TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

Школа Инженерная школа ядерных технологий
Направление подготовки 01.03.02 «Прикладная математика и информатика»
Отделение школы (НОЦ) Отделение экспериментальной физики

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Построение моделей кредитного скоринга с учетом несбалансированности данных

УДК 330.4:519.876:336.77.067

Студент

Группа	ФИО	Подпись	Дата
ОВ8Б	Осорова Милена Николаевна		

Руководитель ВКР

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

Со-руководитель (по разделу «Концепция стартап-проекта»)

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ШИП	Спицын Владислав Владимирович	к.э.н.		

КОНСУЛЬТАНТЫ:

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Сечин Андрей Александрович	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Руководитель ООП 01.03.02 «Прикладная математика и информатика»	Крицкий Олег Леонидович	к.ф.-м.н.		

Томск – 2022 г.

Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение

высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Инженерная школа ядерных технологий
Направление подготовки 01.03.02 «Прикладная математика и информатика»
Отделение экспериментальной физики

УТВЕРЖДАЮ:

Руководитель ООП

_____ _____ Крицкий О.Л.

(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

Студенту:

Группа	ФИО
0В8Б	Осоровой Милене Николаевне

Тема работы:

Построение моделей кредитного скоринга с учетом несбалансированности данных	
Утверждена приказом директора (дата, номер)	

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	*Набор данных, представленных индустриальным партнером ООО «Эко-Томск». Набор состоит из розничного кредитного портфеля банка из 25 465 наблюдений по заемщикам, коэффициент дисбаланса равен 18. Каждый заемщик характеризуется 37 различными параметрами,*

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. *Обзор литературы;* 2. *Реализация предварительной обработки данных и отбор наиболее значимых признаков;* 3. *Построить модели кредитного скоринга и их ансамбли для оценки вероятности дефолта заемщика по несбалансированной и сбалансированной выборке;* 4. *Реализация программного продукта на языке программирования;* 5. *Подготовка и проведение тестирования программы;* 6. *Сравнение метрик качества построенных моделей;* 7. *Анализ влияния дисбаланса на точность модели кредитного скоринга;* 8. *Выбор оптимальной модели кредитного скоринга.*
Перечень графического материала	1. *Скриншоты работы программного продукта;* 2. *Блок-схема алгоритма работы программы;* 3. *Визуализация значений метрик качества, матриц ошибок, ROC-кривых.*

Консультанты по разделам выпускной квалификационной работы

(если необходимо, с указанием разделов)

Раздел	Консультант
Концепция стартап-проекта	Спицын Владислав Владимирович
Социальная ответственность	Сечин Андрей Александрович

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

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

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0В8Б	Осорова Милена Николаевна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«КОНЦЕПЦИЯ СТАРТАП-ПРОЕКТА»

Студенту:

Группа	ФИО
0В8Б	Осорова Милена Николаевна

Школа	ИЯТШ	Направление	01.03.02 «Прикладная математика и информатика»
Уровень образования	Бакалавриат		

Перечень вопросов, подлежащих разработке:	
Проблема конечного потребителя, которую решает продукт, который создается в результате выполнения НИОКР (функциональное назначение, основные потребительские качества)	Увеличение дохода за счет решения проблемы финансовых убытков из-за невозвратных займов и увеличения объема «хороших» заемщиков
Способы защиты интеллектуальной собственности	Патентование программы для ЭВМ, регистрация товарного знака и домена, приобретение исключительного права на сайт, депонирование его страниц и использование на сайте знака Copyright.
Объем и емкость рынка	Объем рынка услуг оценки кредитоспособности заемщика: 7,2 млн. руб/год
Современное состояние и перспективы отрасли, к которой принадлежит представленный в ВКР продукт	Прогнозируется рост спроса на продукт
Себестоимость продукта	141190 рублей
Конкурентные преимущества создаваемого продукта	Индивидуальная настройка моделей, быстрая интеграция, онлайн-услуга
Сравнение технико-экономических характеристик продукта с отечественными и мировыми аналогами	На основании конкурентных преимуществ
Целевые сегменты потребителей создаваемого продукта	Микрофинансовые организации
Бизнес-модель проекта	Модель по А. Остервальду и схема продаж по SaaS-модели
Производственный план	В первый год: 14 компаний
План продаж	Чистая прибыль в первый год продаж: около 1,01 млн. рублей
Перечень графического материала:	
При необходимости представить эскизные графические материалы(например, бизнес-модель)	Бизнес-модель по А. Остервальдеру, таблицы расчета юнит-экономики, таблица конкурентного анализа, таблицы плана продаж и оценки прибыли

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант по разделу «Концепция стартап-проекта» (со-руководитель ВКР):

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ШИП	Спицын Владислав Владимирович	к.э.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
ОВ8Б	Осорова Милена Николаевна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО		
0В8Б	Осорова Милена Николаевна		
Школа	**Инженерная школа ядерных технологий**	**Отделение (НОЦ)**	**Отделение экспериментальной физики**
Уровень образования	Бакалавриат	**Направление/ специальность**	01.03.02 Прикладная математика и информатика

Тема ВКР:

Построение моделей кредитного скоринга с учетом несбалансированности данных	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места).	*Объект исследования:* математические модели кредитного скоринга с учетом несбалансированности данных; *Область применения:* принятие решений в кредитных организациях; *Рабочая зона:* офисное помещение; *Размеры помещения:* 27,2 м^2 *Количество и наименование оборудования рабочей зоны:* 1 персональный компьютер *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне:* алгоритмическая и программная разработка с использованием персонального компьютера
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.	– Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 25.02.2022); – 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. – ГОСТ 21889-76. Система «человек-машина». Кресло-человека-оператора. Общие эргономические требования. – СанПиН 2.2.4.3359-16. Санитарно-эпидемиологические требования к физическим факторам на рабочих местах
2. Производственная безопасность: – Анализ выявленных вредных и опасных производственных факторов	**Вредные и опасные факторы:** – недостаточная освещенность рабочей зоны; – перенапряжение зрительных анализаторов; – отклонение показателей микроклимата; – повышенный уровень шума; – повышенный уровень статического электричества; – повышенное значение напряжения в электрической цепи, замыкание;

3. Экологическая безопасность:	– анализ воздействия при работе на ПЭВМ на атмосферу, гидросферу и литосферу; – наличие отходов (компьютерная техника, бумага, канцелярия и т.д.); – методы утилизации отходов.
4. Безопасность в чрезвычайных ситуациях:	*Возможные ЧС:* пожар на рабочем месте *Наиболее типичная ЧС:* пожар на рабочем месте
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Сечин Андрей Александрович	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0В8Б	Осорова Милена Николаевна		

Реферат

Выпускная квалификационная работа содержит 99 страниц, 32 рисунка, 30 таблиц и 35 источников.

Ключевые слова: кредитный скоринг, несбалансированные данные, машинное обучение, бинарная классификация, ансамблевые модели.

Объектом исследования являются математические модели кредитного скоринга.

Цель работы – построение оптимальной модели кредитного скоринга с учетом несбалансированности данных.

Методы проведения исследования: теоретические (изучение литературы, обзор методов построения моделей) и практическое (применение методов для построения моделей).

В процессе исследования проводились обзор литературы, предварительная обработка данных, отбор наиболее значимых признаков для прогнозирования, построение моделей кредитного скоринга и их ансамблей по несбалансированным и сбалансированным данным, количественная оценка адекватности построенных моделей, их сравнение, анализ влияния дисбаланса данных и выбор оптимальной модели.

Бакалаврская работа написана в Microsoft Word 2016. Программный код моделей кредитного скоринга разработан на языке программирования Python, облачный сервис Colab Google/Yandex.DataSphere.

Область применения: банки, кредитные организации, маркетплейсы.

Экономическая эффективность/значимость работы: значительное снижение экономических потерь из-за невозвратных кредитов и увеличения количества одобренных кредитов.

Оглавление

Введение

Основой деятельности для банков и кредитных организаций являются услуги кредитования. Одним из решающих факторов при выдаче кредита является оценка кредитных рисков. Стабильность банка зависит от качества выданных кредитов, поэтому банк, способный наиболее точно «отфильтровать» клиентов с высоким кредитным риском, сможет предложить рынку более привлекательные кредитные продукты с оптимальными кредитными ставками и, следовательно, получить определенные конкурентные преимущества. Для этого нужно выбрать “лучшую” модель прогнозирования дефолта для кредитного скоринга. При этом нужно иметь ввиду, что данных о дефолтах намного меньше чем о тех, кто вернул кредит. Таким образом, при разработке математической модели приходится учитывать особенность исходных данных, а именно их несбалансированность.

Методика кредитного скоринга использует математическую модель, которая позволяет решить задачу бинарной классификации, так как наблюдаемые объекты относятся к одному из двух классов. Существует несколько методов реализации бинарной классификации, такие как: логистическая регрессия, метод k-ближайших соседей, метод опорных векторов, дерево решений, а также ансамблевые методы на основе вышеприведенных методов: голосование, бэггинг, случайный лес, метод градиентного бустинга, метод адаптивного бустинга и др. Они в свою очередь являются методами машинного обучения. Для решения проблемы дисбаланса существуют методы корректировки выборки (сэмплирование), с помощью которых решается проблема дисбаланса: недосэмплирование (уменьшение большего класса), пересэмплирование (увеличение малого класса) и комбинирование вышеуказанных методов. Также проблему работы с несбалансированными данными можно решить с помощью использования ансамблевых методов (объединение нескольких базовых классификаторов, которые будут компенсировать ошибки базовых моделей). Отметим, что на

практике эффективнее оказалось использование гибридных методов, объединяющих преимущества классификаторов и ансамблевого подхода.

Объектом исследования являются математические модели кредитного скоринга.

Цель работы – построение оптимальной модели кредитного скоринга с учетом несбалансированности данных.

Работа включает следующие задачи:

- обзор литературы и описание математической составляющей;
- предварительную обработку данных и отбор наиболее значимых признаков для прогнозирования дефолта заемщика;
- построение моделей и их ансамблей для оценки вероятности дефолта заемщика по несбалансированным и сбалансированным данным;
- количественная оценка адекватности построенных моделей по метрикам качества и их сравнение;
- анализ влияния дисбаланса на точность модели кредитного скоринга;
- выбор оптимальной модели.

1. Обзор литературы

1.1. Понятие кредитного скоринга

Кредитная оценка – это процедура выявления платежеспособности потенциального заемщика и рисков непогашения обязательств в будущем, которая включает в себя сбор, анализ и выявление риска. Качественная работа данной процедуры является одним из ключевых факторов, влияющих на конкурентоспособность, стабильность и прибыльность банков и кредитных организаций. Самым распространенным и объективным методом кредитной оценки является *кредитный скоринг*.

Существуют разные определения кредитного скоринга.

В статье авторов Алешина В.А. и Рудаевой О.О. написано, что кредитный скоринг (скоринг заявителя) — это вид скоринга, являющийся алгоритмом или методикой, позволяющая на основе данных о потенциальном заемщике оценить вероятность того, что новый клиент не выплатит кредит. То есть по сути это является оценкой риска невозврата кредита заемщиком. Данная вероятность позволяет повысить качество банковского риск-менеджмента за счет контроля параметра процентов невозврата кредитов.

Также в данной работе сделаны выводы о том как влияет применение кредитного скоринга на банк и что позволяет улучшить: «увеличить кредитный портфель за счет уменьшения количества необоснованных отказов по кредитным заявкам; повысить точность оценки заемщика; уменьшить уровень невозвратов; ускорить процедуру оценки заемщика; создать централизованное накопление данных о заемщиках; снизить формируемые резервы на возможные потери по кредитным обязательствам; быстро и качественно оценить динамику изменений кредитного счета индивидуального заемщика и кредитного портфеля в целом» [1].

В дополнение к предыдущему определению можно рассмотреть следующее понятие, представленным в работе Никаненковой В.В.: «Скоринг представляет собой математическую или статистическую модель, с помощью

которой на основе кредитной истории «бывших» заемщиков банк пытается определить, насколько высока вероятность, что потенциальный заемщик вернет (либо не вернет) кредит в срок» [2].

Обобщая приведенные определения, можно сказать, что кредитный скоринг позволяет определить вероятность невозврата или возврата кредита путем использования статистических и математических моделей.

1.2. Несбалансированные данные

В статье [3] рассмотрены особенности работы с несбалансированными данными в задачах классификации (бинарной). Такая проблема часто встречается в задачах с очень редко наблюдаемыми или диагностируемыми явлениями (например, кредитный скоринг, болезни, мошенничество и т.п.).

Данные называются *несбалансированными*, если в обучающей выборке доли объектов разных классов существенно различаются. Проблему дисбаланса требуется решить, если оно влияет на результат классификации. Для этого необходимо сначала выявить причину дисбаланса [3]: иногда бывает, что для решения проблемы нужно всего лишь проанализировать и обработать данные или же дисбаланс меняется со временем. Рассмотрим случай, когда пропорции классов остаются неизменными с течением времени и к данным применена необходимая обработка.

В [3] описаны следующие методы корректировки выборки (сэмплирование), с помощью которых решается проблема дисбаланса: недосэмплирование (простой, NearMiss-1, NearMiss-2, Tomek links, Edited nearest neighbors (ENN)) и пересэмплирование (простой, SMOTE, ADASYN).

Недосэмплирование – это метод сэмпирования, когда б**о**льший класс заменяется своей подвыборкой с мощностью как малый класс. А *пересэмплирование* – это метод, в котором малый класс увеличивается до мощности б**о**льшего класса. Но также автор [3] отметил, что данные можно предварительно не обрабатывать, если выбрать подходящий функционал

качества (метрика качества), порог бинаризации или настроить веса объектов. *Метрика качества* отвечает за точность оценки распределения объектов к какому-то из классов. *Порог бинаризации* отвечает за распределение результатов классификации к определенному классу.

В работе [4] указано, что проблему работы с несбалансированными данными также можно решить с помощью использования ансамблевых методов, т.е. с помощью создания сильного классификатора путем объединения нескольких базовых классификаторов, которые будут компенсировать ошибки базовых моделей. Но также отмечено, что эффективнее, когда используется гибридные методы, такие как объединение описанных выше методов и ансамблевого метода. Например, в данной статье автор предложил алгоритм ансамбля с взвешенной гибридной выборкой. То есть, сначала необходимо сбалансировать исходную выборку двумя способами сэмплирования (сначала пересэмплирование, потом недосэмплирование) и по ним провести обучение ансамблевой модели.

Авторы [3, 4] сходятся во мнении, что из метрик качества следует рассматривать F-score – среднее гармоническое точности и полноты, так как максимизация данного показателя приводит к максимизации и точности и полноты.

1.3. Ансамблевые методы

В источнике [5] подробно рассмотрены различные методы ансамблевого обучения, такие как бэггинг, бустинг и стекинг. Описаны различия между методами, алгоритм их работы и приведены основные термины.

Ансамблевые методы – это алгоритм объединения нескольких моделей (слабых учеников или базовых моделей) в одну метамодель для получения лучших результатов. Цель ансамблевой модели (метамодели) состоит в уменьшении смещения и/или разброса базовых моделей. Существуют три основных типа мета-алгоритмов.

Бэггинг – обучение однородных (т.е. одного) базовых моделей происходит параллельно, потом они объединяются и усредняются по некоторому детерминированному процессу. Цель бэггинга состоит в получении модели с меньшим разбросом (дисперсией). Для него нужно ввести также термин *бутстрэп* – метод генерации выборок указанного размера из исходной выборки путем случайного выбора элементов с повторениями в каждом из наблюдений. Базовая модель обучается на каждой такой бутстрэп выборке, потом результаты объединяются и общее усредняется. Например, случайный лес также является методом бэггинга, он состоит из глубоких (глубиной в множество узлов) деревьев решений.

Бустинг – обучение однородных базовых моделей происходит последовательно, и они объединяются по детерминированной стратегии. Цель бустинга состоит в создании модели с меньшим смещением. Базовые модели для создания метамодели должны обладать низким разбросом и высоким смещением. Существуют несколько алгоритмов бустинга, в данной работе будут рассмотрены: адаптивный (Adaboost) и градиентный. Оба алгоритма являются итеративными.

Стекинг – обучение разнородных базовых моделей происходит параллельно, и они объединяются, обучаясь для вывода прогноза, основываясь на предсказаниях различных базовых моделей.

Также существует еще один способ объединения независимых моделей в одну метамодель – *голосование* (voting). Его суть состоит в том, что он усредняет выходные результаты отдельных моделей для получения общего вывода метамодели. Это можно сделать по двум способам: мажоритарным или мягким голосованием. *Мажоритарное голосование* (hard voting) в качестве ответа метамодели выводит тот класс, который встречался больше всех среди выводов отдельных моделей. А *мягкое голосование* (soft voting) рассматривает вероятности принадлежности к какому-либо классу по отдельным моделям,

усредняет их и по значению полученной вероятности определяет метку класса, полученной на выходе метамодели.

2. Теоретическая часть

2.1. Математические модели кредитного скоринга

В настоящее время существуют различные математические модели кредитного скоринга. Методика кредитного скоринга использует математическую модель, которая позволяет решить задачу классификации. *Классификация* – это задача машинного обучения, которая заключается в соотношении наблюдаемого объекта к одному из нескольких известных классов. В случае двух классов (например, 0 или 1; хороший или плохой; кошка или собака и т.д.) и решается вопрос к какому классу отнести изучаемый объект, то подразумевается *бинарная классификация* [6].

Существует несколько методов бинарной классификации, из них рассмотрим следующие: из базовых классификаторов – *логистическая регрессия, метод k-ближайших соседей, метод опорных векторов, дерево решений,* из ансамблевых методов – *случайный лес, метод градиентного бустинга и метод адаптивного бустинга.* Они в свою очередь являются методами машинного обучения. Стоит отметить, что в работе приведены реализации как вышеперечисленных моделей, так и метод бэггинга базовых моделей и метод голосования.

