высота H = 3,5 м. Высота рабочей поверхности над полом hp = 1,0 м. Согласно СНиП 23-05-95 необходимо создать освещенность не ниже 300 лк.

Площадь помещения:

$$S = A \times B = 6 \times 4 = 24 \text{ м}^2 \quad (1)$$

Коэффициенты отражения стен и потолка составляют соответственно $\rho_С$ = 10% и $\rho_П$ = 30%. Коэффициент запаса, учитывающий загрязнение светильника, для помещений с малым выделением пыли равен $К_З$ =1,5. Коэффициент неравномерности для люминесцентных ламп Z = 1,1.

Выбираем лампу дневного света ЛД-40, световой поток которой равен ФЛД = 2600 лм.

Выбираем светильники с люминесцентными лампами типа ОДОР-2-40. Этот светильник имеет две лампы мощностью 40 Вт каждая, длина светильника равна 1227 мм, ширина – 265 мм.

Интегральным критерием оптимальности расположения светильников является величина λ, которая для люминесцентных светильников с защитной решёткой лежит в диапазоне 1,1–1,3. Принимаем λ = 1,1, расстояние светильников от перекрытия (свес) $h_с$ = 0,3 м.

Высота светильника над рабочей поверхностью определяется по формуле:

$$h = h_п - h_р, \quad (2)$$

где $h_п$ – высота светильника над полом, высота подвеса, $h_р$ – высота рабочей поверхности над полом.

Наименьшая допустимая высота подвеса над полом для двухламповых светильников ОДОР: $h_п$ = 3,5 м.

Высота светильника над рабочей поверхностью определяется по формуле:

$$h = H - h_p - h_c = 3.5 - 1 - 0.5 = 2 \text{ м} \quad (3)$$

Расстояние между соседними светильниками или рядами определяется по формуле:2

$$L = \lambda \cdot h = 1.1 \cdot 2 = 2.2 \text{ м} \quad (4)$$

Число рядов светильников в помещении:

$$N_b = \frac{(B - \frac{2}{3}L)}{L} + 1 = \frac{(4 - \frac{2}{3} \cdot 2.2)}{2.2} + 1 = 2{,}15 \approx 2, \quad (5)$$

Число светильников в ряду:

$$N_a = \frac{(A - \frac{2}{3}L)}{l_{св} + 0.5} = \frac{(6 - \frac{2}{3} \cdot 2.2)}{1.227 + 0.5} = 2.97 \approx 3, \quad (6)$$

где у – расстояние от края ряда (м).

Общее число светильников:

$$N = N_a \cdot N_b = 3 \cdot 2 = 6 \quad (7)$$

Расстояние от крайних светильников или рядов до стены определяется по формуле:

$$l = \frac{L}{3} = \frac{2.2}{3} = 0.7 \quad (8)$$

Размещаем светильники в три ряда. На рисунке 18 изображен план помещения и размещения светильников с люминесцентными лампами.

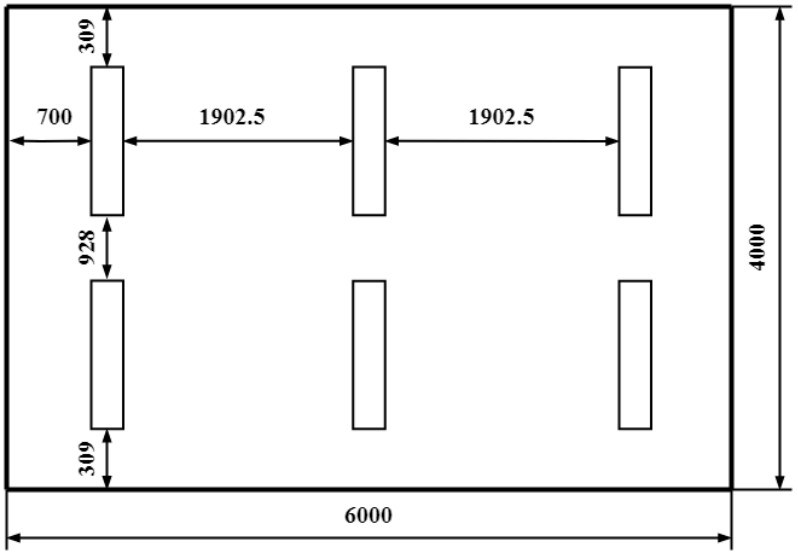


Рисунок 18 – План помещения и размещения светильников с люминесцентными лампами

Индекс помещения определяется по формуле:

$$i = \frac{A \cdot B}{h \cdot (A + B)} = \frac{6 \cdot 4}{2 \cdot (6 + 4)} = 2 \quad (9)$$

Коэффициент использования светового потока, показывающий какая часть светового потока ламп попадает на рабочую поверхность, для светильников типа ОДОР с люминесцентными лампами при $\rho_С$ = 10% и $\rho_П$ = 30% и индексе помещения i = 2 равен η = 0.6.

Световой поток группы люминесцентных ламп светильника определяется по формуле:

$$\Phi_{\text{П}} = \frac{E \cdot S \cdot K_{з} \cdot Z}{2 \cdot N \cdot \eta} = \frac{300 \cdot 24 \cdot 1.5 \cdot 1.1}{12 \cdot 0.44} = 2250\,\text{лм} \qquad (10)$$

Делаем проверку выполнения условия:

$$-10\% \le \frac{\Phi_{л} - \Phi_{п}}{\Phi_{л}} \cdot 100\% \le 20\% \qquad (11)$$

$\frac{\Phi_{л} - \Phi_{п}}{\Phi_{л}} \cdot 100\% = \frac{2600 - 2250}{2600} \cdot 100\% = 13.5\%$, что удовлетворяет условию.

Таким образом, мы получили, что необходимый световой поток не выходит за пределы требуемого диапазона. Теперь рассчитаем мощность осветительной установки: $P = 12 \cdot 40 = 480$ Вт.

Умственное перенапряжение и перенапряжение зрительного анализатора

Разработка математической модели требует значительного умственного напряжения, а также длительной работы с ЭВМ, что создает перенапряжение зрительного анализатора, рук. Эти нагрузки приводят к переутомлению функционального состояния центральной нервной системы, нервно-мышечного аппарата рук.

При пятидневной рабочей неделе и 8-ми часовой смене продолжительность обеденного перерыва составляет 30 мин, а регламентированные перерывы рекомендуется устанавливать через 2 ч от начала рабочей смены и через 2 ч после обеденного перерыва продолжительностью 5-7 мин каждый. Во время регламентированных перерывов с целью снижения нервно-эмоционального напряжения, утомления зрительного и других анализаторов целесообразно выполнять комплексы физических упражнений, включая упражнения для глаз, в первой половине смены, а в конце рабочего дня показана психологическая разгрузка в специально оборудованных помещениях [25].

Кроме того, корректно регулировать основных параметры монитора (яркость, контрастность и так далее), а также частоту обновления (при частоте

меньше 75 Гц глаза человека устают быстрее). Также благоприятнее для глаз подходят мониторы с IPS матрицей.

Воздействие переменных электромагнитных полей

Использование персонального компьютера может привести к наличию таких вредных факторов, как электромагнитные и электростатические поля. Электромагнитные поля обладает способностью биологического, специфического и теплового воздействия на организм человека.

Помещение для работы с ПЭВМ с жидкокристаллическим экраном должно соответствовать требованиям СанПиН 2.2.4.3359-16 [20]. Временные допустимые уровни ЭМП, создаваемых ПЭВМ на рабочих местах пользователей представлены в таблице 26.

Таблица 26 – Временные допустимые уровни ЭМП, создаваемых ПЭВМ на рабочих местах

Наименование параметров	Диапазон	ДУ ЭМП
Напряженность электрического поля	в диапазоне частот 5 Гц - 2 кГц	25 В/м
	в диапазоне частот 2 кГц - 400 кГц	2,5 В/м
Плотность магнитного потока	в диапазоне частот 5 Гц - 2 кГц	250 нТл
	в диапазоне частот 2 кГц - 400 кГц	25 нТл
Напряженность электростатического поля		15 кВ/м

Защита человека от опасного воздействия электромагнитного излучения осуществляется следующими способами:

– выбор рациональных режимов работы оборудования;

– рациональное размещение в рабочем помещении оборудования;

– снижение интенсивности излучения непосредственно в самом источнике излучения;

– экранирование источника.

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

Поражение током может произойти в следующих случаях:

– при прикосновении к токоведущим частям во время ремонта ПЭВМ;

– при однофазном (униполярным) касанием неизолированного человека от земли к неизолированным токоведущим частям электрических установок, находящихся под напряжением;

– при прикосновении к токоведущим частям, находящимся под напряжением, то есть в случае повреждения изоляции;

– при контакте с полом и стенами, которые оказались под напряжением;

– в случае возможного короткого замыкания в высоковольтных блоках: блок питания, блок развертки монитора.

Помещение кабинета по электробезопасности сухое, хорошо отапливаемое помещение с токонепроводящими полами, с температурой 18-21° и влажностью 40-50, поэтому оно относится к помещению без повышенной опасности, согласно ГОСТ Р 12.1.019-2017 ССБТ [26].

Нормы на допустимые токи и напряжения прикосновения в электроустановках должны устанавливаться в соответствии с предельно допустимыми уровнями воздействия на человека токов и напряжений прикосновения и утверждаться в установленном порядке.

Электробезопасность обеспечивается конструкцией электроустановок; техническими способами и средствами защиты; организационными и техническими мероприятиями.

Для обеспечения защиты от случайного прикосновения к токоведущим частям необходимо применять следующее:

– изоляция токопроводящих частей;

– защитное заземление;

– зануление;

– защитное отключение;

– предупредительная сигнализация и блокировки.

На рабочем месте запрещается прикасаться к тыльной стороне дисплея, вытирать пыль с компьютера при его включенном состоянии, работать на компьютере во влажной одежде и влажными руками.

Помимо этого, проводится ряд организационных мероприятий – специальное обучение, аттестация и переаттестация лиц электротехнического персонала, инструктажи и т. д.

4.3 Экологическая безопасность

4.3.1 Анализ влияния процесса исследования на окружающую среду

Математическая модель кредитного скоринга – является программным алгоритмом и не наносит вреда окружающей среде. С точки зрения влияния на окружающую среду можно рассмотреть влияние утилизации компьютерной техники, использованной при его разработке.

С точки зрения потребления ресурсов компьютер потребляет сравнительно небольшое количество электроэнергии, что положительным образом сказывается на общей экономии потребления электроэнергии в целом.