2.1.1. Логистическая регрессия

Логистическая регрессия получает оценку вероятности принадлежности объекта к классу путем построения линейного классификатора. Данную модель следует использовать, если требуется решить задачу бинарной классификации и характеристические признаки независимы друг от друга. Данную модель часто используют в задаче кредитного скоринга, так как в

качестве зависимой переменной выступает бинарная переменная. Логистическая регрессия является статистической моделью [7].

Пусть у нас есть n признаков $f_i: X \to R, j = 1, \dots, n$, которые описывают заемщиков. Следовательно, $X = R^n$ – пространство признаков. Пусть Y – конечное пространство номеров классов ($Y = \{-1, +1\}$). Возьмем обучающую выборку, состоящую из m – наблюдений: $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$. Тогда линейный классификатор имеет следующий вид [8]:

$$a(x, \omega) = sign\left(\sum_{j=1}^{n} \omega_j f_j(x) - \omega_0\right) = sign\langle x, \omega\rangle,$$

где ω_j – вес j-го признака, ω_0 – порог принятия решения, $\omega = (\omega_0, \dots, \omega_n)$ – вектор весов, $\langle x, \omega\rangle$ – скалярное произведение признакового описания объекта на вектор весов.

Обучение линейного классификатора выполняется по обучающей выборке с целью настройки вектора весов ω признаков заемщиков. Этот процесс называется *задачей минимизации* весов и выполняется следующим образом [7]:

$$Q(x, y, \omega) = \sum_{i=1}^{m} ln\ ln\ (1 + exp\ exp\ (-y_i\langle x_i, \omega\rangle)) \to min$$

После получения решения задачи нужно провести классификацию $a(x, \omega) = sign\langle x, \omega\rangle$ и результат привести к виду вероятности $P(x)$ принадлежности к тому или иному классу с помощью сигмоидной функции [11]:

$$P(x) = \frac{1}{1 + e^{-(y\langle x, \omega\rangle)}}.$$

2.1.2. Метод k-ближайших соседей (K-Nearest Neighbors)

Метод k-ближайших соседей является метрическим методом (используются для построения классификаторов на основе машинного обучения). Суть данного алгоритма состоит в том, что вокруг исследуемого

объекта выбираются k-ближайших соседей и анализируя их, определяется к какому классу относится объект. Для данной модели следует изначально нормировать и сбалансировать обучающие данные [8].

Пусть имеется набор из n наблюдений $X_i(i = 1, \dots, n)$, которые описывают заемщиков и имеется $C_i(j = 1, \dots, m)$ классов. Необходимо выбрать правило, по которому объекты относятся к какому-либо классу. Для этого используется *функция сочетания*, которая делится на два типа. Простое невзвешенное голосование – расстояние до классифицируемого объекта не учитывается. Взвешенное голосование – расстояние до классифицируемого объекта учитывается и вводится взвешивание объектов в зависимости от расстояния.

Взвешенное голосование вводит штрафы для класса, учитывая удаленность соседа от классифицируемого объекта. Данный штраф называется *показателем близости* и вычисляется как сумма следующих величин:

$$Q_j = \sum_{j=1}^{n_j} \frac{1}{D^2(x, a_{ij})},$$

где D – функция вычисления расстояния, x – вектор признаков классифицируемого объекта, a_{ij} – i-ый пример j-го класса. Чем больше величина Q_j для одного из нескольких классов, тот и является искомым.

Функция вычисления расстояния выбираются в зависимости от данных, например, евклидова метрика, расстояние Хэмминга.

Заметим, что для того, чтобы расстояния между объектами не искажались, нужно перед применением метода нормализовать (масштабировать) данные [8].

2.1.3. Метод опорных векторов (Support Vector Machines)

Метод опорных векторов изначально был предназначен для бинарной классификации, поэтому именно для такого рода задачи он работает очень хорошо. Данный метод также является метрическим. Суть метода состоит в построении на плоскости с объектами линии или гиперплоскости, относительно которых объекты группируются в классы.

Пусть имеется выборка $\{(x_i, y_i)\}(i = 1, \dots, m), x_i \in R^n\ y_i \in \{-1,1\}$.

В данном методе классифицирующая функция имеет вид $F(x) = sign(\langle w, x\rangle + b)$, где $\langle\ ,\rangle$ — скалярное произведение, w — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр.

Если $F(\boldsymbol{x})$ = 1, то объект относится к одному классу, а при $F(\boldsymbol{x})$ = -1 – к другому. Далее нужно выбрать такие w и b, чтобы расстояние до каждого класса было максимальным. Это расстояние равно $\frac{1}{\|w\|}$. Поиск этого расстояния сходится к поиску минимума $\|w\|$ – задача оптимизации, которая решается с помощью множителей Лагранжа [9]:

$$\{arg\ \|w\|^2\ \ y_i(\langle w, x_i\rangle + b) \geq 1, \qquad i = 1, \dots, m$$

Часто бывает так, что невозможно разделить выборку линейно, тогда элементы выборок преобразуют с помощью отображения

$$\varphi\colon R^n \to X$$

так, чтобы в пространстве X выборку можно было разделить линейно: $F(x) = sign(\langle w, \varphi(x)\rangle + b)$. Отображение $\varphi(x)$ выбирается таким образом, чтобы выполнялось следующее выражение: положительно определенная и симметричная функция двух переменных – ядро - $k(x, x') = \langle \varphi(x), \varphi(x')\rangle$.

2.1.4. Дерево решений (Decision Tree)

В данном случае будем рассматривать дерево решений как метод классификации. Суть заключается в построении правил для принятия решений по исходным данным. То есть, цель дерева решений состоит в составлении

наборов логических условий «если-то», по которым классифицируются объекты. В дереве решений есть следующие понятия: объект – пример, шаблон, наблюдение; атрибут – признак, независимая переменная, свойство; целевая переменная – метка класса, зависимая переменная; узел – внутренний узел дерева, узел проверки условий; корневой узел – начальный узел дерева решений, лист – конечный узел дерева, метка класса; решающее правило – условие в узле, проверка [10].

Перед построением дерева решений стоит следующая задача: какой атрибут нужно выбрать первым для проверки условия, то есть какое решающее правило будет стоять в корневом узле. При последующем добавлении узлов данную задача выбора атрибута также нужно решать. При построении дерева решений нужно с каждым последующим добавлением узлов уменьшать неопределенность, то есть увеличивать прирост информации. Для этого используются два способа, их еще называют *критериями качества разбиения*: критерий прироста информации и неопределенность Джини. Также нужно определить критерий остановки обучения, чтобы дерево решений не переобучалось. Обычно для этого задают глубину деревьев решений или минимальное допустимое число примеров в узле.

Пусть имеется обучающий набор S, состоящий из n примеров, для каждого задана метка класса C_i $(i = 1, 2,\ldots, k)$ и m атрибутов A_j $(j = 1, 2,\ldots, m)$. Возможны три случая: а) все примеры относятся к одному классу, тогда смысла в обучении нет из-за принадлежности примеров только к одному классу; б) примеров совсем нет; в) множество с примерами содержит все метки классов C_k. Представим алгоритм выбора атрибутов для данного случая.

Выбирается один из атрибутов A_j с двумя и более уникальными значениями. Выбор данного атрибута происходит по критерию качества разбиения.

а) Критерий прироста информации $Gain(A)$.

$$Gain(A) = Info(S) - Info(S_A),$$

где $Info(S)$ – информация до разбиения подмножества S, $Info(S_A)$ – информация после разбиения по атрибуту A. Этот критерий основан на уменьшении энтропии Шеннона:

$$H = -\sum_i^n \frac{N_i}{N} log \frac{N_i}{N},$$

где n – количество меток класса, N_i – число примеров i-го класса, N – общее число примеров в подмножестве. Задача выбора атрибутов состоит в максимизации $Gain(A)$.

$$Info(S) = H,\ Info(S_A) = \sum_i^n \frac{N_i}{N} H_i.$$

b) Неопределенность Джини *Gini*.

$$Gini = 1 - \sum_i^n (\frac{N_i}{N})^2.$$

Задача выбора атрибутов состоит в минимизации *Gini*.

Выбор атрибутов останавливается, если подмножество пустое или подмножество принадлежит только одному классу [10].

2.1.5. Случайный лес (Random Forest)

Случайный лес является представителем бэггинга. Суть метода заключается в объединении множества деревьев решений в одну модель (композиция). Случайный лес основан на том, что объединение множества деревьев будет в любом случае работать лучше, чем одно дерево решений. Это связано с тем, что отдельные деревья решений обучаются по разным данным и за счет этого получаются разнообразные результаты и эти результаты в конце объединяются и выдают один общий результат. Также важным фактором является то, что такие деревья слабо коррелированы друг с другом [17].

Пусть случайный лес состоит из N деревьев. Тогда алгоритм имеет следующий вид [11]:

Для каждого дерева $n = 1, 2, \ldots, N$:

1) Сгенерируем выборку X_n с помощью бутстрэпа.

2) Построим решающее дерево b_n по выборке X_n:

a) Выберем лучший признак по заданному критерию, далее сформируем разбиение в дереве по нему и так до конца выборки

b) Дерево будем строить, пока в каждом листе не более n_{min} объектов или пока не будет достигнута определенная высота дерева. Под *высотой дерева* будем понимать максимальную длину пути от корня до листа.

c) При каждом разбиении сначала выбирается m случайных признаков из n исходных и оптимальное разделение выборки ищется среди них.

3) В итоге получаем предсказание путем выбора наиболее часто встречаемого класса.

2.1.6. Метод градиентного бустинга (Gradient Boosting)

Градиентный бустинг является представителем ансамблевых методов. Суть метода состоит в итеративном обучении базовых моделей с целью минимизации функции потерь по их градиенту.

Пусть имеется следующий набор пар признаков *x* и целевых переменных *y*: $\{(x_i, y_i)\}(i = 1,2, \dots, n)$, на которых нужно восстанавливать зависимость вида $y = f(x)$. Но восстанавливаться будет приближение $\hat{f}(x)$ и для оценки этого приближения будет использоваться функция потери $L(y, f)$, которую нужно минимизировать.

Предположим, что $\hat{f}(x) = \sum_{i=0}^{M} \hat{f}_i(x) = h(x, \theta)$ – это некоторое параметризованное семейство функций $\hat{f}(x, \theta)$, где *M* – количество итераций для минимизации функции потерь, θ – параметр оптимизации. За итеративный алгоритм для функции потерь возьмем градиентный спуск. Для него нужно выписать градиент $\nabla L_f(\hat{f}) = \left[\frac{\partial L(y, f(x))}{\partial f(x)}\right]$. Также нужно учесть, что на каждом шаге нужно еще и выбирать оптимальный параметр $\rho \in R$.

Модель нужно обучать таким образом, чтобы предсказания были наиболее отрицательно скоррелированы с градиентом. То есть в процессе все

время будет минимизироваться квадрат разности между псевдо-остатками *r* и нашими предсказаниями.

Собственно, сам алгоритм [12]:

1) Инициализировать начальное приближение как

$$\hat{f}(x) = \hat{f}_0 = \gamma, \gamma \epsilon R: \hat{f}_0 = arg \sum_{i=1}^{n} \quad L(y_i, \gamma)$$

2) Повторяющаяся итерация *t* = 1, 2, …, *M*:

a) Вычислить псевдо-остатки r_t:

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)}, i = 1, \dots, n.$$

b) Построить новый базовый алгоритм $h_t(x)$ как регрессию на псевдо-остатках $\{(x_i, r_{it})\}(i = 1, \dots, n)$.

c) Найти оптимальный параметр ρ_t при $h_t(x)$ относительно исходной функции потерь: $\rho_t = arg \sum_{i=1}^{n} \quad L(y_i, \hat{f}(x_i) + \rho * h(x_i, \theta))$.

d) Сохранить $\hat{f}_t(x) = \rho_t * h_t(x)$.

e) Обновить текущее приближение

$$\hat{f}(x): \hat{f}(x) \leftarrow \hat{f}(x) + \hat{f}_t(x) = \sum_{i=0}^{t} \quad \hat{f}_i(x).$$

3) Собрать модель: $\hat{f}(x) = \sum_{i=0}^{M} \quad \hat{f}_i(x)$.

2.1.7. Метод адаптивного бустинга (Adaboost)

Суть адаптивного бустинга состоит в последовательном повышении точности серии базовых моделей. На каждой итерации больший вес дается тем наблюдениям, которые были неправильно классифицированы на предыдущей итерации, т.е. тем самым в приоритете обучения будут они.

Сам алгоритм [13]:

1) В начале инициализируются одинаковые веса объектов наблюдения:

$$w_i = \frac{1}{N}, i = 1,2, \dots, N;$$

2) Повторяющаяся итерация *t* = 1, …, *M*:

a) Обучить базовую модель $G_t(x)$ на тренировочной выборке умноженной на веса w_i;

b) Вычислить значение частоты ошибок в выборке:

$$err_t = \frac{\sum_{i=1}^{N} w_i \, I(y_i \neq G_t(x_i))}{\sum_{i=1}^{N} w_i};$$

c) Вычислить вес базовой модели: $\alpha_t = log((1 - err_t)/err_t)$;

d) Обновить значения весов объектов

$$w_i \leftarrow w_i \, exp\left[\alpha_t I\left(y_i \neq G_t(x_i)\right)\right], i = 1, 2, \ldots, N;$$

3) Получить метамодель $G(x) = sign[\sum_{t=1}^{M} \alpha_t \, G_t(x)]$.

2.2. Математические методы сэмплирования

В основе методов недосэмплирования лежит идея выбора из б**о**льшего класса объектов для решения проблемы дисбаланса. Рассмотрим из них четыре метода [3].

NearMiss-1 – данный метод позволяет выбирать объекты, среднее расстояние которых до *N* ближайших объектов малого класса наименьшее.

NearMiss-2 – в данном методе наоборот выбираются объекты, среднее расстояние которых до *N* дальних объектов малого класса наименьшее.

Tomek links – данный объект позволяет удалить объекты большего класса, если они образуют связи Томека. Данная связь образовывается, если между объектами из разных классов E_i и E_j не найдется ни одного объекта E_l, так чтобы выполнялись следующие неравенства:

$$d(E_i, E_l) < d(E_i, E_j) \text{ и } d(E_j, E_l) < d(E_i, E_j)$$

ENN – данный метод проводит скользящий контроль по нескольким подвыборкам, затем в них удаляет объекты большего класса, на которых метод ближайшего соседа нашел отличающиеся метки класса.

В основе методов оверсэмплирования лежит идея умного увеличения малого класса.

SMOTE (Synthetic Minority Oversampling Technique) – метод увеличения малого класса путем генерации искусственных объектов, похожих на объекты из малого класса, но не дублирующих их. Алгоритм работы данного метода следующий. Сначала случайным образом выбирается объект

x_1 из малого класса, где x_1 – это вектор, содержащий в себе числовые характеристики признаков объекта. Затем методом ближайших соседей выбираются *k* объектов из того же класса и из них выбирается один случайный объект x_2. Потом находится новый объект малого класса по следующей формуле [14]:

$\propto (x_1 - x_2) + x_2$, где $\propto$ – случайное число из промежутка (0, 1) .

ADASYN (Adaptive Sythetic) – является модифицированной версией SMOTE [14]:

1) Вычисляем степень дисбаланса: $d = \frac{m_s}{m_l}$, где m_s и m_l – количество объектов в малом и большом классе соответственно;

2) Если $d < d_{th}$, где d_{th} – порог максимально допустимой степени дисбаланса;

a) Находим количество объектов, необходимых для генерирования малого класса: $G = \beta(m_l - m_s)$;

b) Для каждого объекта x_i из малого класса найдем k ближайших соседей и посчитаем значение $r_i = \frac{\Delta_i}{k}$, где Δ_i – количество соседей которые принадлежат большему классу;

c) Нормируем r_i: $rn_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i}$;

d) Вычислим количество генерируемых примеров для каждого объекта малого класса x_i: $g_i = rn_i G$;

e) Для каждого объекта малого класса x_i вычислим g_i новых объектов. Для каждого x_i выполняем следующую итерацию g_i раз :

А) Случайно выбираем из k ближайших соседей объекта x_i объект малого класса x_{zi};

Б) Генерируем новый объект малого класса $s_i = \propto (x_{zi} - x_i) + x_i$.

2.3. Показатели эффективности моделей кредитного скоринга

После построения модели кредитного скоринга оценивается эффективность, а она рассматривается как совокупность разных показателей оценок.

Двумерная матрица ошибок (так как в выходе у нас два класса, табл. 1):

Таблица 1 – Матрица ошибок

Прогноз	Эмпирические данные	
	$y = 1$	$y = 0$
$\hat{y} = 1$	TP	FP
$\hat{y} = 0$	FN	TN

где $\hat{y}$ – метка класса модели, y – истинная метка класса, TP (True Positive) – количество надежных заемщиков, которым модель выдала кредит, FP (False Positive) – количество надежных заемщиков, которым модель не выдала кредит, FN (False Negative) – количество ненадежных заемщиков, которым модель выдала кредит, TN (True Negative) – количество ненадежных заемщиков, которым модель не выдала кредит. С помощью данной матрицы можно количественно оценить работу алгоритмов и на их результатах рассмотреть метрики качества [15].

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ – доля (точность) правильных ответов алгоритма. Данный показатель является не точным при несбалансированных данных.

$Recall = \frac{TP}{TP+FN}$ – доля положительных событий (в нашем случае, события дефолтов), которые правильно предсказаны (полнота).

$Precision = \frac{TP}{TP+FP}$ – доля ожидаемых положительных событий, которые на самом деле являются положительными (точность).