4.3.2 Обоснование мероприятий по защите окружающей среды

Большинство компьютерной техники содержит бериллий, кадмий, мышьяк, поливинилхлорид, ртуть, свинец, фталаты, огнезащитные составы на основе брома и редкоземельные минералы, которые не должны попадать на свалку после истечения срока использования, а должны правильно утилизироваться [27].

В списываемом имуществе могут содержаться вредные для жизни и здоровья человека вещества (ртуть, свинец и т.д.). В зависимости от степени негативного воздействия на окружающую среду оно может быть отнесено к одному из классов опасных отходов (ст. 1, 4.1 Закона № 89-ФЗ). Обезвреживание и размещение отходов I – IV классов опасности проводятся организациями, имеющими лицензию на осуществление этой деятельности.

Порядок отнесения отходов I – IV классов опасности к конкретному классу утвержден приказом Минприроды России от 05.12.2014 № 541. Приказом Росприроднадзора от 22.05.2017 № 242 утвержден федеральный классификационный каталог отходов (ФККО).

Таким образом, учреждение на основании ФККО может определить класс опасности отходов. Компьютеры (системный блок, монитор, клавиатура), утратившие потребительские свойства, относятся к IV классу опасности (малоопасные отходы). Бытовая техника, отходы деревянной офисной мебели также относятся к IV классу опасности, отходы мебели деревянной офисной (с содержанием недревесных материалов не более 10%) - к V классу опасности.

Утилизация компьютерного оборудования осуществляется по специально разработанной схеме, которая должна соблюдаться в организациях:

1. На первом этапе необходимо создать комиссию, задача которой заключается в принятии решений по списанию морально устаревшей или не рабочей техники, каждый образец рассматривается с технической точки зрения.

2. Разрабатывается приказ о списании устройств. Для проведения экспертизы привлекается квалифицированное стороннее лицо или организация.

3. Составляется акт утилизации, основанного на результатах технического анализа, который подтверждает негодность оборудования для дальнейшего применения.

4. Формируется приказ на утилизацию. Все сопутствующие расходы должны отображаться в бухгалтерии.

5. Утилизацию оргтехники обязательно должна осуществлять специализированная фирма.

6. Получается специальная официальной формы, которая подтвердит успешность уничтожения электронного мусора.

После оформления всех необходимых документов, компьютерная техника вывозится со склада на перерабатывающую фабрику. Все полученные в ходе переработки материалы вторично используются в различных производственных процессах [28].

Обращение с люминесцентными лампами в лаборатории должен осуществлять специализированный персонал, ответственный за организацию и проведение работ по сбору, хранению и утилизации отработанных ртутьсодержащих ламп. Услуги по утилизации предоставляются на основании КОГСУ (Классификация операций сектора государственного управления) и ОКПД 90.02.140149 (Общероссийский классификатор продукции по видам экономической деятельности).

Все люминесцентные лампы содержат в себе ртуть и относятся к отходам 1-го класса опасности (чрезвычайно опасные отходы).

Чтобы избежать этих последствий необходимо соблюдать определенные правила обращения с ртутьсодержащими лампами. Для образовательных учреждений эти правила перечислены в следующих нормативно-правовых документах:

1. СанПиН 2.4.2.2821-10 «Санитарно-эпидемиологические требования к условиям и организации обучения в общеобразовательных учреждениях»;

2. Постановление Правительства РФ от 3 сентября 2010 г. N 681 «Об утверждении Правил обращения с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание, транспортирование и

размещение которых может повлечь причинение вреда жизни, здоровью граждан, вреда животным, растениям и окружающей среде»;

3. КоАП РФ, Статья 8.2. Несоблюдение требований в области охраны окружающей среды при обращении с отходами производства и потребления.

Сбор отработанных ртутьсодержащих ламп у потребителей отработанных ртутьсодержащих ламп осуществляют специализированные организации.

4.4 Безопасность в чрезвычайных ситуациях

4.4.1 Анализ вероятных ЧС, которые могут возникнуть в лаборатории при проведении исследований

В лаборатории проводятся исследования с эксплуатацией электрооборудования, имеются твердые горючие материалы (столы, шкафы, ПК). Возможными причинами загорания является неправильная эксплуатация электроустановок. По степени пожароопасности помещение относится к классу П-IIа [28]. Зоны класса П-IIа – зоны, расположенные в помещениях, в которых обращаются твердые горючие вещества.

Наиболее характерной ЧС для помещения, оборудованных ЭВМ, является пожар.

Причинами возникновения данного вида ЧС являются:

– возникновением короткого замыкания в электропроводке;

– возгоранием устройств ПК из-за неисправности аппаратуры;

– возгоранием устройств искусственного освещения;

– возгоранием мебели по причине нарушения правил пожарной безопасности, а также неправильного использования дополнительных бытовых электроприборов и электроустановок.

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

Наиболее типичной чрезвычайной ситуацией для нашего объекта является пожар. Эта аварийная ситуация может возникнуть в случае короткого замыкания в проводке оборудования, обрыва провода, несоблюдения мер пожарной безопасности в офисе и т. д.