$F1\text{-}score = 2 * \frac{Precision*Recall}{Precision+Recall}$ – среднее гармоническое *Precision* и *Recall.*

Также для оценки модели в целом используется такой показатель как *AUC-ROC (Area Under Curve)* – площадь под кривой ошибок *(Receiver Operating Characteristic curve).* Кривая строится как линия от (0, 0) до (1, 1) по

координатам *TPR (True Positive Rate)* – полнота и *False Positive Rate (FPR)* – доля негативных событий, которые были неправильно предсказаны [15]:

$$TPR = \frac{TP}{TP + FN}; \qquad FPR = \frac{FP}{FP + TN}$$

Также используется такая метрика качества как *коэффициент Джини* – используется в задачах бинарной классификации при условии несбалансированности классов целевой переменной. Данный показатель связан с показателем AUC-ROC [16]:

$$Gini\ coefficient = 2AUCROC - 1.$$

2.4. Отбор значимых признаков

Перед построением моделей кредитного скоринга необходимо выделить значимые признаки, чтобы при обучении моделей не возникало излишних связей, которые будут влиять на точность прогноза. А также необходимо избавиться от зависимых друг от друга признаков. Для этого воспользуемся ANOVA F-test, многофакторным и корреляционны анализами.

ANOVA F-test часто используется для отбора предикторов в случае, когда предикторы являются количественными, а зависимая переменная категориальная. Тест проверяет гипотезу о равенстве средних значений рассматриваемых групп путем анализа дисперсий. Пусть $X -$ предиктор, $Y -$ зависимая категориальная переменная. Сформулируем основную гипотезу.

H0: Средние значения предикторов в разных группах (дефолты/не дефолты) одинаковы, т.е. переменная не значима для модели.

F-статистика и p-value:

$$F = \frac{\sum_{j=1}^{J} N_j\left(\underline{x_j} - \underline{\underline{x}}\right)^2/(J-1)}{\sum_{j=1}^{J} (N_j - 1)s_j^2/(N-J)}, \ p\_value = P\{F(J-1, N-J) > F\},$$

где $N -$число наблюдений в выборке; $N_j -$число наблюдений в категории $Y = j$, т. е. число случаев, когда $Y = 1$ (число дефолтов); $J -$ количество категорий Y; $\underline{x_j} -$ выборочное среднее предиктора X для категории $Y = 1$; $s_j^2 =$

$\frac{\sum_{i=1}^{N_j} \left(x_{i,j} - \underline{x_j}\right)^2}{N_j - 1}$ – выборочная дисперсия предиктора X для категории $Y = 1$; $\underline{\underline{x}} = \frac{\sum_{j=1}^{J} N_j \underline{\underline{x_j}}}{N}$ – выборочное среднее предиктора X.

Гипотеза H0 отвергается на уровне значимости α (1% или 5%) при $F > F(J-1, N-J, \alpha)$ или $p\_value \leq \alpha$. Отклонение H0 говорит о статистически значимой разнице между средними значениями групп, то есть признак имеет влияние на зависимую переменную.

Значимыми признаками считаются признаки с p-value меньшим заданного уровня значимости.

Многофакторный анализ – метод в математической статистике, который позволяет выявить влияние нескольких независимых переменных (факторов) на зависимую переменную [18]. Один из вариантов проведения многофакторного анализа в среде Python – это метод ранжирования с рекурсивным устранением объектов (Recursive Feature Elimination, RFE) из библиотеки sklearn. Данный метод позволяет выбрать признаки путем рекурсивного рассмотрения все меньших и меньших наборов данных. Первым элементом, необходимым для рекурсивного исключения признаков (*recursive feature elimination*), является *оценщик*, например, линейная модель. У такой модели есть коэффициенты для линейных моделей. Для выбора оптимального количества признаков нужно обучить оценщика и выбрать признаки с помощью коэффициентов. Наименее важные признаки будут удаляться. Этот процесс будет повторяться рекурсивно о тех пор, пока не будет получено оптимальное число признаков [19].

Корреляционный анализ – метод обработки статистических данных, заключающийся в изучении коэффициентов корреляции между переменными. С помощью данного анализа исключаются сильно зависимые признаки (факторы), которое определяется граничными значениями корреляции (табл. 2) [20].

Таблица 2. Граничные значения коэффициентов корреляции

Коэффициент корреляции	**Значение**
$\|r\| \leq 35\%$	Корреляция низкая
$35\% < \|r\| \leq 70\%$	Корреляция средняя
$\|r\| > 70\%$	Корреляция высокая

3. Практическая часть

3.1. Разработка алгоритма построения моделей кредитного скоринга

В данной работе на вход подается набор данных состоящий из *N* записей по *m* признаков. Реализуется два алгоритма, блок схемы приведены в приложении А: без балансировки данных (рис. 1) и с балансировкой (рис. 2).

3.2. Среда разработки для реализации алгоритма

Для программной реализации практической части выбран язык программирования *Python* на облачном сервисе Colab Google/Yandex.DataSphere. Данный язык программирования является наиболее простым, популярным и востребованным при решении задач машинного обучения и анализа данных. Это обусловлено тем, что он содержит много библиотек, упрощающих работу с данными, а также содержащих готовые алгоритмы машинного обучения. Наиболее популярными библиотеками являются *Numpy, Scipy, Pandas, Scikit-learn, Matplotlib, Seaborn, Keras* и т.д.

3.3. Подготовка и интерпретация данных

В данной работе используется набор данных (рис. 3), представленных компанией ООО «Эко-Томск». Он содержит всего 25465 наблюдений. Перед построением модели кредитного скоринга необходимо проверить и подготовить данные перед дальнейшей работой.

Сначала подключим необходимые библиотеки, импортируем файл с данными в среду программирования:

	id	vintage_year	monthly_installment	loan_balance	num_bankrupt_iva	time_since_bankrupt	time_since_ccj	ccj_amount	num_bank
0	6670001	2005	746.70	131304.44	0.0	0.0	0.0	0.0	
1	9131199	2006	887.40	115486.51	0.0	0.0	0.0	0.0	
2	4963167	2004	1008.50	128381.73	0.0	0.0	36.0	459.0	
3	3918582	2005	458.23	35482.96	0.0	0.0	0.0	0.0	
4	5949777	2006	431.20	77086.31	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	
25460	1409465	2004	1457.50	190414.54	0.0	0.0	48.0	11578.0	
25461	3951203	2005	1021.32	270842.68	0.0	0.0	0.0	0.0	
25462	3932376	2007	62.12	16416.69	0.0	0.0	0.0	0.0	
25463	6631009	2006	782.80	106171.27	0.0	0.0	0.0	0.0	
25464	1918539	2006	434.51	55693.09	0.0	0.0	0.0	0.0	

25465 rows × 38 columns

Рисунок 3 - Набор данных

Необходимо провести разведочный анализ данных: проверить данные на наличие дубликатов, пустых значений, проверить, что целевой признак с дефолтами заемщика состоит только из 0 и 1, посмотреть на типы данных и привести основные статистические показатели для каждого из признаков. А также необходимо рассмотреть соотношение дефолтных и не дефолтных наблюдений в полной выборке [1].

Данные об дефолте распределены в соотношении 24129 – не дефолтных случаев и 1336 – дефолтных случаев, то есть всего в 5,25% случаях, заемщики не возвращали кредит. Очевидно, что данные являются несбалансированными.

У каждого заемщика есть свой собственный идентификационный номер, а также 37 различных признаков (табл. 3). Признак *'default_flag'* является зависимой (целевой) переменной, которая принимает категориальные значения «0» и «1» и обозначает «хороший» и «плохой» заемщик соответственно. Остальные признаки являются независимыми переменными. Разделим зависимые переменные и целевую переменную в отдельные массивы, удалим столбец с идентификационными номерами, так как данный столбец не несет в себе никакой информации, влияющей на результат скоринга.

Перед дальнейшей работой с данными и выбором признаков для построения модели необходимо нормализовать значения независимых переменных. Это необходимо для того, чтобы значения признаков находились в одном диапазоне от 0 до 1 и тем самым не искажали восприятия взаимоотношений между независимыми признаками и зависимой переменной. Сделаем это с помощью функции *MinMaxScaler()* из пакета *sklearn.preprocessing.*

3.4. Выбор признаков для построения модели

Представленное количество признаков избыточно для построения моделей, поэтому необходимо провести процесс отбора признаков для дальнейшего использования. Для этого воспользуемся многофакторным и корреляционным и F-тест анализами.

Сначала проведем однофакторный дисперсионный анализ. Будем использовать метод *SelectKBest(score_func=f_classif)* из пакета *sklearn.feature_selection* для выбора наилучших признаков по значению *F-score*, где *f_classif* – функция для оценки *F-score* из этого же пакета. В итоге получим ранжированный список признаков (рис. 4) с их значением F-score и p-значения, отсортированный по убыванию значений F-score. А также отобразим график (рис. 5) значений F-статистики для признаков с пороговым значением F-score = 100 для выбора признаков, лежащих выше данного уровня.

	col	F-score	p-value
14	arrears_months	4719.892815	0.000000e+00
17	arrears_segment	3466.467190	0.000000e+00
16	arrears_status	2872.208073	0.000000e+00
32	woe_cc_util	2730.444319	0.000000e+00
30	woe_max_arrears_12m	2673.470091	0.000000e+00
24	worst_arrears_status	2254.000417	0.000000e+00
27	max_mia_6m	2196.935770	0.000000e+00
25	avg_mia_6m	1399.307336	3.466409e-298
36	woe_annual_income	1111.809372	1.212515e-238
29	avg_bureau_score_6m	748.188463	2.211788e-162
31	woe_bureau_score	639.752554	1.993530e-139
35	woe_months_since_recent_cc_delinq	363.069643	2.193032e-80
33	woe_num_ccj	305.487112	5.244516e-68
34	woe_emp_length	260.823851	2.214393e-58
6	time_since_ccj	124.927187	6.165909e-29
20	loan_term	105.744007	9.380414e-25
19	remaining_mat	94.952586	2.133971e-22
13	ltv	25.847097	3.721765e-07
4	num_bankrupt_iva	22.905085	1.711551e-06
15	repayment_type	22.751773	1.853541e-06

Рисунок 4 – Список признаков, отсортированных по значению F-score

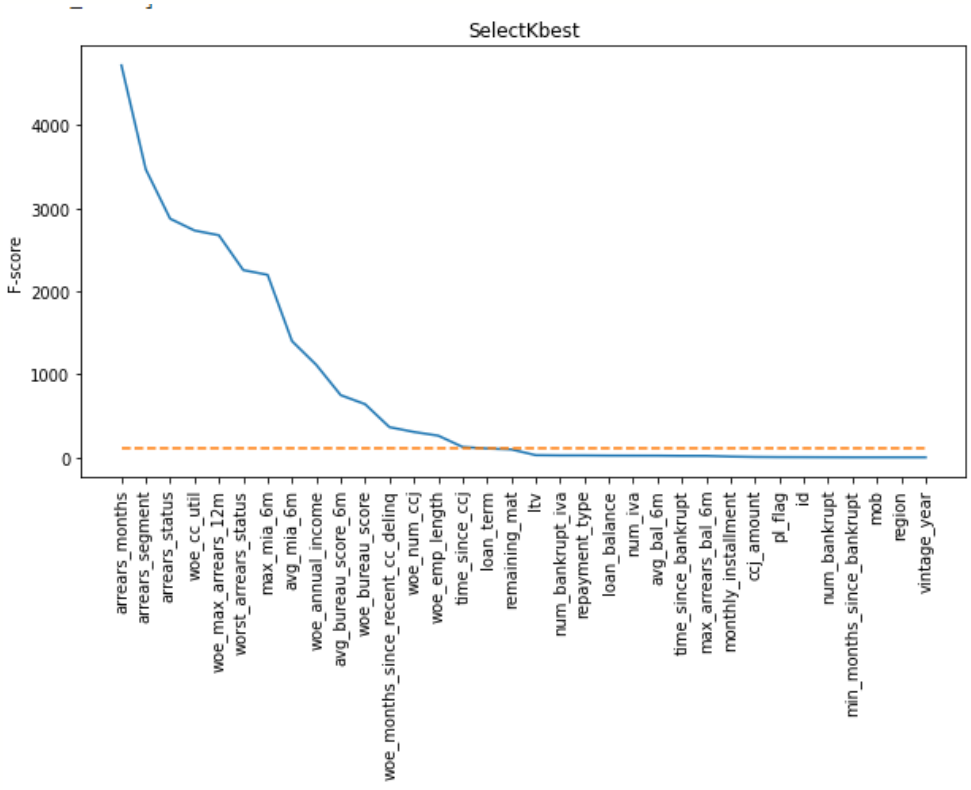

Рисунок 5 – График значений F-score для признаков

В результате мы выбрали 16 признаков под номерами 14, 17, 16, 33, 31, 24, 27, 25, 37, 29, 32, 36, 34, 35, 6, 20 (рис. 4).

Для выбранных 16 признаков проведем корреляционный анализ. Для этого построим матрицу парных корреляций, чтобы наглядно увидеть, какие признаки сильно коррелированы. Сделаем это с помощью функции *corr_data.corr()* из библиотеки *pandas*, где *corr_data* – данные по выбранным признакам из предыдущего анализа. Затем с построим корреляционную матрицу с помощью библиотеки *seaborn*:

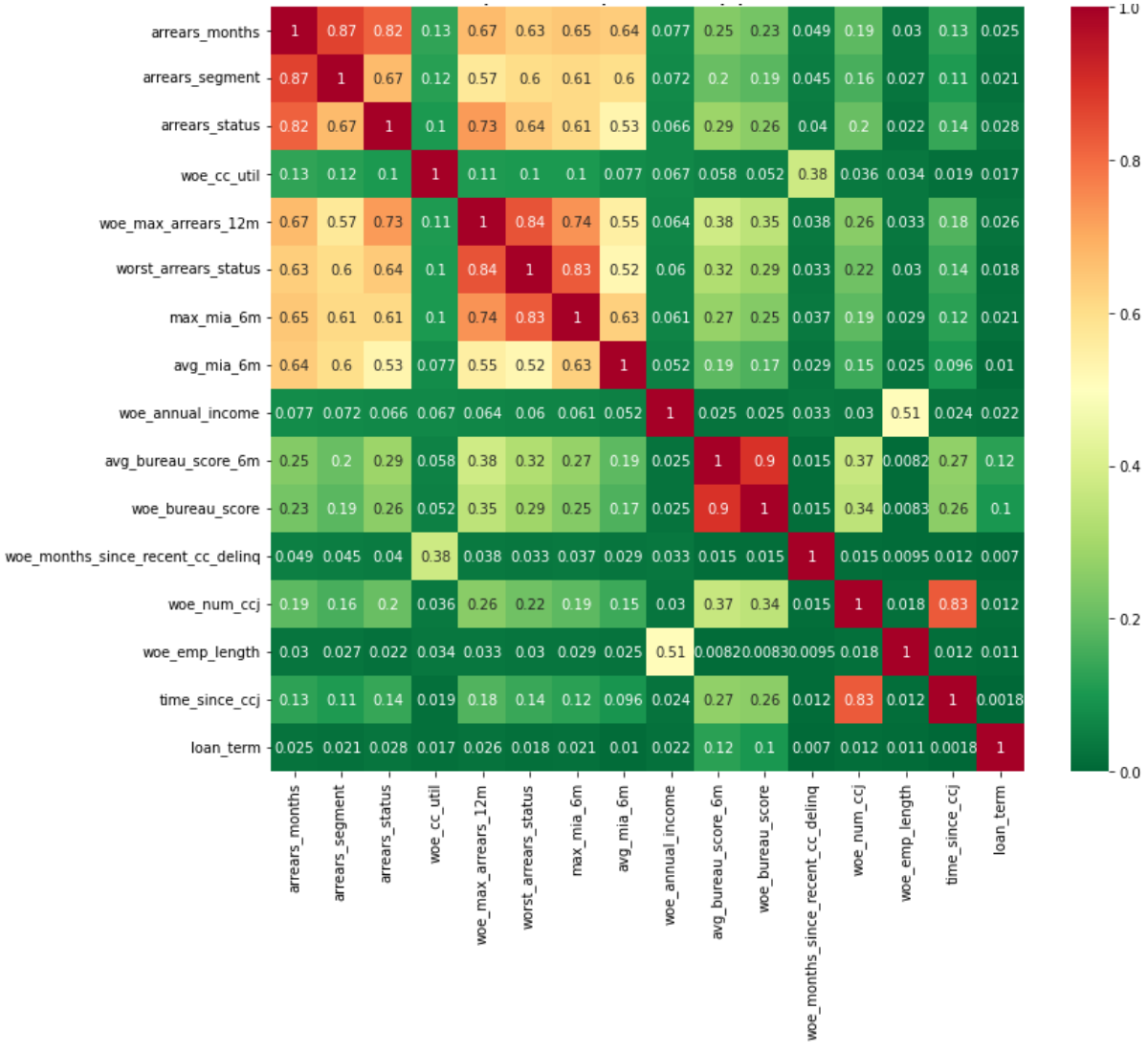

Рисунок 6 – Матрица парных корреляций

Признаки 'arrears_months' и 'woe_max_arrears_12m' – являются сильно коррелированными с другими, поэтому исключим их. Для каждой пары признаков с сильной корреляцией (рис. 4) удалим те признаки, у которых

значение F-score меньше. В итоге мы оставили 9 признаков под номерами 33, 24, 25, 37, 29, 36, 34, 35, 20.

Проведем многофакторный анализ методом ранжирования с рекурсивным устранением объектов (RFE) при помощи функции *feature_selection.RFE()* из библиотеки *sklearn*. По данному методу зададим оптимальное количество признаков равным 7. В итоге мы получили финальный список признаков:

1. 'woe_cc_util' – оценка использования кредитной карты
2. 'worst_arrears_status' – наихудший статус просроченной задолженности
3. 'avg_mia_6m' – среднее месяцев просрочки
4. 'woe_annual_income' – оценка годового дохода
5. 'woe_months_since_recent_cc_delinq' – оценка количества месяцев с момента последней просрочки по кредитной карте
6. 'avg_bureau_score_6m' – средняя оценка бюро за 6 последних месяцев
7. 'loan_term' – рок кредита.

Далее будем использовать эти 7 наиболее значимых признаков для решения задачи бинарной классификации.

3.5. Разработка моделей кредитного скоринга на несбалансированной выборке

Как и упоминалось, в данной работе будем пользоваться функциями из библиотеки *sklearn.*

Сформируем тестовые и тренировочные выборки из общей выборки случайным образом, но с соблюдением условия равномерности (одинаковая доля дефолтности 5,25%). Для этого воспользуемся функцией *train_test_split* из пакета *sklearn.model_selection.* В итоге получили следующие выборки (табл. 4):

Таблица 4. Соотношение выборок

Выборка	y = 0	y = 1
Тренировочная	16980	935
Тестовая	7239	401

Далее построим по отдельности модели кредитного скоринга на основе моделей бинарных классификаторов из библиотеки sklearn: LogisticRegression(), KNeihborsClassifier(), SVC(), DecisionTreeClassifier(), RandomForestClassifier(), GradientBoostingClassifier(), AdaBoostClassifier(DecisionTreeClassifier()), BaggingClassificator() и VotingClassificator().

Для каждой отдельной модели будем подбирать гиперпараметры (параметры для настройки работы модели, в течении обучения не меняются), чтобы она возвращала нам наиболее точный результат. Для этого мы будем пользоваться следующими инструментами из библиотеки *sklearn* для автоматического подбора гиперпараметров – *GridSearchCV (перекрестный поиск по сетке)* и *RandomizedSearchCV (случайный поиск)*. Разница данных методов в том, что *GridSearchCV* производит поиск по всем возможным комбинациям гиперпараметров и проводит несколько экспериментов с перекрестной проверкой, а *RandomizedSearchCV* производит поиск по *n_iter* случайным комбинациям из всех комбинаций гиперпараметров [21]. Второй метод работает быстрее, но первый метод точнее за счет полной проверки. В нашем случае будем использовать второй метод только для случайного леса, так как там производится подбор большего количества параметров.

После данной процедуры с помощью метода fit(X_train, y_train) мы обучаем наши оптимизированные модели, затем с помощью методов *predict(X_test), predict.proba(X_test)* находим прогноз отношения к классу 0 или 1 и прогнозные вероятности отнесения объекта к классу 0 или 1 соответственно. После этого необходимо оценить полученные результаты и сравнить их.

3.6. Разработка моделей кредитного скоринга на сбалансированной выборке

Применим перечисленные в предыдущем разделе выше процедуры для сбалансированных данных. Отдельно рассмотрим методы пересэмплирования и недосэмплирования, затем выберем налучший и построим комбинированный метод. Балансировать будем тренировочные данные и по ним для каждой модели будем подбирать оптимальные гиперпараметры. Все методы указанные методы реализованы в библиотеке *imblearn.*

Из методов пересэмплирования рассмотрим *SMOTE()* и *ADASYN().* Будем для каждого отдельного метода строить модели классификации с теми же гиперпараметрами, что и при работе с несбалансированными данными.

Из методов недсэмплирования рассмотрим *NearMiss(), TomekLinks(), EditedNearestNeighbours().*

Оценки построенных моделей, сбалансированных по разным методам, будут проводиться отдельно для каждой модели. Из ансамблевых методов рассматриваются те, которые показали лучшие результаты по несбалансированным данным. Полученные результаты сравниваются и выбирается лучшая модель классификатора и методы сэмплирования. Затем оценивается каждая модель по комбинированному методу, сравниваются и выбирается лучшая модель.

3.7. Оценка качества построенных моделей

На тестовых выборках были проведены оценки качества построенных моделей. Для этого будем использовать такие инструменты, как *матрица ошибок, ROC кривая* и *метрики качества (Precision, Recall, F1-score, Accuracy, Gini, AUROC).* Воспользуемся готовыми функциями *classification_report(), roc_auc_score(), roc_curve(), confusion_matrix(), plot_confusion_matrix()* из пакета *sklearn.metrics* для выполнения оценок.

3.7.1. Несбалансированные данные

Для начала рассмотрим базовые модели и самостоятельные ансамблевые модели (случайный лес, градиентный бустинг и адаптивный бустинг). Сравнение матриц ошибок при прогнозировании (рис. 7):

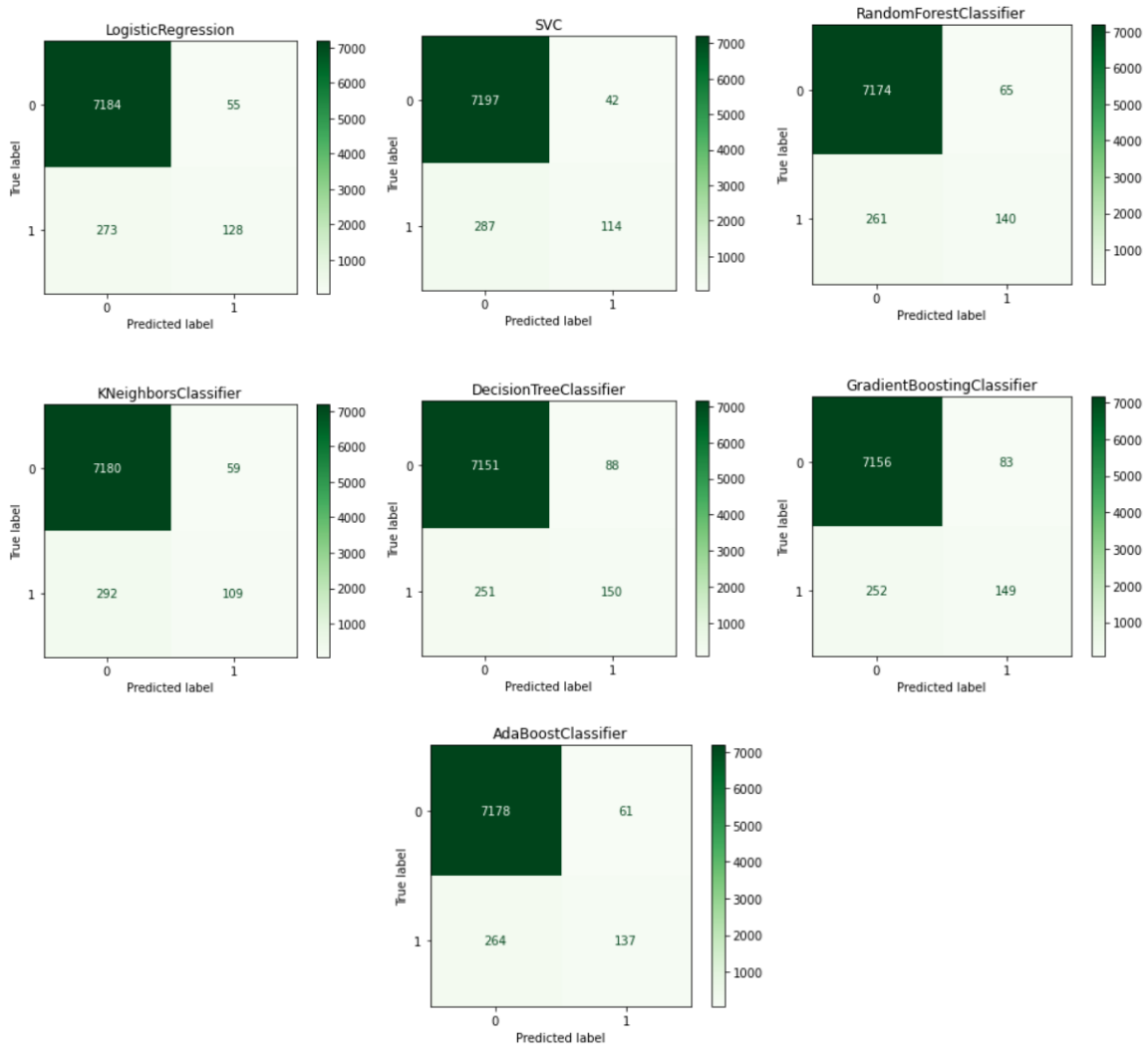

Рисунок 7 – Матрицы ошибок для построенных моделей

По результатам построения матриц ошибок приходим к выводу о том, что построенные модели уверенно предсказывают «хороших» заемщиков, а «плохих» заемщиков предсказывают намного хуже. Число ложно «хороших» заемщиков почти в два раза больше чем истинных «плохих» заемщиков. То есть модели одобрили кредиты большей части тех, кто не вернет кредит. Делаем предположение о том, что на полученные результаты повлияла несбалансированность данных.

Рассмотрим результаты метрик качества (табл. 5). Желтым отмечены лучшие модели по отдельным показателям метрик качеств (лучшие показатели по столбцам).

Таблица 5 – Метрики качества для моделей без балансировки данных

Модель	*Precision 0*	*Recall 0*	*F-мера 0*	*Precision 1*	*Recall 1*	*F-мера 1*	*Accuracy*	*AUCROC*	*Gini*
Логистическая регрессия	0,963	0,992	0,978	0,699	0,319	0,438	0,957	0,911	0,822
Метод k-ближайших соседей	0,961	0,992	0,976	0,649	0,272	0,383	0,954	0,883	0,766
Метод опорных векторов	0,962	0,994	0,978	0,731	0,284	0,409	0,957	0,767	0,535
Дерево решений	0,966	0,988	0,977	0,630	0,374	0,469	0,956	0,905	0,811
Случайный лес	0,965	0,991	0,978	0,683	0,349	0,462	0,957	0,925	0,849
Метод градиентного бустинга	0,966	0,989	0,977	0,642	0,372	0,471	0,956	0,926	0,851
Метод адаптивного бустинга	0,965	0,992	0,978	0,692	0,342	0,457	0,957	0,908	0,815

Как можно заметить из результатов, все модели очень хорошо предсказывают «хороших» заемщиков. Лучшей моделью среди таких моделей по показателю F1-мера (среднее гармоническое между полнотой и точностью) являются методы логистической регрессии, опорных векторов, случайного леса и адаптивного бустинга. А лучшей моделью среди моделей, неплохо предсказывающих «плохих» заемщиков является метод градиентного бустинга, но также неплохой результат показали метод случайного леса и дерево решений. Метод градиентного бустинга также показал высокие коэффициенты Джини и AUCROC, поэтому делаем вывод о том, что наиболее точным из представленных методов является данный метод.

Рассмотрим ROC-кривую и визуально определим лучшую модель (рис. 8).

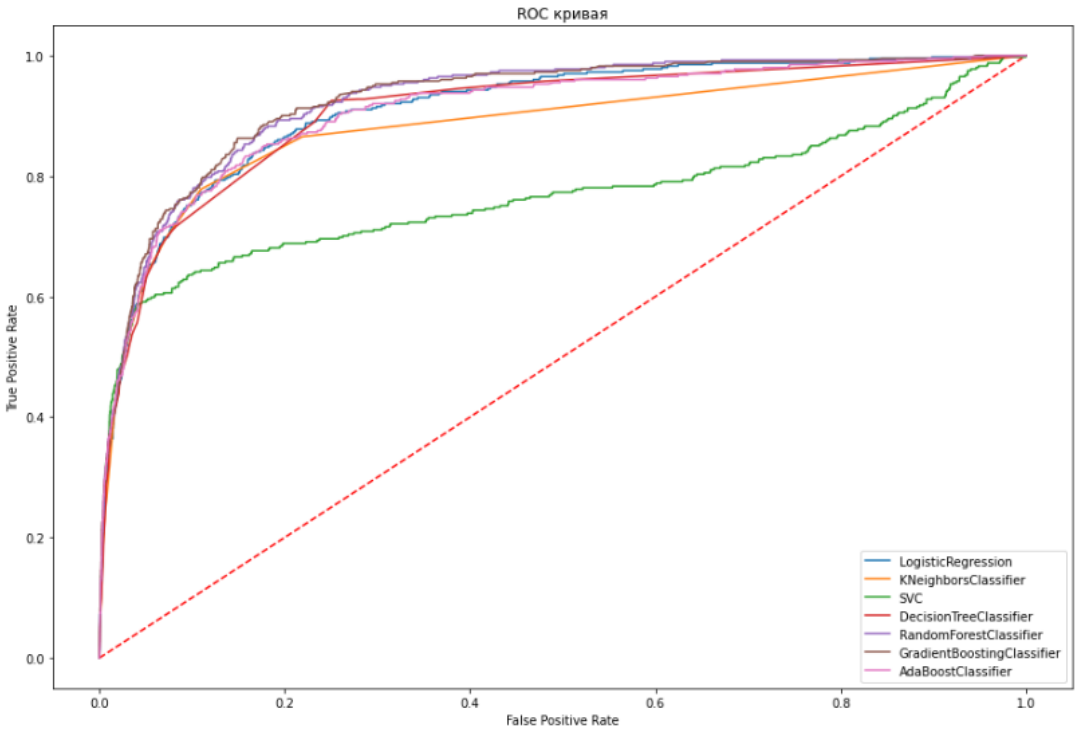

Рисунок 8 - ROC-кривая моделей по несбалансированным данным

По рисунку 8 видно, что метод градиентного бустинга и случайного леса имеют более высокую точность, так как они находятся ближе всего к левому верхнему углу (0,1).

Теперь рассмотрим ансамблевые модели, построенные по методу бэггинга на основе логистической регрессии, метода *k*-ближайших соседей и метода опорных векторов. Рассмотрим матрицы ошибок (рис. 9):

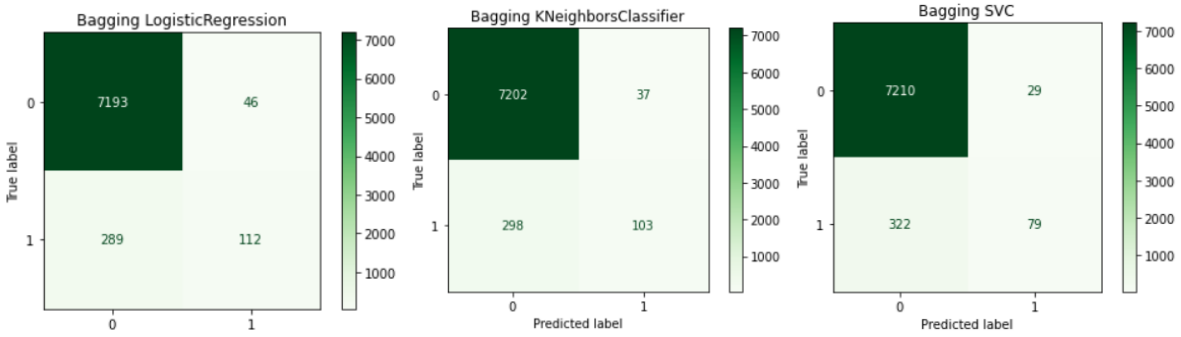

Рисунок 9 - Матрица ошибок моделей бэггинга по несбалансированным данным

Как видно из результатов, модели научились лучше предсказывать «хороших» заемщиков, но стали хуже предсказывать «плохих» заемщиков. Рассмотрим результаты метрик качества (табл. 6) и ROC-кривую (рис. 10):

Таблица 6 - Метрики качества моделей бэггинга по несбалансированным данным

Модель/ Бэггинг на основе:	*Precision 0*	*Recall 0*	*F-мера 0*	*Precision 1*	*Recall 1*	*F-мера 1*	*Accuracy*	*AUCROC*	*Gini*
Логистическая регрессия.	0,961	0,994	0,977	0,709	0,279	0,401	0,956	0,914	0,828
Метод k-ближайших соседей.	0,960	0,995	0,977	0,736	0,257	0,381	0,956	0,919	0,839
Метод опорных векторов.	0,957	0,996	0,976	0,731	0,197	0,310	0,954	0,861	0,722

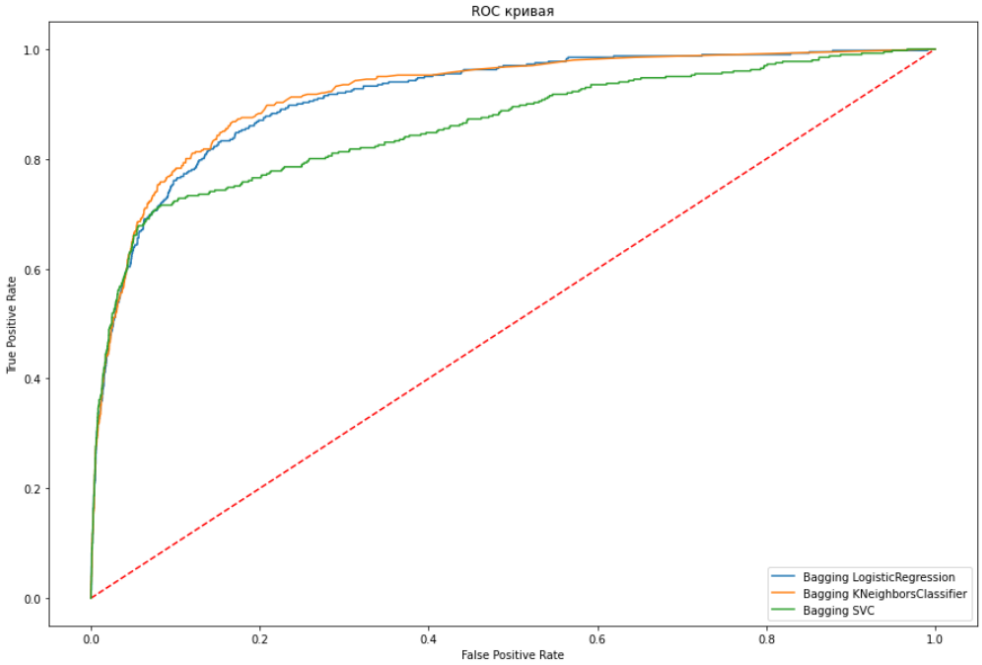

Рисунок 10 - ROC-кривая моделей бэггинга по несбалансированным данным

По результатам видно, что при бэггинге моделей точность предсказывания «плохих» заемщиков уменьшилась, но в целом точность предсказывания увеличился (по показателям *AUCROC* и *Gini*). Из данных моделей лучшим оказался модель бэггинга на основе логистической регрессии по показателям полноты и средней гармонической для «плохих» заемщиков, хоть и способность предсказывать «плохих» заемщиках значительно снизилась чем у одиночной модели.