Следующие меры относятся к противопожарным мерам в помещении:

1. Помещение должно быть оборудовано: средствами тушения пожара (огнетушителями, ящиком с песком, стендом с противопожарным инвентарем); средствами связи; должна быть исправна электрическая проводка осветительных приборов и электрооборудования.

2. Каждый сотрудник должен знать место нахождения средств пожаротушения и средств связи; помнить номера телефонов для сообщения о пожаре и уметь пользоваться средствами пожаротушения.

Помещение обеспечено средствами пожаротушения в соответствии с нормами и иметь: пенный огнетушитель ОП-10 – 1 шт; углекислотный огнетушитель ОУ-5 – 1 шт.

Принудительная эвакуация при пожаре происходит в условиях усиливающегося действия опасных факторов пожара. Короткая продолжительность процесса аварийной эвакуации достигается наличием аварийных маршрутов и выходов, количество, размеры и конструктивно-планировочные решения которых регламентированы строительными нормами СНиП 2.01.02-85 [30].

Для предотвращения возникновения пожара необходимо проводить следующие профилактические работы, направленные на устранение возможных источников возникновения пожара:

– периодическая проверка проводки;

– отключение оборудования при покидании рабочего места;

– проведение с работниками инструктажа по пожарной безопасности.

Для увеличения устойчивости помещения к ЧС необходимо устанавливать системы противопожарной сигнализации, реагирующие на дым и другие продукты горения. Оборудовать помещение огнетушителями, планами эвакуации, а также назначить ответственных за противопожарную безопасность. Согласно НПБ 166-97 [31] необходимо проводить своевременную проверку огнетушителей. Два раза в год (в летний и зимний период) проводить учебные тревоги для отработки действий при пожаре.

Одними из наиболее вероятных видов чрезвычайных ситуаций являются пожар, а также взрыв на рабочем месте.

Всякий работник при обнаружении пожара должен:

1. Незамедлительно сообщить об этом в пожарную охрану;

2. Принять меры по эвакуации людей, каких-либо материальных ценностей согласно плану эвакуации;

3. Отключить электроэнергию, приступить к тушению пожара первичными средствами пожаротушения.

Лаборатория, в которой проводились исследования, оснащена ручным углекислотным огнетушителем ОУ-2, а также аптечкой первой помощи согласно требованиям ГОСТ Р 51057-01 [32].

В результате проделанной работы можно сделать вывод, что социальная ответственность – ответственность разработчика научных технологий за безопасность их применения для людей и окружающей среды, а также обеспечение безопасного проведения исследований для испытуемых.

Полученные результаты исследования правовых и организационных вопросов обеспечения безопасности, производственной, экологической безопасности и безопасности в чрезвычайных ситуациях должны быть учтены при реализации разработки математической модели кредитного скоринга. Основное внимание необходимо обратить на работу с компьютером, который используется для проведения исследования.

Заключение

В результате исследования был проведен литературный обзор по теме бинарной классификации.

В качестве исходного набора данных был использован розничный кредитный портфель банка из 25 906 наблюдений по заемщикам, включающий 42 признака.

Предварительная обработка и анализ данных выполнен с использованием Python библиотек *numpy*, *pandas*, *matplotlib*, *seaborn*, *sweetviz* и *feature_selection*, по итогам которого определен список 6 предикторов:

- просроченные месяцы;
- использование кредитной карты;
- срок с шестого месяца просрочки;
- количество обращений в суд;
- кредитные условия;
- стаж работы.

Для указанных предикторов построены 6 моделей:

- K-ближайших соседей (*K-Nearest Neighbors*)
- Метод опорных векторов (*Support Vector Machines*)
- Логистическая регрессия (*Logistic Regression*)
- Стохастический градиентный спуск (*SGD Classifier*)
- Наивный байесовский классификатор (*Gaussian Naive Bayes*)
- Дерево решений (*Decision Tree*)

Лучшей моделью оказалась логистическая регрессия, параметры которой *AUC* = 0,917, коэффициент Джини = 0,835 и *Accuracy* = 0,958.

Дальнейшие исследование были направлены на построение ансамблей моделей, лучшей комбинацией оказался ансамбль *AdaBoost* на основе деревьев решений. Данный ансамбль позволил повысить качество бинарной классификации *AUC*=0,934 (+2%), коэффициент Джини 0,868 (+3%) и *Accuracy*=0,961 (+1%).

Список использованных источников

1. Сахабиева Г. А. Скоринг как способ снижения кредитного риска // Аудит и финансовый анализ. – 2017. – № 2. – с. 125-129.

2. Побединская Т.Д. Применение кредитного скоринга для повышения эффективности управления кредитными рисками // Вестник современных исследований. – 2018. – №10.8 (25). – с. 167-170.

3. Tripathi D., Edla D.R., Cheruku R. Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification / D. Tripathi, D.R. Edla, R. Cheruku // J. Intell. Fuzzy Syst. – 2018. – №34 (3). – pp. 1543-1549.

4. Louzada F., Ara A., Fernandes G.B. Classification methods applied to credit scoring: Systematic review and overall comparison. / F. Louzada, A. Ara, G.B. Fernandes // Surveys in Operations Research and Management Science. – 2016. – vol. 21 – № 2. – pp. 117–134.

5. Barddal JP., Loezer L., Enembreck F., Lanzuolo R. Lessons learned from data stream classification applied to credit scoring / JP. Barddal, L. Loezer, F. Enembreck, R. Lanzuolo // Expert Systems with Applications. – №162. – 2020.

6. Mukid M. A., Widiharih T., Rusgiyono A., Prahutama A. Credit scoring analysis using weighted k nearest neighbor / M. A. Mukid, T. Widiharih, A. Rusgiyono, A. Prahutama // Journal of Physics Conference Series 1025. – 2018.

7. Krichene A. Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank // Journal of Economics, Finance and Administrative Science. – 2017. – vol. 22. – № 42. – pp. 3-24.

8. L. Shi, Y. Liu, X. Ma Credit Assessment with Random Forests / Shi L., Liu Y., Ma X. // Emerging Research in Artificial Intelligence and Computational Intelligence. – 2011. – pp. 24-28.

9. Волкова Е.С., Гисин В.Б., Соловьев В.И. Современные подходы к применению методов интеллектуального анализа данных в задаче кредитного скоринга // Финансы и кредит. – 2017. – т. 23, № 34. – с. 2044 – 2060.

10. Agresti, Alan An introduction to categorical data analysis / Alan Agresti. – 1996. – 394 p.

11. Logistic Regression [Электронный ресурс]. URL: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html – Machine Learning Glossary. – Дата обращения: 20.03.2021.

12. Nikolic N. et al. The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. // Expert Systems with Applications. – 2013. – vol. 40. – № 15. – pp. 5932–5944.

13. Chen N., Ribeiro B., Chen A. Financial credit risk assessment: a recent review. / N. Chen, B. Ribeiro, A. Chen // Artificial Intelligence Review. – 2016. – vol. 45. – №1. – pp. 1-23.

14. Ala'raj M., Abbod M.F. A New Hybrid Ensemble Credit Scoring Model Based on Classifiers Consensus System Approach / M. Ala'raj, M.F. Abbod // Expert Systems with Applications. – 2016. – vol.64. – pp. 36-55.

15. Ghodselahi A. A Hybrid Support Vector Machine Ensemble Model for Credit Scoring // International Journal of Computers and Applications. – 2011. – vol. 17. – №5. – pp. 1-5.

16. Kumari S., Kumar D., Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier / S. Kumari, D. Kumar, M. Mittal // International Journal of Cognitive Computing in Engineering. – 2021. – vol. 2. – pp. 40-46.

17. S. Gonzalez, S. García, J.Del Ser, L. Rokach A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities / Gonzalez S., García S., Del Ser J., Rokach L. // Information Fusion. – 2020. – vol. 60. – pp. 205-237.

17. Панин В.Ф., Сечин А.И., Федосова В.Д. Экология для инженера // под ред. проф. В.Ф. Панина. – М: Изд. Дом «Ноосфера». – 2000. – 284 с.

18. ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования

19. ГОСТ 12.2.049-80 ССБТ. Оборудование производственное. Общие эргономические требования

20. СанПиН 2.2.4.3359-16 Санитарно-эпидемиологические требования к физическим факторам на рабочих местах

21. ТОИ Р-45-084-01. Типовая инструкция по охране труда при работе на персональном компьютере

22. ГОСТ 12.0.003-2015 «Опасные и вредные производственные факторы. Классификация»

23. СН 2.2.42.1.8.562-96 Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки

24. СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*

25. МР 2.2.9.2311-07. Состояние здоровья работающих в связи с состоянием производственной среды

26. ГОСТ 12.1.019-2017 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты

27. Грязная и опасная сторона технологий [Электронный ресурс]. URL: https://www.osp.ru/pcworld/2013/06/13035804 – Дата обращения: 27.05.2021.

28. ГОСТ Р 55102-2012 Ресурсосбережение. Обращение с отходами. Руководство по безопасному сбору, хранению, транспортированию и разборке отработавшего электротехнического и электронного оборудования, за исключением ртутьсодержащих устройств и приборов

29. Правила устройства электроустановок (ПУЭ)

30. СНиП 2.01.02-85. «Противопожарные нормы»

31. НПБ 166-97. «Пожарная техника. Огнетушители. Требования к эксплуатации»

32. ГОСТ Р 51057-01. «Техника пожарная. Огнетушители переносные. Общие технические требования. Методы испытаний»

Приложение А

(справочное)

Literature Review on Credit Scoring Systems

Студент

Группа	ФИО	Подпись	Дата
0ВМ92	Шеров Ш.		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОЭФ	Семенов М.Е.	к.ф.-м.н.		

Консультант-лингвист отделения иностранных языков ШБИП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОИЯ	Утятина Я.В.			

1 Literature Review on Credit Scoring Systems

1.1 Definition of Credit Scoring

The credit risk is a crucial factor affecting the business performance of enterprises and banks. More importantly, extremely high credit risks may directly lead to the bankruptcy of enterprises or banks. Hence, it is of substantial practical significance to develop customer credit scoring models that can be used to predict possible credit default accurately and effectively.

Customer credit scoring is a binary classification problem. In recent years, scholars have used different classification models to deal with this problem, including single classification models and ensemble classification models. The commonly used single classification models in customer credit scoring include decision tree (DT), artificial neural network (ANN), naive Bayes (NB), support vector machine (SVM), logistic regression (LR) and so on.