Построим ансамблевые модели по методу мягкого голосования. Будем объединять в пары самостоятельные модели с методом градиентного

бустинга, так как он показал лучшие результаты. А также объединим три лучшие самостоятельные модели: градиентный бустинг, случайный лес и логистическая регрессия и три разные модели: градиентный бустинг, метод опорных векторов и дерево решений.

Рассмотрим матрицу ошибок (рис. 11, прил. В), показатели метрик качества (табл. 7) и ROC-кривую (рис. 12, прил. В).

Таблица 7 - Метрики качества моделей голосования по несбалансированным данным

Модель / Градиентный бустинг и:	*Precision 0*	*Recall 0*	*F-мера 0*	*Precision 1*	*Recall 1*	*F-мера 1*	*Accuracy*	*AUC ROC*	*Gini*
Логистическая регрессия.	0,964	0,991	0,978	0,678	0,337	0,450	0,957	0,922	0,844
Метод k-ближайших соседей.	0,965	0,991	0,978	0,686	0,354	0,467	0,958	0,92	0,84
Метод опорных векторов.	0,962	0,994	0,978	0,739	0,297	0,423	0,958	0,926	0,852
Дерево решений	0,966	0,989	0,977	0,641	0,369	0,468	0,956	0,922	0,844
Случайный лес.	0,966	0,990	0,978	0,665	0,362	0,468	0,957	0,926	0,852
Метод адаптивного бустинга	0,966	0,990	0,978	0,668	0,372	0,478	0,957	0,923	0,845
Логистическая регрессия и случайный лес.	0,965	0,992	0,978	0,699	0,342	0,459	0,958	0,925	0,849
Метод опорных векторов и дерево решений	0,964	0,993	0,978	0,722	0,337	0,459	0,958	0,924	0,848

Объединение градиентного бустинга и других моделей не увеличило точность предсказывания «плохих заемщиков», но увеличило точность предсказывания «хороших» заемщиков. Все показатели «хороших» заемщиков и общие показатели приблизительно находятся на одном уровне. Но по показателям «плохих» заемщиков Recall 1 и F-score 1 лучшим оказалось объединение методов градиентного и адаптивного бустинга.

3.7.2. Сбалансированные данные

Логистическая регрессия. Рассмотрим матрицы ошибок для модели при разных способах балансировки данных.

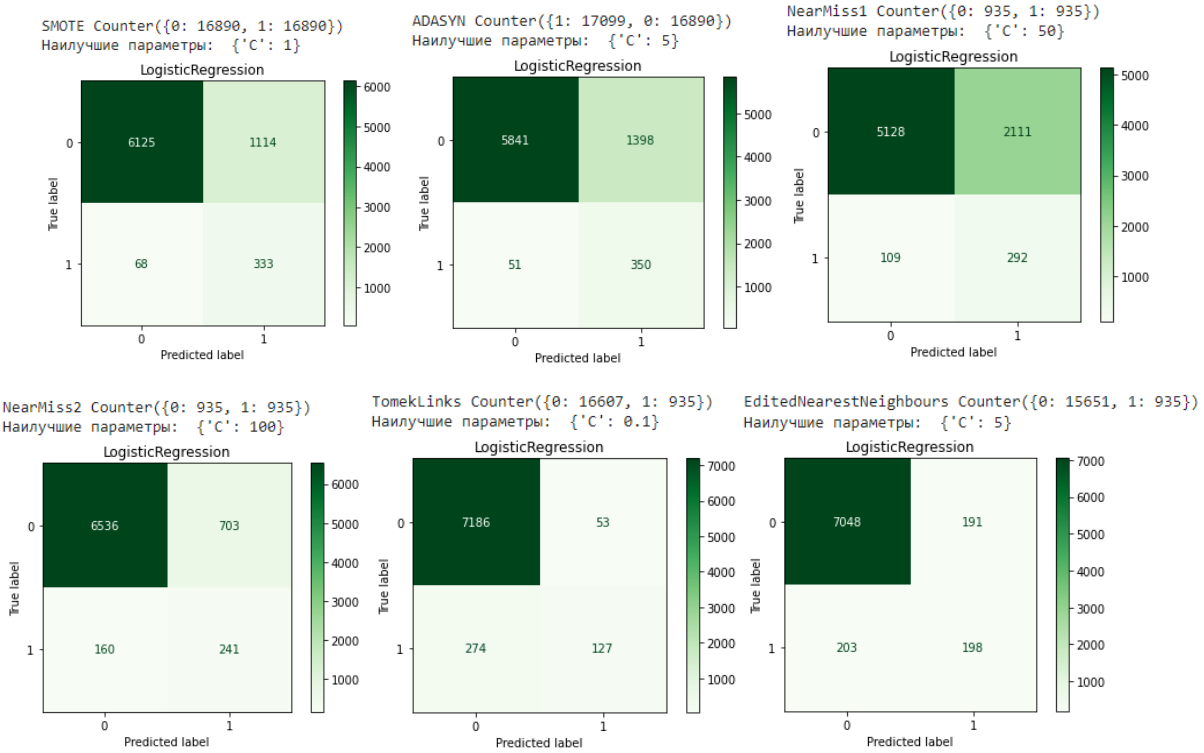

Рисунок 13 – Матрица ошибок лог. регрессии при разных методах сэмплирования

В результате видим, что количество заемщиков, правильно отнесенных к «плохим» заемщикам увеличилось. Но также заметим, что количество отказов «хорошим» заемщикам тоже увеличилось. Из методов пересэмплирования на первый взгляд лучшим себя показал метод Smote, а из методов недосэмплирования – метод ENN. Рассмотрим метрики качества:

Таблица 8 – Логистическая регрессия при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,845	0,810	0,709	0,887	0,957	0,948
Precision 0	0,989	0,991	0,979	0,976	0,963	0,972
Recall 0	0,846	0,807	0,708	0,903	0,993	0,974
F1-score 0	0,912	0,890	0,822	0,938	0,957	0,973
Precision 1	0,230	0,200	0,122	0,255	0,706	0,509
Recall 1	0,830	0,873	0,728	0,601	0,317	0,494
F1-score 1	0,360	0,326	0,208	0,358	0,437	0,501
AUROC	0,913	0,912	0,799	0,847	0,91	0,912
Gini	0,825	0,825	0,597	0,694	0,82	0,825

Если судить по F1-мере для невозвратных случаев, то из методов пересэмплирования большее значение имеет SMOTE, но ADASYN так же показал неплохие результаты, а из методов недосэмплирования – ENN и TomekLinks. Теперь рассмотрим ROC-кривые (рис. 14): на графике повторяются вышеприведенные выводы по метрикам качества.

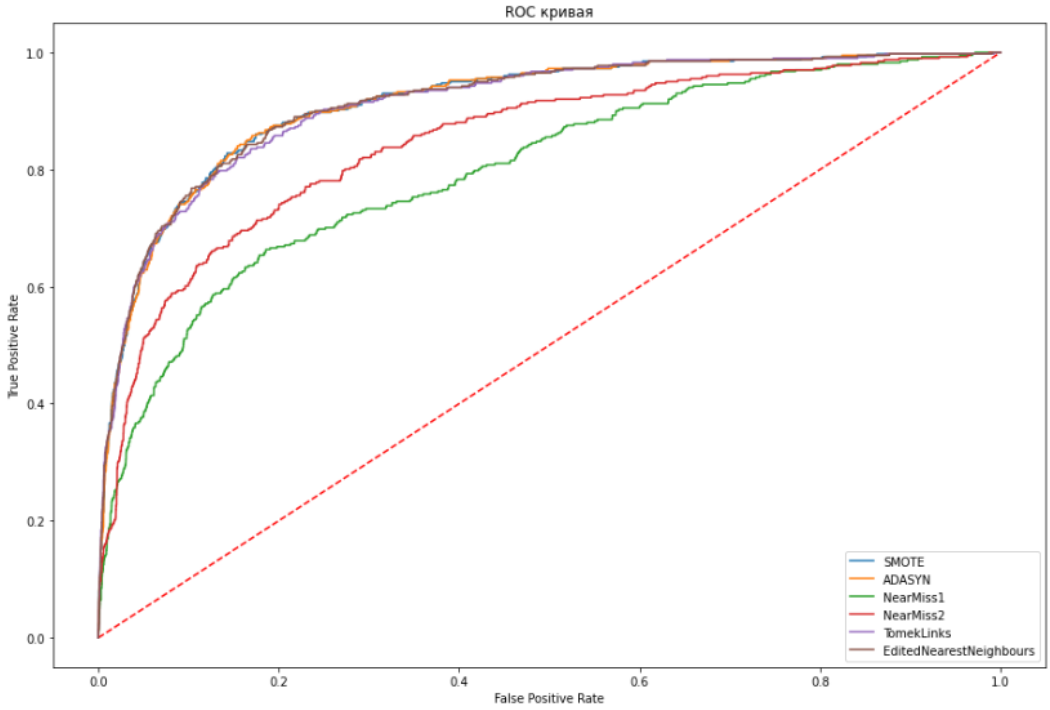

Рисунок 14 - ROC-кривые лог. регрессии по разным методам сэмплирования

Метод k-ближайших соседей. По виду матриц ошибок (рис. 15) и по графикам ROC-кривых (рис. 16) оптимальными методами являются SMOTE и ENN. Приведем Таблицу с метриками качества:

Таблица 9 – Метод k-ближайших соседей при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,889	0,876	0,468	0,237	0,953	0,949
Precision 0	0,984	0,985	0,975	0,933	0,965	0,971
Recall 0	0,898	0,883	0,451	0,209	0,987	0,975
F1-score 0	0,939	0,931	0,616	0,342	0,976	0,973
Precision 1	0,285	0,262	0,074	0,049	0,597	0,515
Recall 1	0,733	0,753	0,788	0,731	0,344	0,481
F1-score 1	0,410	0,389	0,135	0,091	0,437	0,497
AUROC	0,873	0,868	0,585	0,514	0,874	0,875
Gini	0,745	0,737	0,170	0,028	0,748	0,749

При сравнении методов пересэмплирования и недосэмплирования выяснили, что методы SMOTE и TomekLinks лучшие по нескольким показателям, но по важной для нас характеристике (F1-мера) ENN – лучше.

Метод опорных векторов. По виду матриц ошибок (рис. 17) и по графикам ROC-кривых (рис. 18) оптимальными методами являются SMOTE и ENN. Рассмотрим таблицу с метриками качества:

Таблица 10 – Метод опорных векторов при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,862	0,813	0,300	0,119	0,957	0,951
Precision 0	0,989	0,991	0,970	0,879	0,965	0,972
Recall 0	0,864	0,810	0,269	0,081	0,990	0,977
F1-score 0	0,922	0,892	0,421	0,148	0,978	0,974
Precision 1	0,253	0,203	0,061	0,046	0,673	0,536
Recall 1	0,830	0,873	0,850	0,798	0,354	0,484
F1-score 1	0,388	0,329	0,113	0,087	0,464	0,509
AUROC	0,902	0,897	0,549	0,580	0,773	0,83
Gini	0,805	0,795	0,098	0,160	0,546	0,661

При сравнении методов пересэмплирования и недосэмплирования выяснили, что метод SMOTE лучше, а ENN и TomekLinks очень похожи, но по важной для нас характеристике (F1-мера) ENN – лучше.

Дерево решений. По виду матриц ошибок (рис. 19) и по графикам ROC-кривых (рис. 20) оптимальными методами являются SMOTE и ENN. Рассмотрим таблицу с метриками качества:

Таблица 11 – Дерево решений при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,889	0,868	0,544	0,883	0,956	0,945
Precision 0	0,982	0,984	0,983	0,975	0,967	0,974
Recall 0	0,899	0,876	0,527	0,900	0,987	0,968
F1-score 0	0,939	0,927	0,686	0,936	0,977	0,971
Precision 1	0,278	0,247	0,089	0,245	0,624	0,481
Recall 1	0,701	0,738	0,838	0,589	0,401	0,531
F1-score 1	0,398	0,371	0,162	0,346	0,489	0,505
AUROC	0,874	0,873	0,802	0,742	0,908	0,910
Gini	0,747	0,745	0,605	0,485	0,816	0,82

При сравнении методов пересэмплирования и недосэмплирования выяснили, что метод SMOTE лучше, а из ENN и TomekLinks у второго показатели выше, но по важной для нас характеристике (F1-мера) ENN – лучше.

Случайный лес. По виду матриц ошибок (рис. 21) оптимальными методами являются SMOTE и ENN, а по графикам ROC-кривых (рис. 22) не все так очевидно. Рассмотрим таблицу с метриками качества:

Таблица 12 – Случайный лес при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,894	0,873	0,304	0,351	0,956	0,947
Precision 0	0,986	0,987	0,980	0,955	0,966	0,973
Recall 0	0,901	0,877	0,271	0,331	0,988	0,971
F1-score 0	0,942	0,929	0,424	0,492	0,977	0,972
Precision 1	0,299	0,262	0,064	0,056	0,635	0,493
Recall 1	0,761	0,786	0,898	0,721	0,382	0,506
F1-score 1	0,430	0,393	0,119	0,104	0,477	0,499
AUROC	0,917	0,914	0,763	0,68	0,924	0,924
Gini	0,834	0,829	0,526	0,361	0,847	0,849

При сравнении методов пересэмплирования и недосэмплирования выяснили, что метод SMOTE лучше, а из ENN и TomekLinks у второго показатели выше, но по важной для нас характеристике (F1-мера) ENN – лучше.

Метод градиентного бустинга. По виду матриц ошибок (рис. 23) оптимальными методами являются SMOTE и ENN, и по графикам ROC-кривых (рис. 24) та же картина. Рассмотрим таблицу с метриками качества:

Таблица 13 – Метод градиентного бустинга при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,948	0,946	0,401	0,140	0,955	0,948
Precision 0	0,971	0,970	0,982	0,914	0,968	0,974
Recall 0	0,974	0,974	0,374	0,102	0,985	0,970
F1-score 0	0,972	0,972	0,542	0,184	0,976	0,972
Precision 1	0,501	0,487	0,072	0,048	0,601	0,501
Recall 1	0,476	0,451	0,875	0,825	0,416	0,539
F1-score 1	0,488	0,468	0,133	0,092	0,492	0,519
AUROC	0,899	0,891	0,709	0,478	0,926	0,925

Gini	0,799	0,783	0,418	-0,044	0,851	0,851

При сравнении методов пересэмплирования и недосэмплирования выяснили, что метод SMOTE лучше, а из ENN и TomekLinks у второго показатели выше, но по важной для нас характеристике (F1-мера) ENN лучше.

Метод адаптивного бустинга. По виду матриц ошибок (рис. 25) оптимальными методами являются SMOTE и ENN, а по графикам ROC-кривых (рис. 26) SMOTE и TomekLinks. Рассмотрим таблицу с метриками качества:

Таблица 14 – Метод адаптивного бустинга при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,952	0,951	0,678	0,166	0,958	0,944
Precision 0	0,966	0,966	0,986	0,921	0,966	0,975
Recall 0	0,983	0,983	0,669	0,131	0,991	0,965
F1-score 0	0,975	0,975	0,797	0,229	0,978	0,970
Precision 1	0,558	0,553	0,122	0,048	0,695	0,472
Recall 1	0,384	0,377	0,830	0,798	0,364	0,561
F1-score 1	0,455	0,448	0,213	0,091	0,478	0,513
AUROC	0,885	0,889	0,818	0,552	0,924	0,887
Gini	0,771	0,778	0,636	0,105	0,848	0,774

При сравнении методов пересэмплирования и недосэмплирования выяснили, что метод SMOTE лучше, а из ENN и TomekLinks у второго показатели намного выше, поэтому оставим его.

Бэггинг логистической регрессии. По виду матриц ошибок (рис. 27) оптимальными методами являются SMOTE и ENN, а по графикам ROC-кривых (рис. 28) SMOTE и TomekLinks. Рассмотрим таблицу с метриками качества:

Таблица 15 – Бэггинг логистической регрессии при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,844	0,807	0,717	0,910	0,957	0,951
Precision 0	0,989	0,991	0,979	0,977	0,963	0,968
Recall 0	0,844	0,803	0,717	0,927	0,993	0,981
F1-score 0	0,911	0,887	0,828	0,951	0,978	0,975
Precision 1	0,229	0,197	0,125	0,312	0,703	0,549
Recall 1	0,833	0,873	0,728	0,599	0,307	0,421
F1-score 1	0,359	0,322	0,213	0,410	0,427	0,477

AUROC	0,913	0,913	0,804	0,861	0,914	0,914
Gini	0,826	0,825	0,608	0,722	0,827	0,829

Из таблицы можно заметить, что по сравнению с одиночной логистической регрессией значительных улучшений показателей не появилось. Также методы SMOTE и ENN по важным показателям показали себя лучше.