Credit scoring is a set of decision models and their underlying techniques that aid credit lenders in the granting of credit. A broader definition of credit scoring is a numerical expression based on a level analysis of customer credit worthiness, a helpful tool for assessment and prevention of default risk, an important method in credit risk evaluation, and an active research area in financial risk management.

At the same time, the modern statistical and data mining techniques have given a significant contribution to the field of information science and are capable of building models to measure the risk level of a single customer conditioned to his characteristics, and then classify him as a good or a bad payer according to his risk level. Thus, the main idea of credit scoring models is to identify the features that influence the payment or the non-payment behavior of the customer as well as his default risk, occurring as the classification into two distinct groups characterized by the decision on the acceptance or rejection of the credit application [1].

1.2 Scoring Mathematical Models

Over the past decades, logistic regression has become the standard method of analysis in various fields where the outcome variable of interest is a discrete binary variable. Given a training set logistic regression estimates the probability of default, $p(+1|x)$ for a loan x, as follows:

$$p(+1 \mid x) = \frac{1}{1 + \exp(-(\omega_0 + \omega^t x))}, \quad (1)$$

where ω is the parameter vector and the scalar ω_0 is the intercept.

Decision tree algorithms are classification algorithms which apply a recursive partitioning on a given data set so as to come up with a tree-like structure representing patterns in underlying data by sorting them based on values of the variables present in the data. Decision trees aim at partitioning the data set into groups that are as homogeneous as possible in terms of the variable to be predicted. The C4.5 algorithm is one of the most popular ones and uses entropy to calculate the homogeneity within a sample to decide upon a partitioning. The algorithm then greedily favors splits with the largest normalized gain in entropy. The tree is then constructed by recursively repeating this procedure over the subsets created. This method often yields a complex tree structure with many internal nodes which can result in a solution that overfits the data, that is the model starts modelling the noise in the data. To counter this the algorithm prunes the resulting tree after it has been fully grown by removing nodes that have resulted from noise in the training sample.

Artificial neural networks (ANNs) are networks of simple processing elements called neurons. Neurons are simple computational units that take an arbitrary number of weighted inputs (optionally including a bias input) and are able to return a single output through an activation function. The idea of a neuron can be generalized to a multilayer perceptron (MLP) neural network by adding multiple layers containing multiple neurons to this network, where each neuron processes its inputs and generates one output value that is transmitted to all neurons in the

following layer. The basic structure of a multilayer perceptron neural network has one hidden layer and one output layer [3].

The linear regression analysis has been used in credit scoring applications even though the response variable is a two-class problem. The technique sets a linear relationship between the characteristics of borrowers $X = \{X_1, \ldots, X_p\}$ and the target variable Y, as follows,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon, \quad (2)$$

where ε is the random error and independent of X. Ordinary least squares is the traditional procedure to estimate $\beta = \beta_0, \ldots, \beta_p$, being $\hat{\beta}$ the estimated vector. Once Y is a binary variable, the conditional expectation $E(Y \mid X) = x'\beta$ may be used to segregate good borrowers and bad borrowers. Since $-\infty < x'\beta < \infty$, the output of the model cannot be interpreted as a probability.

Genetic Programming is based on mathematical global optimization as adaptive heuristic search algorithms, its formulation is inspired by mechanisms of natural selection and genetics. Basically, the main goal of a genetic algorithm is to create a population of possible answers to the problem and then submit it to the process of evolution, applying genetic operations such as crossover, mutation and reproduction. The crossover is responsible for exchanging bit strings to generate new observations.

Discriminant analysis (DA) is based on the construction of one or more linear functions involving the explanatory variables. Consequently, the general model is given by

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p, \quad (3)$$

where Z represents the discrimination score, α the intercept, β_i represents the coefficient responsible for the linear contribution of the i explanatory variable X_i, where $i = 1, 2, \ldots, p$.

This technique has the following assumptions: the covariance matrices of each classification subset are equal; each classification group follows a multivariate normal distribution.

Bayesian networks (BN) is based on calculating a posterior probability of each observation belongs to a specific class. In other words, it finds the posterior probability distribution $P(Y|X)$, where $Y = (y_1, y_2,\ldots, y_k)$ is a random variable to be classified featuring categories, and is a set of explanatory variables. A Bayesian classifier may be seen as a Bayesian network (BNs): a directed acyclic graph (DAG) represented by the triplet (N, E, P), where N are the nodes, E are the edges and P is a set of probability distributions and its parameters. In this case, the nodes represent the domain variables and edges the relations between these variables [3].

Hybrid methods combine different techniques to improve the performance capability. In general, this combination can be accomplished in several ways during the credit scoring process.

The ensemble procedure refers to methods of combining classifiers, thereby multiple techniques are applied to solve the same problem in order to boost credit scoring performance. There are three popular ensemble methods: bagging, boosting, and stacking [4].

1.3 Logistic Regression. Basic Concepts and Advantages

Probably the most common used technique for default prediction is logistic regression. It is employed in solving problems of assigning probability to an event where there is binary dependent target variable to predict. The primary difference between linear and logistic regression is the use of a binary variable as modeling target. The main reason for continuing usage of logistic regression over other methods of estimation is that it provides suitable balance of: accuracy, efficiency and interpretability of the results [7].

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

In order to map predicted values to probabilities used the sigmoid function. The function maps any real value into another value between 0 and 1 (figure 1).

$$S(z) = \frac{1}{1 + e^{-z}}, \tag{4}$$

where $S(z)$ – output between 0 and 1 (probability estimate; z – input to the function; e – base of natural log.

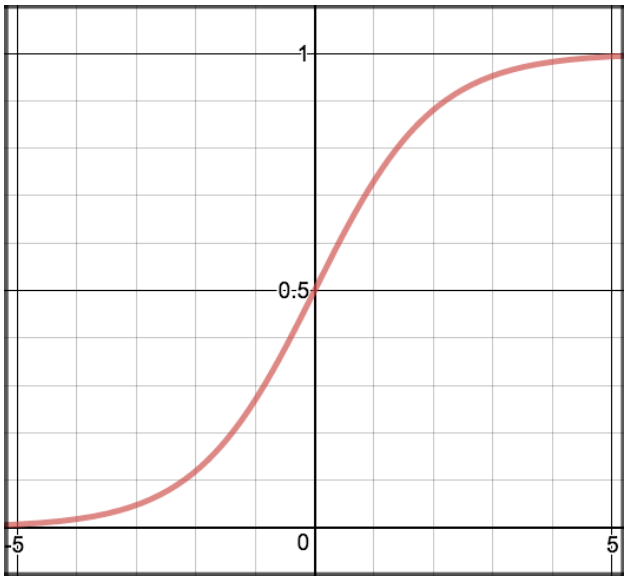


Figure 1 – Graph of sigmoid function [4]

Current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class, selected a threshold value or tipping point above which will classify values into class 1 and below which classify values into class 2.

A prediction function in logistic regression returns the probability of our observation being positive, True, or "Yes". As the probability gets closer to 1, model is more confident that the observation is in class 1.

The logistic regression coefficients are estimated on training dataset using the maximum likelihood method. The maximum likelihood method is applied by constructing a likelihood function which expresses the probability of the observed data as a function of the unknown model coefficients. The solution for the unknown coefficients is recognized when maximum likelihood function is maximized. The resulting beta coefficients will agree most closely with the observed training dataset.

Logistic regression is easy to implement, interpret, and very efficient to train. It makes no assumptions about distributions of classes in feature space. It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions. It not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative). It is very fast at classifying unknown records. Good accuracy for many simple data sets and it performs well when the dataset is linearly separable. It can interpret model coefficients as indicators of feature importance. Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets. One may consider regularization techniques to avoid overfitting in these scenarios.

A disadvantage of it is that we can't solve non-linear problems with logistic regression since its decision surface is linear. Logistic regression is also not one of the most powerful algorithms out there and can be easily outperformed by more complex ones. Another disadvantage is its high reliance on a proper presentation of your data. This means that logistic regression is not a useful tool unless you have already identified all the important independent variables. Since its outcome is discrete, Logistic regression can only predict a categorical outcome. It is also an algorithm that is known for its vulnerability to overfitting [4].

1.4 Credit Scoring Performance Evaluation Criteria

Performance evaluation criteria, such as the confusion matrix or the Average Correct Classification (ACC) rate, the estimated misclassification cost, mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), the receiver operating characteristics (ROC) curve, GINI coefficient, and other criteria are all used in credit scoring applications under different fields. The following is a discussion of some of these performance criteria.

Confusion matrix (average correct classification rate criterion) is one of the most widely used criteria in the area of accounting and finance (for credit scoring applications) in particular, and other fields, such as marketing and health in general.

The average correct classification rate measures the proportion of the correctly classified cases as good credit and as bad credit in a particular data-set. The average correct classification rate is a significant criterion in evaluating the classification capability of the proposed scoring models. The idea of correct classification rates comes from a matrix, which is occasionally called “a confusion matrix”, otherwise called a classification matrix. A classification matrix presents the combinations of the number of actual and predicted observations in a data-set [13].

A traditional procedure is to build a confusion matrix, as shown in Table 1, where *M* is the model prediction, *D* is the real value in dataset, *TP* the number of true positives, *FP* the number of false positives, *FN* the number of false negatives and *TN* the number of true negatives. Naturally, $TP+FP+FN+TN = N$, where *N* is the number of observations. Through the confusion matrix, some measures are employed to evaluate the performance on test samples.

Table 1 – Confusion matrix

D	*M*	
	1	0
1	*TP*	*FP*
0	*FN*	*TN*

Accuracy (*ACC*): the ratio of correct predictions of a model, when classifying cases into class {1} or {0}.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Sensitivity (*SEN*): also known as Recall or True Positive Rate is the fraction of the cases that the technique correctly classified to the class {1} among all cases belonging to the class {1}.

$$SEN = \frac{TP}{TP + FN} \tag{6}$$

Specificity (*SPE)*: also known as True Negative Rate is the ratio of observations correctly classified by the model into the class {0} among all cases belonging to the class {0}.

$$SPE = \frac{TN}{TN + FP} \tag{7}$$

Precision (*PRE*): is the fraction obtained as the number of true positives divided by the total number of instances labeled as positive.

$$PRE = \frac{TN}{TN + FP} \tag{8}$$

False Negative Rate (*FNR*) also known as Type I Error is the fraction of {0} cases misclassified as belonging to the {1} class.

$$FNR = \frac{FN}{TP + FN} \tag{9}$$

False Positive Rate (*FPR*) also known as Type II Error is the fraction of {1} cases misclassified as belonging to the {0} class.

$$FPR = \frac{FP}{TN + FP} \tag{10}$$

Other traditional measures used in credit scoring analysis are *F*-Measure and two-sample *K–S* value. The *F*-Measure combines both Precision and Recall, while the *K–S* value is used to measure the maximum distance between the distribution functions of the scores of the «good payers» and «bad payers».