Объединение голосованием методов градиентного и адаптивного бустинга. По виду матриц ошибок (рис. 29) оптимальными методами являются SMOTE и ENN, а по графикам ROC-кривых (рис. 30) ADASYN и ENN. Рассмотрим таблицу с метриками качества:

Таблица 16 – Объединение голосованием методов градиентного и адаптивного бустинга при сбалансированных данных

	Методы сэмплирования					
Метрики качества	*SMOTE*	*ADASYN*	*NearMiss1*	*NearMiss2*	*TomekLinks*	*ENN*
Accuracy	0,935	0,935	0,334	0,234	0,951	0,938
Precision 0	0,975	0,975	0,975	0,952	0,968	0,973
Recall 0	0,956	0,955	0,305	0,201	0,981	0,961
F1-score 0	0,965	0,965	0,464	0,332	0,974	0,967
Precision 1	0,411	0,410	0,064	0,054	0,542	0,425
Recall 1	0,551	0,559	0,858	0,815	0,406	0,524
F1-score 1	0,471	0,473	0,119	0,100	0,464	0,469
AUROC	0,911	0,911	0,722	0,582	0,914	0,921
Gini	0,822	0,822	0,444	0,164	0,827	0,843

Из методов недосэмплирования лучшим оказался метод ENN, а из методов пересэмплирования ADASYN оказался лучше. По сравнению с одиночной моделью адаптивного бустинга показатели выше, а по сравнению с градиентным бустингом показатели снизились, но по методам пересэмплирования результаты получились выше.

В итоге приходим к выводу, что из методов сэмплирования самые лучшие методы – это SMOTE и ENN. А из моделей кредитного скоринга хорошо себя показали случайный лес, метод градиентного бустинга, объединение голосованием. Наиболее оптимальным показал себя метод градиентного бустинга, так как в данном методе при балансировке данных

сохранялся высокий уровень F1-меры. Заметим, что высокие метрики качества показали ансамблевые методы.

Комбинированный метод. Теперь построим модели кредитного скоринга с комбинированным методом балансировки. В качестве комбинированного метода возьмем комбинацию лучших методов сэмплирования по предыдущим результатам – SMOTE и ENN. Существует такая готовая функция *SMOTEENN()* из пакета *imblearn.combine.*

Сначала рассмотрим матрицы ошибок:

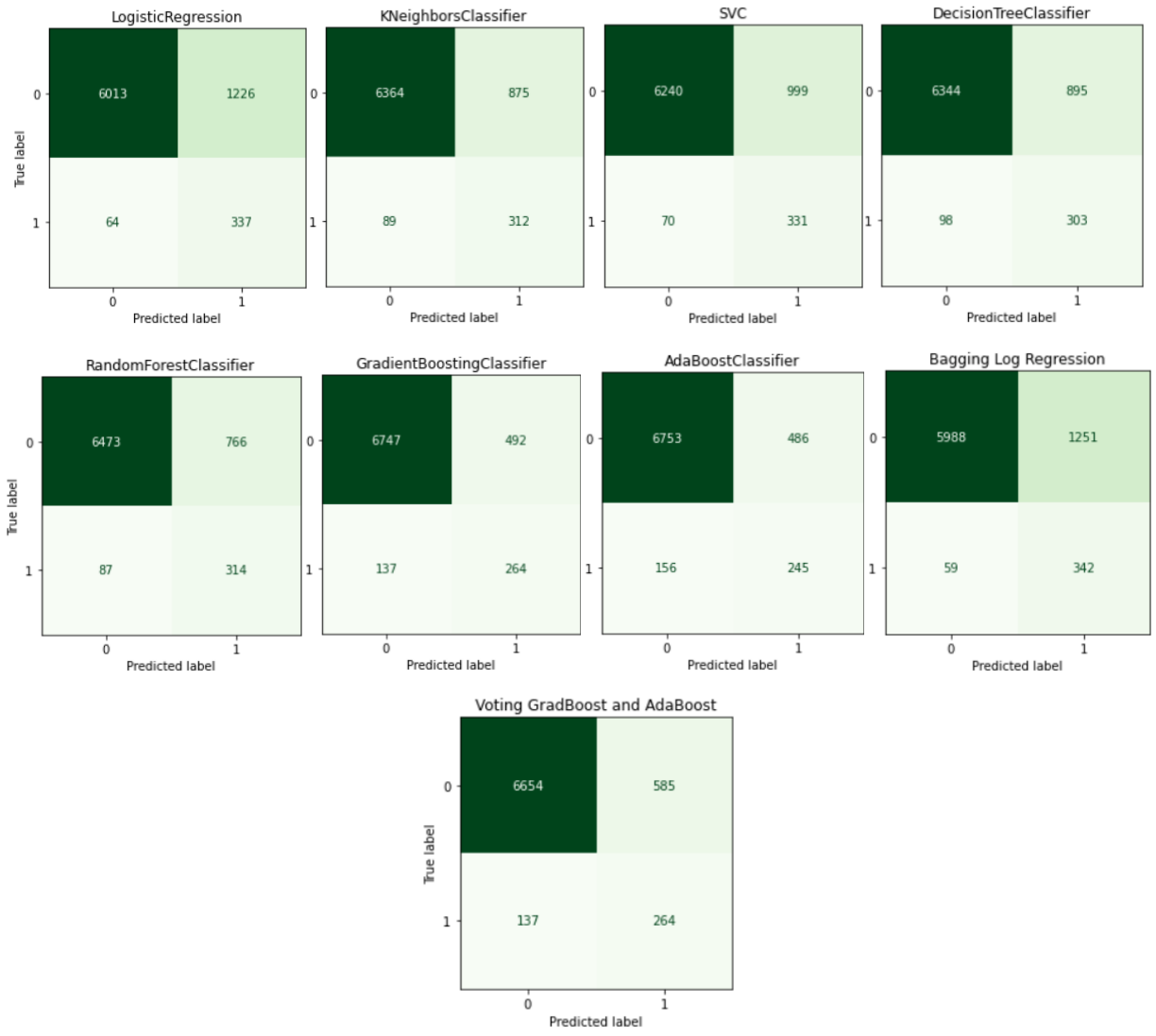

Рисунок 31 – Матрица ошибок для моделей, сбалансированных по SMOTEENN

По виду матриц ошибок делаем вывод о том, что наиболее точно «плохих» заемщиков предсказывает бэггинг логистической регрессии, но при этом она теряет много «хороших» заемщиком. Оптимальной моделью является метод градиентного бустинга, так как в этом случае у нас меньше

всего потерь «хороших» заемщиков и выявляется большая часть «плохих» заемщиков.

Теперь рассмотрим метрики качества:

Таблица 17 - Метрики качеств для моделей, сбалансированных по SMOTEENN

Модель	*Precision 0*	*Recall 0*	*F1-мера 0*	*Precision 1*	*Recall 1*	*F-мера 1*	*Accuracy*	*AUCROC*	*Gini*
Логистическая регрессия.	0,989	0,831	0,903	0,216	0,840	0,343	0,831	0,913	0,826
Метод k-ближайших соседей.	0,986	0,879	0,930	0,263	0,778	0,393	0,874	0,869	0,738
Метод опорных векторов.	0,989	0,862	0,921	0,249	0,825	0,382	0,860	0,904	0,809
Дерево решений	0,985	0,876	0,927	0,253	0,756	0,379	0,870	0,875	0,750
Случайный лес.	0,987	0,894	0,938	0,291	0,783	0,424	0,888	0,921	0,842
Метод градиентного бустинга.	0,980	0,932	0,955	0,349	0,658	0,456	0,918	0,896	0,792
Метод адаптивного бустинга	0,977	0,933	0,955	0,335	0,611	0,433	0,916	0,889	0,778
Бэггинг логистической регрессии	0,990	0,827	0,901	0,215	0,853	0,343	0,829	0,913	0,827
Объединение голосованием	0,980	0,919	0,949	0,311	0,658	0,422	0,905	0,919	0,838

Как можно заметить по полученным оценкам, по общей точности Accuracy, по среднему гармоническому предсказывания «хороших» и «плохих» заемщиков лучшей моделью оказался метод градиентного бустинга. Если говорить про наиболее точное предсказывание «плохих» заемщиков по показателю Recall 1 лучшим оказался бэггинг логистической регрессии, но у него низкий показатель (Recall 0) предсказывания «хороших» заемщиков. По показателям AUCROC и Gini лучшим оказался метод случайного леса.

Рассмотрим ROC-кривые:

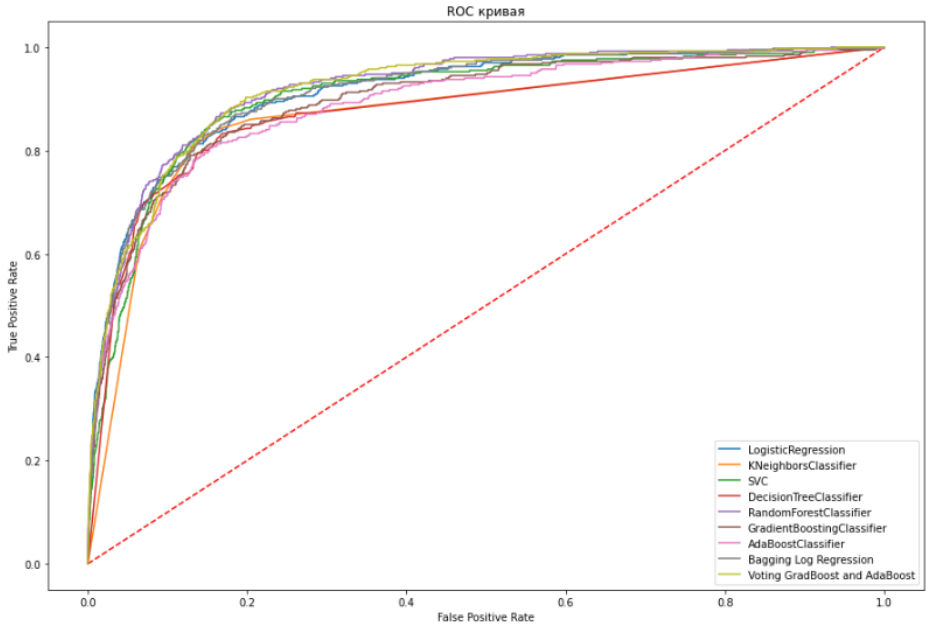

Рисунок 32 - ROC-кривые для моделей, сбалансированных по методу SMOTEENN

По построенным кривым можно сделать вывод о том, что модели, обученные на сбалансированных по комбинированному методу данных, показали наиболее стабильные результаты в отличие от других моделей. Все кривые находятся примерно в одинаковом расположении. Но ближе к левому верхнему углу располагается кривая случайного леса, как и показывали метрики AUCROC и Gini.

Сравним все показатели для разных случаев с методом градиентного бустинга:

Таблица 18 – Сравнение показателей метода градиентного бустинга при разных случаях

Модель	*Precision 0*	*Recall 0*	*F1-мера 0*	*Precision 1*	*Recall 1*	*F-мера 1*	*Accuracy*	*AUCROC*	*Gini*
Метод градиентного бустинга	0,966	0,989	0,977	0,642	0,372	0,471	0,956	0,926	0,851
Метод градиентного бустинга с SMOTE	0,971	0,974	0,972	0,501	0,476	0,488	0,948	0,899	0,799
Метод градиентного бустинга с ENN	0,974	0,970	0,972	0,501	0,539	0,519	0,948	0,925	0,851

Объединение голосованием с ADASYN	0,975	0,955	0,965	0,410	0,559	0,473	0,935	0,911	0,822
Объединение голосованием с ENN	0,973	0,961	0,967	0,425	0,524	0,469	0,938	0,921	0,843
Метод градиентного бустинга с SMOTEENN	0,980	0,932	0,955	0,349	0,658	0,456	0,918	0,896	0,792
Объединение голосованием с SMOTEENN	0,980	0,919	0,949	0,311	0,658	0,422	0,905	0,919	0,838

Как видим из результатов, значения метрик Accuracy, AUCROC, Gini, Recall 0, F1-мера 0, Precision 1 уменьшились, но нашей задаче данные показатели не являются решающими, так как они направлены на предсказывание «хороших» заемщиков. А вот показатели Precision 0, Recall 1, F1-мера 1 увеличились, что более важно. Самое большое значение F1-меры у метода градиентного бустинга с ENN, также у него показатели Accuracy, AUCROC, Gini выше по сравнению с методом градиентного бустинга с SMOTEENN, но у него низкий показатель предсказывания «плохих» заемщиков по сравнению с тем же методом. Нам нужна модель, которая одинаково хорошо предсказывает и «хороших» и «плохих» заемщиков. Данную цель достигает модель кредитного скоринга, построенная по методу градиентного бустинга с методом ENN.

Сравним полученные результаты с результатами диссертации [22]. В данной работе был сделан вывод, что ансамблевая модель адаптивного бустинга основанный на классификаторе деревьев решений показывает наилучший результат. Сравним результаты:

Таблица 19 – Сравнение результатов

Модель	*Precision 0*	*Recall 0*	*F1-мера 0*	*Precision 1*	*Recall 1*	*F-мера 1*	*Accuracy*	*AUCROC*	*Gini*
Метод градиентного бустинга с ENN	0,974	0,970	0,972	0,501	0,539	0,519	0,948	0,925	0,851

Модель другого автора	0,968	0,991	0,979	0,705	0,406	0,515	0,961	0,934	0,968

Как можно заметить по показателям Accuracy, AUCROC, Gini общая способность предсказывать объекты из большего класса выше, чем в нашей работе. Если сравнивать по способности предсказывать объекты из малого класса, то по показателям Precision 1 модель данного автора предсказывает лучше, но показатели Recall 1 и F1-мера в нашей работе выше.

Полученные результаты в дальнейшем можно улучшить, используя другие методы балансирования данных или оптимизируя параметры моделей классификаторов.

Вывод по основной части

В ходе проведенной работы был проведен литературный обзор, по теме исследования.

Построены модели кредитного скоринга для исходных (несбалансированных) и сбалансированных различными методами данными.

В работе проведен анализ влияния несбалансированности на результат прогнозирования случаев дефолтов по исходным данным. Для использованного набора данных было проведено исследование и выяснено, что при несбалансированных и сбалансированных данных метод градиентного бустинга дает наиболее точные оценки, по сравнению с остальными. Ансамбль метода градиентного бустинга и метода сэмплирования ENN оказался самой оптимальной метамоделью. Точность предсказания данной модели по коэффициенту Джини составило 0,851, при среднем гармоническом между точностью и полнотой для случаев дефолта равной 0,519 и показателем точности предсказывания случаев дефолта 0,539. При несбалансированных данных показатели были следующие: коэффициент Джини также равен 0,851, а среднее гармоническое между точностью и полнотой для случаев дефолта равна 0,471 и показатель точности предсказывания случаев дефолта 0,372.

4. Концепция стартап-проекта

4.1. Описание продукта как результата НИР

В настоящее время автоматизируются различные рутинные работы человека. Данные изменения касаются и сектора кредитования, где необходимо оценивать риск невозврата кредита заемщиком. Однако не все кредитные организации могут себе позволить создавать собственные программы для такой автоматизации.

Результатом НИР является программа для оценки риска невозврата займа заемщиком. Данная программа является основой продукта «MONScore» - это онлайн-сервис по оценке риска невозврата займов для МФО. Программа реализована на основе методов машинного обучения, а именно на классификаторах, так как они могут находить связи и закономерности в большом объеме данных, в отличие от статистических. Также среди моделей есть ансамблевые методы, которые позволяют получать более точные результаты.

Онлайн-сервис работает по модели SaaS-сервиса (Software as a Service – «программное обеспечение как услуга» или «информация как сервис») – предоставления программного обеспечения по подписке, то есть сама программа будет находиться в облачном сервисе и ничего на компьютер устанавливать не нужно. Доступ к онлайн-сервису происходит через браузер. Наш продукт предлагает следующую услугу: первым делом мы настраиваем модель оценки заемщика в индивидуальном порядке под требования заемщика, а затем предоставляем клиентам доступ к настроенной модели.

Программа работает в два этапа. Первый этап - это индивидуальная настройка самой программы по известному алгоритму (программа индивидуальна для каждого клиента, так как итоговая модель обучается на исторических данных клиента). На вход подаются исторические данные по заемщикам, возвративших и невозвративших кредиты (кредитный портфель), затем данные обрабатываются, выделяются значимые параметры оценки, строятся разные модели, настраиваются и из них выбирается наиболее

оптимальный вариант модели кредитного скоринга, отвечающего требованиям клиента. На выходе получаем готовую модель – программа, по которой оцениваются заемщики.

Второй этап – использование построенной модели. Через веб-интерфейс заполняется анкета заемщика, далее данная информация направляется в программу для подтверждения данных и сбора дополнительной информации по заемщику и потом все полученные данные направляются в программу оценки. Программа получает данные, обрабатывает запрос и на выходе подает значение вероятности риска невозврата займа с рекомендацией об выдаче, отказе или подбору других условий выдачи кредита.

4.2. Интеллектуальная собственность

Онлайн-сервис включает в себя несколько составляющих объектов: сайт (посредством которого осуществляется деятельность), логотип (товарный знак) сервиса и программное обеспечение (ПО, к которому клиенты получают доступ). Отдельного закрепления и защиты прав интеллектуальной собственности SaaS-сервисов нет, но можно защитить отдельные объекты онлайн-сервиса.

В ходе реализации проекта разрабатывается программа оценки риска невозврата займа. Для того, чтобы защитить программу, необходимо зарегистрировать его как программу для ЭВМ в федеральном органе исполнительной власти по интеллектуальной собственности Российской Федерации – Роспатент.

Логотип (товарный знак) сервиса можно защитить, зарегистрировав его в Роспатенте. Для этого необходимо подать заявку в национальное/региональное ведомство по товарным знакам и оплатить требуемые пошлины. Стоит отметить, что товарные знаки регистрируются только на классы товаров или услуг, входящие в список МКТУ (Международная классификация товаров и услуг).

Также в ходе реализации проекта необходимо будет создать веб-сайт. Сайт является сложным объектом права, так как он состоит из нескольких охраняемых результатов интеллектуальной собственности (программный код, тексты, фото-, видео-, аудиоматериалы, база данных и т.п.). Для защиты авторских прав на сайт необходимо приобрести у разработчиков сайта исключительные права на все его составные элементы по договору, а также рекомендуется зафиксировать авторское право на сайт помощью депонирования – загрузки всех его страниц на хранение в виртуальную ячейку распределенного реестра данных IPChain. Программный код и базу данных сайта можно зарегистрировать в Роспатенте. Также стоит зарегистрировать доменное имя сайта как товарный знак, а еще не лишним будет установить на сайте знак Copyright.

4.3. Объем и емкость рынка

Рынок продукта – это рынок услуг оценок кредитоспособности заемщика. Так как информации о данном рынке мало, то также будем рассматривать рынок финансовых технологий. Продукт является инструментом при принятии решений в сфере микрофинансирования. То есть потребителями данного рынка являются микрофинансовые организации.

Объем рынка будем считать с расчетом на год по методу TAM-SAM-SOM «снизу-вверх» и «сверху-вниз»: TAM (Total Addressable Market) – общий объем целевого рынка, SAM (Serviced/Serviceable Available Market) – доступный объем рынка и SOM (Serviceable & Obtainable Market) – реально достижимый объем рынка.

«Снизу-вверх». Для расчета объема рынка определим количество микрофинансовых организаций (МФО) в России. Нам необходимо рассматривать легально действующие МФО. Регулятором финансового рынка страны, в том числе деятельности МФО, является Центробанк России. Поэтому список таких МФО можно найти в Государственном реестре микрофинансовых организаций. Данный реестр находится в свободном

доступе на официальном сайте Центрального банка Российской Федерации (ЦБ РФ).

Для того, чтобы увидеть информацию необходимо сначала скачать файл с базой данных в формате Excel. Документ содержит четыре листа. На первом размещен общий список действующих МФК и МКК, на втором – только МФК, на третьем – только МКК. Четвертый лист включает информацию о микрофинансовых организациях, исключенных ЦБ РФ из реестра. Реестр часто обновляется сотрудниками регулятора в режиме онлайн, поэтому на сайте всегда представлены актуальные данные на момент посещения сайта.

По состоянию на 15.05.2022 в реестре зарегистрировано всего 1283 микрофинансовых организаций: из них 37 микрофинансовых компаний и 1245 микрокредитных компаний. Данное количество является общим объемом целевого рынка, то есть TAM = 1283. Потребителями нашего продукта являются не все микрофинансовые организации, мы нацелены на микрофинансовые организации, которые выдают займы физическим лицам только онлайн или и онлайн и оффлайн. Из информационно-аналитического материала от ЦБ РФ «Обзор ключевых показателей микрофинансовых институтов» за 3 квартал 2021 года получаем примерное количество МФО, которые в течении 9 месяцев 2020-го года и 9 месяцев 2021-го года выдавали займы физическим лицам (и офлайн и онлайн). Их количество составило 724 МФО.

Допустим, что из этого количества МФО у около 70% есть свои программы оценки рисков невозврата займов заемщиком. Получается, что допустимый объем рынка SAM = 217 МФО.

Для определения достижимого объема рынка учитываем, что нужно до продажи необходимо пройти следующие этапы: связаться с организациями, выйти на ЛПР, презентовать продукт, предоставить тестовый продукт, заключить сделку. То есть до конкретной продажи у нас есть 4 этапов, и будем считать, что в итоге первого этапа будет 50% отказа, в итоге второго и третьего этапа процент отказа от продукта будет равен 30%, а в итоге четвертого будет

10% (расчет продаж первого года). Тогда в итоге останется 217*0,5*0,7*0,7*0,9 = 48 компаний. Тогда получится, что достижимый объем рынка в первый год SOM = 48 МФО. Мы рассчитали по пессимистичному прогнозу.

«Сверху-вниз». По данным «Анализа рынка финансовых технологий (FinTech) в России», подготовленного BuisinessStat в 2022 году, общий объем рынка финансовых технологий для кредитных организаций (банков) в 2021 году составило 9 млрд. рублей. Так как нас интересуют именно технологии, связанные со скорингом для МФО, то предположим, что из общего объема оно составляет где-то 3%. Тогда общий объем рынка TAM = 9 млрд. руб. * 0,03 = 270 млн. рублей.

Допустим, что приблизительная доля МФО, готовых приобрести наш продукт 40% из общего объема рынка. Тогда допустимый объем рынка SAM = 270 млн. руб. * 0,4 = 108 млн. рублей.

Для определения достижимого объема рынка найдем, сколько примерно конкурентов имеется на нашем рынке. Их количество около 15-ти, учитывая не только прямых, но и косвенных конкурентов. Значит для определения SOM делим допустимый объем рынка на 15. В итоге получаем SOM = 108 млн. руб. : 15 = 7,2 млн. рублей.

Полученный объем достижимого рынка насчитывает 65 МФО, если в деньгах, то 7,2 млн. рублей, но это пока примерное допущение на данный период. Так как количество МФО, переходящих на онлайн-займы с каждым кварталом больше, то и количество потенциальных потребителей может увеличиться.

Данные вычисления объема рынка являются приближенными и рассчитаны из личных допущений, так как точной информации об рынке услуг оценки рисков невозврата займов не имеется.

4.4. Анализ современного состояния и перспектив развития отрасли

Рынок услуг по оценке кредитоспособности заемщиков на данный момент является развивающимся рынком в силу того, что количество МФО, переходящих на выдачу онлайн-займов увеличивается с каждым годом. А это означает, что для стабильной работы таким организациям понадобится развивать технологическую инфраструктуру, которая будет помогать снижать риски финансовых потерь. В особенности это касается скоринговых систем, которые оценивают риск невозврата займа тому или иному заемщику, ведь от возвратности займов напрямую зависит доход МФО. Скоринговые системы помогают решить две проблемы: автоматизацию процесса оценки заемщика и увеличение точности оценки. Последнее имеет место быть, так как при выдаче онлайн-займов сотрудник не обладает достаточной информацией для оценки и не может лично взаимодействовать с заемщиком, а при использовании скоринговых систем есть возможность найти дополнительные данные по заемщику и автоматически учитывать их при оценке заемщика. Также переход на выдачу онлайн-займов позволяет увеличить доходы благодаря продаже дополнительных продуктов и услуг (СМС-информирование, страхование и др.).

В сфере скоринга сейчас развиваются два направления: технологическое (новые источники данных и применение новых алгоритмов) и регуляторные (повышение стабильности моделей и снижение модельного риска). В этих целях используются методы машинного обучения и потенциал их использования только увеличивается. Основной упор делается на автоматизацию процессов принятия решений по кредитной заявке. В настоящий момент из-за изменений экономической и рыночной ситуации скоринговые модели потеряли свою предсказательную силу из-за того, что они не настроены под кризисные данные и не хватает новых данных для их переобучения. Но как только ситуация стабилизируется и появятся достаточно

данных, то потребность в построении моделей под новые данных только возрастет.

Сфера микрофинансирования также терпит множество изменений из-за регуляторных ограничений. Например, планируется снизить максимальную ежедневную процентную ставку и максимальный размер суммы всех платежей по договору для PDL-займов. А это значит, что бизнес-модель PDL-займов в будущем может показать отрицательную рентабельность. Из-за этого сейчас многие организации перестраивают свои процессы под другие виды займов, в особенности на IL-займы. При перестройке бизнес-процессов также будут меняться модели оценки заемщиков, поэтому программа оценки кредитоспособности заемщиков будет актуальна и в будущем.

4.5. Планируемая стоимость продукта

Для определения планируемой стоимости продукта воспользуемся методом расчета юнит-экономики. В данном проекте ест два юнита - базовые доходные единицы: индивидуальная настройка программы и оценка одного заемщика.

1. *Индивидуальная настройка программы для одного клиента.* В данном случае юнитом является клиент, оплачивающий настройку программы. Стоит отметить, что выплата разовая, но также необходимо сопровождение программы, которое предоставляется клиентам бесплатно, но в себестоимость учитывается. Оно будет включать в себя базовый оклад разработчика программы. Себестоимость такой настройки программы будет включать в себе премию разработчика программы оценки риска невозврата займов за одну продажу, затраты на подключение интернета и платежную комиссию. Также будут необходимы затраты на привлечение данного клиента и его удержание (премия менеджера по продажам за одну продажу и затраты на рекламу). Менеджер по продажам будет работать на удаленно. К зарплатам сотрудников также необходимо

учитывать районный коэффициент 1.3 и страховые взносы в размере 30,2%. Также необходимо рассчитать постоянные издержки для определения размера инвестиций. Считаем, что в месяц при 8 часовом режиме работы тратится 196 руб. за электроэнергию. Приведем вычисления с расчетом на год работы (табл. 20).

Таблица 20 - Юнит-экономика индивидуальной настройки программы

Юнит-экономика	**Описание**	**Сумма, руб**
Прибыль от продажи (Revenue)	Доход, полученный от одного клиента (стоимость)	60000
Затраты на обеспечение продажи (COGS)	Премия разработчику за каждую продажу (фиксированная)	15000*1,3+15000*0,32 = 24300
	Затраты на подключение интернета	50 + 2990 = 3040
	Затраты на платежную комиссию	60000*0,025 = 1500
		28840
Валовая прибыль (Gross Profit)	Как заработали на продаже	60000 – 28840 = 31160
Затраты на привлечение одного клиента (CAC)	Премия менеджера по продажам (за привлечение одного клиента, 35% от оклада)	10500*1,3+10500*0,32 = 17010
	Затраты на рекламу	3000
		17310
Маржинальная прибыль (Contribution Margin)	Деньги после продажи услуги и привлечения клиента	31160-17310 = 13850
Постоянные издержки в год	ЗП разработчика настройки индивидуальной модели (оклад), ЗП менеджера по продажам (оклад)	(10000*1,3+10000*0,32 + 30000*1,3+30000*0,32) * 12 = 777600
	Затраты на электроэнергию в год	196*12 = 2352
	Затраты на оплату интернета	675 * 12 = 8100
		786452
Инвестиции	Сколько инвестиций необходимо для продажи одного юнита	786452 – 13850 = 772602

2. *Обработка одного запроса об оценке риска невозврата.* После заключения сделки на первую услугу необходимо будет оплачивать за оценку одного заемщика. Рассчитаем юнит-экономику с учетом минимального возможного объема обрабатываемых заявок (табл. 21). За себестоимость обработки одной оценки будем брать стоимость кредитной истории (КИ) из БКИ и платежные комиссии. А стоимость одного запроса определили из конкурентных предложений на рынке.

Таблица 11 - Юнит-экономика обработки одной заявки

Юнит-экономика	**Описание**	**Сумма, руб**
Прибыль от продажи (Revenue)	Стоимость обработки одной заявки	20
Затраты на обеспечение продажи (COGS)	Стоимость КИ из БКИ	3,46
	Платежная комиссия	20 * 0,025 = 0,5
		3,96
Валовая прибыль (Gross Profit)	Сколько заработали на продаже	20 – 3,96 = 16,04
Затраты на привлечение одного клиента (CAC)	Затрат нет, так как программой пользуются уже привлеченные клиенты	0
Маржинальная прибыль юнита (Contribution Margin)	Деньги после продажи услуги и затраты на маркетинг	16,04 – 0 = 16,04
Постоянные издержки в год	Аренда онлайн-сервера	11980 * 12 = 143760
	ЗП специалиста по тех. поддержке	(33000 * 1,62) * 12 = 641520
		785280
Инвестиции	Создание веб-сервиса	250000
	Инвестиции для одного юнита	785280 – 16,04 = 785263,96
		1035263,96

В итоге стоимость нашего онлайн-сервиса «MONScore» за первый год с расчетом на минимальное количество заявок составляет 60000+20*2000*12 = 540000 рублей, а в следующие годы 480000 рублей. А себестоимость составляет 141190 рублей в год. Такая разница между себестоимостью и стоимостью объясняется за счет разницы себестоимости и стоимости оценки одной заявки. Стоимость оценки одной заявки взята по сравнению с конкурентами.

4.6. Конкурентные преимущества создаваемого продукта, сравнение технико-экономических характеристик с отечественными и мировыми аналогами.

На рынке оценки кредитоспособности заемщиков имеется достаточное количество решений. Но способы работы, оценки рисков у них разные. Также имеется много мировых аналогов, но в силу того, что кредитные решения в других странах принимаются немного по другим показателям, то будем

рассматривать только российский рынок. Сравним наш продукт с двумя решениями, известными в России - «Скоринг для МФО 2.0» от БКИ «Эквифакс» и «Loginom Scorecard Modeler». Сравним данные решения по следующим параметрам: быстрота интеграции, наличие индивидуализации моделей, технологии решения, стоимость и гибкость работы:

Таблица 22 - Конкурентный анализ

Решения/характеристики	**Быстрота интеграции**	**Наличие индивидуализации моделей**	**Технологии решения**	**Стоимость**	**Гибкость работы**
Наш проект «MONScore»	7 дней на настройку программы и подключение к сервису через браузер	+	Машинное обучение, различные модели кредитного скоринга	0,54 млн. рублей в год	+
«Скоринг для МФО 2.0» от БКИ «Эквифакс»	Готовая скоринговая карта	-	Математическая модель	0,24 млн. рублей	-
«Loginom Scorecard Modeler»	Срок внедрения 6 месяцев	+/-	Методы машинного обучения, одна модель (логистическая)	0,9 млн. рублей + затраты на обучение сотрудников пользованию + затраты на внедрение	+/-

По всем показателям наш проект показал лучшие результаты, но также стоит учитывать, что если объем заявок будет больше, то стоимость получится больше чем у конкурентов. Это связано с тем, что наша программа имеет более гибкий вариант настройки моделей, все сопутствующие затраты к онлайн-сервису уже включены в стоимость. Например, «Loginom Scorecard Modeler» - это программа, в котором вы сами строите модели скоринга, то есть нужны будут дополнительные расходы на внедрение и обучение. А вот решение от «Эквифакс» это готовая скоринговая карта, то есть она не может адаптироваться под изменяющиеся условия.

По технологии решения оба конкурента используют машинное обучение (БКИ «Эквифакс» предположительно использует эту же технологию, точных данных об этом не найдено). Наша программа выигрывает у конкурентов, так

как наш продукт предлагает несколько различный моделей и из них выбирается наиболее подходящий под специфику работы МФО.

4.7. Целевые сегменты потребителей

Под целевым сегментом на данный момент у нас являются МФО России. Были проведены проблемные интервью с 4 микрофинансовыми организациями, базирующимися в Томске (ООО МКК «СМАЛ», ООО МКК «ТДК», ООО МКК «Автомик» и ООО МКК «Синий Слон»), и экспертные интервью с двумя специалистами в области моделей кредитного скоринга из компаний МТС и Эконофизика.

На основе общения с вышеперечисленными МФО было выяснено, что компании их типа не подходят под нашу целевую аудиторию, что прояснило с какими организациями стоит работать. Выяснено, что автоматизированными системами для оценки кредитоспособности заемщиков пользуются МФО с потоком заявок физических лиц больше 2000 потенциальных заемщиков в месяц. Преимущественно такой поток заявок характерен для МФО с онлайн-займом для физических лиц.

При выдаче онлайн-займа при оценке кредитоспособности заемщика используются только анкетные данные, без возможности вживую оценить внешний вид и состояние человека, поэтому вероятность невозврата займов увеличивается по сравнении с оффлайн-займом. Если в оффлайн-займах можно гибко изменять процентную ставку в целях снижения рисков финансовых потерь от невозврата займов, то в онлайн-займах для того, чтобы оставаться конкурентоспособным и привлекать больше заемщиков необходимо предлагать низкие процентные ставки, что увеличивает риски финансовых потерь. Данные риски можно минимизировать путем отсеивания «плохих» заемщиков. Для этого необходимо точно оценить риск невозврата займа. Но также некоторые МФО работают в режиме и онлайн и оффлайн – такие МФО также пользуются методами скоринга.

В итоге приходим к выводу о том, что для нашего продукта целевым сегментом являются МФО с оналйн-займом или с онлайн и оффлайн займом для физических лиц без залога на сумму не свыше 100 тыс. рублей и со сроком не более 12-ти месяцев и потоком потенциальных заемщиков свыше 2000 человек в месяц. В качестве потенциальных покупателей, подходящих под данное описание целевого сегмента, можно отметить ООО МКК «КапиталЪ-НТ» и ООО МКК «4финанс».

4.8. Бизнес-модели проекта. Производственный план и план продаж

Бизнес-модель эффективнее и понятнее описать по модели А. Остервальдера, так как данный инструмент помогает структурированно описать разные бизнес-процессы проекта. Он представляет собой схему из 9 блоков. Таблица бизнес модели по Остервальдеру приведена ниже (табл. 23). А теперь рассмотрим по отдельности каждый блок.

Потребительские сегменты. Продукт направлен на B2B рынок и выделен один потребительский сегмент – микрофинансовые организации, выдающие займы физическим лицам без залога на сумму не свыше 100 тыс. рублей и со сроком не более 12-ти месяцев и потоком обрабатываемых заявок свыше 2000 человек в месяц.

Данный сегмент будет разделяться на четыре группы, так как услуги для МФО предоставляются по четырем видам займов: PDL (Pay Day Loans) – займы «до зарплаты», выдаются до 30 тыс. рублей и сроком до 30 дней; IL (Installment Loan) – среднесрочные займы сроком от 30 дней, он делится на IL-займы до 30 тыс. рублей и IL-займы свыше 30 тыс. рублей; POS (Point of Sale) – целевые займы на приобретение товаров и услуг. Выбрано такое разделение МФО, потому, что для оценки заемщиков по этим видам займов используются разные скоринговые модели с учетом специфики займов и не все МФО предоставляют услуги по всем четырем видам займов.