The receiver operating characteristic curve (ROC curve) may be geometrically defined as a graph for visualizing the performance of a binary classifier technique. The ROC curve is obtained by measuring the 1-specificity on the first axis and measured the sensitivity on the second axis, creating a plane. Therefore, the more the curve distances from the main diagonal, the better is the model performance. Figure 2 shows an example of ROC Curve [4].

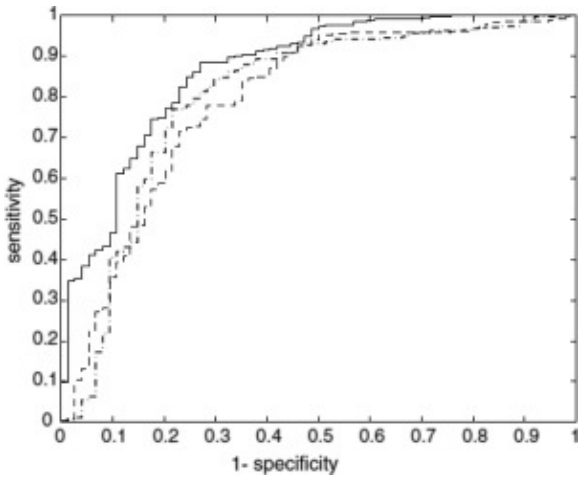


Figure 2 – ROC Curve [4]

AUC (the area under the ROC curve) provides a single value performance measure [0; 1] which quantifies the predictive power of the model. For perfect discrimination AUC value is one. For a credit scoring which has purely random discrimination AUC value is 0.5. For any other scorecard, the AUC curve will be between 0.5 and 1 and the larger the area the better the discrimination. Any value less than 0.5 implies that the model is getting it wrong with some consistency and zero means the predictions are perfectly wrong.

GINI coefficient corresponds to twice the area under the ROC curve as GINI = 2 × AUC − 1. Useless and perfect discriminations are quantified by 0 and 1, respectively. Another possibility to calculate the model performance using the related GINI coefficient is done by Trapezium rule. GINI also enables comparing discrimination power results directly with other studies. By this reason we considered it to be main criterion for choosing the final prediction model.

Kolmogorov–Smirnov (*K–S*) statistic is calculated using the cumulative distribution functions (CDFs) of financial ratio values. Firstly, CDFs are computed separately for goods and bads and the distance between these CDFs is captured using the formula $D = \max|CDS_{goods} - CDS_{bads}|$. *K–S* takes values between 0 and 1 with higher values indicating stronger discriminatory power [12].

1.5 Credit Scoring Software

Statistical software packages SAS, SPSS, Statistica are widely used for software implementation of scoring algorithms. But due to the limited capabilities of these commercial packages in recent years, they have been replaced by the use of open programming environments such as Python, R. In them, more algorithms can be implemented, existing mathematical models can be updated and improved, the use of many libraries significantly increases scoring capabilities, as well as the speed of processing a large amount of data.

For the pre-processing step, Python and R proved to be extremely efficient. Many of the operations could be performed using built-in functions. The scripting nature of these languages also allowed to directly interact with the dataset.

In Python, the data is read from csv files using the *pandas* library, which provides a solution for manipulating tabular data. Analyzing the dataset could directly be performed using simple commands. Python also offers some powerful graphical libraries and plots of variables are also easily generated in one line of code.

For quick data exploration and reports, R allows to get real fast visualizations reports thanks to the libraries *DataExplorer* and esquisse, and to get a touch of the dataset. With only one line of code, we can derive a detailed report to get a first intuition on our data and know where to focus on during the pre-processing phase.

The pre-processing is a bit more tedious in SAS. On the other hand, if the data initially is in SAS, it gives a strong advantage to SAS, since it eases the data flows. For instance, it allows to quickly reparametrize the model with new data, by simply connecting to the relevant datasets owned by the business.

SAS directly provides an algorithm for the stepwise selection with *PROC LOGISTIC*. The algorithm is efficient and all the necessary statistics are included. The stepwise selection is however fully automated and only based on significance levels. What is more, overfitting risk is often reduced using other statistical metrics (such as Bayesian or Akaike information criterions) and expert knowledge (expected signs for instance).

Different modules are providing logistic regressions in Python. The *scikit-learn* module is primarily a machine learning package, and only provides the implementation of regularized logistic regressions. While some basic indicators are directly available, p-values have to be recalculated by the user. For these reasons, the *statsmodels* module was preferred. The module is more focused on traditional statistics, and indeed provides all the classical statistics for a logistic regression. However, the *statsmodels* module does not directly provide a function for calculating the *AUC*.

R provides a large range of package for modelling GLMs. The two most commonly used libraries are stats and *glmnet*. The summary function returns the regression diagnostic. Strongly significant variables are highlighted with a three-star grade. Statistics of the residuals are also returned, as well as the model deviance. Finally, a dedicated function produces a comparative table of different models for the user to pick the best one [10].