Ценностное предложение. Продуктом является онлайн-сервис оценки риска невозврата займа заемщиком. Потребителям предлагаются следующие ценности: индивидуальная программа, ускорение процесса принятия решения о выдаче займа, увеличение доходов за счет снижения потерь из-за невозвратных займов, увеличения объема «хороших» заемщиков, точность оценки, удобство и быстрота внедрения в систему, предоставление бесплатного демо-доступа к программе для оценки 500 заемщиков и бесплатная техническая поддержка и валидация моделей.

Каналы сбыта. Клиенты могут узнать о предоставляемых нами услугах через веб-сайт онлайн-сервиса, тематические конференции и холодные звонки.

Клиенты могут приобрести программу на использование после подачи заявки на приобретение услуг через веб-сайт (облачный сервис), с ними свяжутся эксперты для уточнения требований к разработке индивидуальной модели, обговаривается цена и сроки настройки программы под их данные, затем им передается доступ к программе на использование. Или же с клиентами можно связаться посредством холодного звонка, затем договориться с ними о презентации продукта, предоставлении тестового периода.

Взаимоотношения с клиентами. Клиенты обеспечиваются всем необходимым для работы с облачным сервисом, и они самостоятельно работают с нашими программами. Если возникают проблемы, то можно обращаться в техническую поддержку через электронную почту или быструю форму отправки запроса на сайте. Также перед приобретением программы можно воспользоваться бесплатным демо-доступом к программе.

Потоки доходов. Поступление доходов происходит по двум способам: индивидуальная настройка моделей оценки заемщиков (разовая выплата) и плата за количество оцениваемых заемщиков в месяц. Оплата за объем заявок происходит в конце месяца. Стоит учитывать, что если компания обработает

меньше 2000 заявок, то в любом случае платит за 2000 обработок, так как данный объем оценивается как минимальный объем.

Ключевые ресурсы. Для функционирования бизнес-модели необходимы интеллектуальный ресурс в виде алгоритма работы программы для оценки заемщиков, рабочий персонал в составе разработчика программы оценки, менеджер по продажам и специалист по технологической поддержке, а также платформа облачного сервиса и исторические данные для построения моделей.

Ключевые деятельности. Для реализации ценностных предложений необходимо настраивать индивидуальные модели, проводить мониторинг и обеспечивать их адекватность. А также необходимо быть всегда готовым оказать техническую поддержку при появлении проблем и вопросов при работе с онлайн-сервисом.

Ключевые партнеры. Для создания ценностного предложения необходимо будет сотрудничать с Бюро кредитных историй (БКИ) для получения дополнительных данных о заемщиках, с интернет провайдером и с платформой, поддерживающей облачный сервис. Данный сервис нужен для того, чтобы разместить на нем все программные обеспечения и базы данных, которые будут реализованы в процессе работы онлайн-сервиса.

Структура издержек. Расходы проекта состоят по отдельности из двух групп затрат: на индивидуальную настройку программы и обработку одного запроса.

К переменным издержкам первой группы относятся: премия разработчику программы оценки риска невозврата займов для настройки модели, затраты на подключение интернета, комиссии за платежную систему, затраты на маркетинг. А к постоянным издержкам относятся: зарплата (оклад) разработчика программы оценки риска невозврата займов на сопровождение, зарплата менеджера по продажам, оплата интернета и затраты на электроэнергию. Также требуются инвестиции для перекрытия постоянных издержек.

К переменным издержкам второй группы затрат относятся: покупка кредитных историй из БКИ, платежные комиссии, создание веб-сервиса. К постоянным издержкам относятся: аренда облачного сервера и зарплата специалиста по технической поддержке. Также требуются инвестиции.

Таблица 23 - Бизнес-модель по А. Остервальдеру

<table>
<tr>
<td rowspan="2">Ключевые партнёры

• БКИ – для получения дополнительных данных
• Платформа онлайн-сервера
• Интернет-провайдер</td>
<td>Ключевые виды деятельности

• Настройка программы оценки
• Мониторинг адекватности моделей
• Техническая поддержка сервиса
• Оценка заемщика</td>
<td rowspan="2">Ценностные предложения

• Индивидуальная модель
• Точность оценки
• Увеличение доходов
• Ускорение процесса принятия решения
• Онлайн-сервис
• Быстрое и удобное внедрение в систему МФО
• Бесплатная техническая поддержка сервиса
• Бесплатная валидация моделей скоринга
• Бесплатный демо-доступ</td>
<td>Взаимоотношения с клиентами

• Служба поддержки
• Самообслуживание
• Бесплатный демо-доступ</td>
<td rowspan="2">Потребительские сегменты

B2B рынок

МФО:

• PDL займы
• IL займы до 30 тыс. рублей
• IL займы свыше 30 тыс. рублей
• POS займы</td>
</tr>
<tr>
<td>Ключевые ресурсы

• Платформа
• Рабочий персонал
• Алгоритм программы оценки
• Исторические данные</td>
<td>Каналы сбыта

• Веб-сайт MONScore
• Холодные звонки
• Конференции</td>
</tr>
<tr>
<td colspan="3">Структура издержек

• Переменные издержки: оплата труда разработчика за настройку модели, платежные комиссии, затраты на подключение интернета, покупка кредитных историй БКИ, затраты на маркетинг, создание сайта
• Постоянные издержки: зарплата сотрудникам по найму, аренда облачного сервера, оплата интернета, электроэнергии.
• Инвестиции: на перекрытие затрат в первый год</td>
<td colspan="3">Потоки поступления доходов

Бизнес-модель SaaS
• Настройка индивидуальной модели (разовая выплата)
• Обработка одного запроса об оценке заемщика.</td>
</tr>
</table>

Исходя из вышеописанной бизнес-модели составлен план производств и продаж (табл. 24).

Таблица 24 - План продаж и производства на первый год в тыс. руб.

Показатели	**1 кв**	**2 кв**	**3 кв**	**4 кв**	**1 год**
Количество МФО	2	3	4	5	14
Оплата за настройку программы	60 * 3 = 120	60 * 4 = 180	60 * 5 = 300	60 * 6 = 360	840
Оценка заемщика	0,02 * 2000 * 3 * 2 = 240	0,02 * 2000 * 3 * (2+3) = 600	0,02 * 2000 * 3 * (2 + 3 + 4) = 1080	0,02 * 2000 * 3 * (2 + 3 + 4 + 5) = 1680	3600
Итого доходов	**360**	**780**	**1320**	**1980**	**4440**

При соблюдении данного плана продаж и производства в год можно выручить 4,44 млн. рублей. Теперь рассмотрим какую чистую прибыль можно получить в течении года (табл. 25). Данные для расчета взяты из таблиц 1 и 2.

Таблица 25 - Оценка прибыли в тыс. руб.

Показатели	**1 кв**	**2 кв**	**3 кв**	**4 кв**	**1 год**
Итого доходов	**360**	**780**	**1320**	**1980**	**4440**
Переменные затраты:					
Премия разработчику за каждую продажу	48,6	72,9	97,2	121,5	340,2
Подключение интернета	3,04	0	0	0	3,04
Платежная комиссия (за настройку)	3	4,5	6	7,5	21
Стоимость КИ из БКИ	41,52	103,8	186,84	290,64	622,8
Платежная комиссия (за обработку одной заявки)	6	15	27	42	90
Премия менеджера по продажам	34,02	51,03	68,04	85,05	238,14
Затраты на рекламу	6	9	12	15	42
Создание веб-сервиса	250	0	0	0	250
Итого переменных затрат	**392,18**	**256,23**	**397,08**	**561,69**	**1607,18**
Маржинальный доход	**-32,18**	**523,77**	**922,92**	**1418,31**	**2832,82**
Постоянные затраты:					
ЗП разработчика настройки индивидуальной модели (оклад), ЗП менеджера по продажам (оклад)	194,4	194,4	194,4	194,4	777,6
Затраты на электроэнергию	0,588	0,588	0,588	0,588	2,352
Затраты на оплату интернета	2,025	2,025	2,025	2,025	8,1
Аренда онлайн-сервера	35,94	35,94	35,94	35,94	143,76
ЗП специалиста по тех. поддержке	160,38	160,38	160,38	160,38	641,52
Итого постоянных затрат	**393,33**	**393,33**	**393,33**	**393,33**	**1573,332**
Прибыль до н/о	-425,51	130,44	529,59	1024,98	1259,488
Налог на прибыль 20%	0	26,088	105,918	204,996	251,8976
Чистая прибыль	**-425,51**	**-321,158**	**102,514**	**819,984**	**1007,5904**

Как видим из результатов чистая прибыль составит около 1,01 млн. рублей в первый год при выполнении вышеприведенного плана продаж и производства.

Рассчитаем точку безубыточности (после продажи скольким компаниям можно выйти на положительную прибыль) по формуле:

Точка безубыточности = Постоянные затраты / (Маржинальный доход/Количество МФО) = 8 компаний

Требуемые инвестиции для реализации продукта до точки безубыточности составляют 746,668 тыс. рублей. Данные инвестиции окупятся примерно через полгода.

4.9. Стратегия продвижения продукта на рынок

Продуктом данного стартапа является услуга, онлайн-сервис по оценке кредитоспособности заемщика для МФО. Как и описывалось, на рынке оценки кредитоспособности заемщиков имеется множество конкурентов. Поэтому важно грамотно продумать стратегию выхода на рынок.

Мы уже определили, какой сегмент рынка является нашей целевой аудиторией, оценили рынок, определились со стоимостью, провели конкурентный анализ и построили бизнес-модель проекта. Осталось продумать каким образом мы будет выходить на рынок. Для этого выделим следующие способы продвижения на рынок:

- Email-рассылки – для того, чтобы найти точки касания с потенциальными клиентами им отправляются письма о возможном сотрудничестве с их компанией, коротко описываются преимущества продукта; электронный адрес клиентов можно найти на официальных сайтах МФО;
- Холодные звонки – менеджер по продажам связывается с ЛПР, рассказывает про продукт, договаривается о личной презентации продукта. Контакты также можно находить с официальных сайтов, через личные знакомства или конференции.;

— Контекстная реклама в поисковых запросах – необходимо настроить рекламу в Google Ads и Яндекс.Директе так, чтобы при запросах «кредитный скоринг», «оценка заемщика» и т.п. в первых строках появлялся наш сайт;
— Социальные-сети – для того, чтобы создать привлекательный имидж компании, с которым у потенциальных клиентов возникнет желание поработать. В собственной странице социальной сети (например, ВК) можно публиковать актуальные новости в сфере микрофинансирования, делиться рассуждениями об развивающихся технологиях, рекламировать и рассказывать про собственную продукцию. Также в социальной сети (например, ВК) на страницах группы с тематикой микрофинансирования можно выставлять рекламу продукта или предоставить интервью как эксперт в сфере скоринга. Выбирается социальная сеть ВК, так как оно является российским и многие МФО есть в данной платформе;
— Конференции – можно поучаствовать в форуме «ScoringDay X», «Summit MFO», «MFO Russia Forum» и т.д., где можно познакомиться с потенциальными клиентами и рассказать о своем решении.

5. Социальная ответственность

Введение

Объектом исследования данной ВКР являются математические модели кредитного скоринга с учетом несбалансированности данных, позволяющие оценить риск невозврата кредита (займа) заемщиком.

В данной работе представлена разработка программы оптимальной математической модели кредитного скоринга на персональном компьютере (ПК). Поэтому в данном разделе будут рассматриваться комплексы мер технического, организационного и правового характера, уменьшающие пагубное влияние работы с ПК. Также будут рассматриваться вопросы техники безопасности и пожарной профилактики, охраны окружающей среды и создания оптимальных условий труда.

Объект исследования данного раздела – рабочая зона офисного сотрудника, включая ПК, клавиатура, компьютерная мышь, стол, стул, а также само помещение в котором находится офисное помещение. Рабочая зона находится в 427А аудитории 10 корпуса. 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.

5.1. Правовые и организационные вопросы обеспечения безопасности

5.1.1. Правовые нормы трудового законодательства

В трудовом кодексе РФ содержатся основные положения отношений между работодателем и работником, включая обеспечение справедливых условий труда (безопасность и гигиена, право на отдых, ограничение рабочего времени, ежедневный отдых, нерабочие дни и отпуск), а также оплата и нормирование труда и т.д.

Согласно нормативным актам ТК РФ режим офисного работника, пользующегося персональным электронно-вычислительной машиной (ПЭВМ), в течении рабочего дня не должен превышать при пятидневной рабочей неделе 8 часов с перерывом на обед от 30-ти минут до 2-х часов, а также должны устанавливаться регламентированные перерывы. При восьмичасовой рабочей неделе перерывы составляют 15 минут через два часа после начала работы и через два часа после перерыва на обед [23].

Защита персональных данных работника подробно описана в главе 14 ТК РФ, также действия с персональными данными регулируются настоящим Кодексом и иными федеральными законами.

5.1.2. Эргономические требования к правильному расположению и компоновке рабочей зоны

Согласно нормативному документу [24] 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.

При организации рабочего места основной целью является обеспечение качественного и эффективного выполнения работы при полном использовании оборудования в соответствии с установленными сроками [25]. По санитарным требованиям и нормативам выделены следующие требования к рабочему месту:

Таблица 26 - Параметры рабочего места при работе с ПК

Параметры	Значение параметра	Реальные значения
Высота рабочей поверхности стола	от 600 до 800 мм	740 мм
Высота клавиатуры	600 – 700 мм	600 мм
Удаленность клавиатуры	Не менее 80 мм	83 мм
Удаленность экрана монитора	500 – 700 мм	600 мм
Высота сидения	400 – 500 мм	470 мм
Угол наклона монитора	0 – 30 град	12 град
Наклон подставки ног	0 – 20 град	0 град

Рабочее место в аудитории 427А соответствует санитарным требованиям и нормативам.

5.2. Производственная безопасность

5.2.1. Анализ вредных и опасных факторов, которые могут возникнуть при разработке программы

Выбор факторов был произведен из ГОСТ 12.0.003-2015 «Опасные и вредные производственные факторы. Классификация» [26]. Перечень опасных и вредных факторов, характерных для проектируемой производственной среды представлен в виде таблицы 27.

Таблица 27 - Возможные опасные и вредные производственные факторы при работе с ПЭВМ

Факторы (ГОСТ 12.0.003-2015)	Этапы работ		Нормативные документы
	Разработка	Тестирование	
Недостаточная освещенность рабочей зоны	+	+	СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*
Перенапряжение зрительных анализаторов	+	+	Р 2.2.2006-05. Гигиена труда. Руководство по гигиенической оценке факторов рабочей среды и трудового процесса. Критерии и классификация условий труда
Отклонение показателей микроклимата	+	+	СанПиН 1.2.3685-21. «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания»

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
Воздействие переменных электромагнитных полей	+	+	СанПиН 2.2.4.3359-16 Санитарно-эпидемиологические требования к физическим факторам на рабочих местах
Повышенный уровень статического электричества	+	+	ГОСТ Р 53734.1-2014 «Электростатические явления»
Повышенное значение напряжения в электрической цепи, замыкание	+	+	ГОСТ 12.1.019-2017 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты

5.2.2. Обоснование мероприятий по защите исследователя от действия опасных и вредных факторов

Недостаточная освещенность рабочей зоны

Неудовлетворительное освещение приводит к напряжению зрения, ослаблению внимания и наступлению преждевременной утомленности. Свет на рабочем месте может создать сильные тени или отблески, а также дезориентировать работающего.

В аудитории 427А имеется естественное и искусственное освещение (совмещенное). Естественное освещение односторонние боковое. Общее освещение складывается из естественного источника света и люминесцентных ламп.

При работе с ПЭВМ рабочие столы следует размещать таким образом, чтобы мониторы 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. [27]. Освещение не должно создавать бликов на поверхности экрана.

При недостаточном освещении 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.

Перенапряжение зрительных анализаторов

При разработке программы моделей кредитного скоринга требуется длительная работа с ПЭВМ, что влечет за собой перенапряжение зрительного анализатора. Эта нагрузка приводит к переутомлению функционального состояния центральной нервной системы, нервно-мышечного аппарата рук.

По нормативным документам для безопасной работы с ПЭВМ без вреда для зрительных анализаторов рекомендуется делать регламентированные перерывы и регулировать параметры монитора так, чтобы оно не оказывало негативных эффектов на зрительный анализатор. В нормативном документе [28] описано, что для категории Б1 трудовой деятельности с ПЭВМ необходимо в суммарном 50 минут перерывов. Во время перерывов рекомендуется выполнять комплексы физических упражнений, включая упражнения для глаз.

Кроме того, можно корректно регулировать основные параметры монитора (яркость, контрастность и так далее), а также частоту обновления (при частоте меньше 75 Гц глаза человека устают быстрее).

Отклонение показателей микроклимата

Оптимальные микроклиматические при воздействии на человека в течение рабочей смены обеспечивают сохранение теплового состояния организма и не вызывают отклонений в состоянии здоровья. При отклонении от норм возможно временное (в течение рабочей смены) снижение работоспособности, без нарушения здоровья.

Работа, производимая сидя и сопровождающаяся незначительным физическим напряжением, относится к категории Iа [29]. Допустимые нормы микроклимата приведены в таблице 28.

Таблица 28 – Допустимые нормы микроклимата в рабочей зоне производственных помещений

Период года	Категория работ по уровню энерготрат, ВТ	Температура воздуха, °С		Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с	
		Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин			Для диапазона температур воздуха ниже оптимальных величин, не более	Для диапазона температур воздуха выше оптимальных величин, не более
Холодный	Ia (до 139)	20,0-21,9	24,1-25,0	19,0-26,0	15-75	0,1	0,1
Теплый	Ia (до 139)	21,0-22,9	25,1-28,0	20,0-29,0	15-75	0,1	0,2

По температуре воздуха в помещении аудитория 427А соответствует нормам (23°С). В производственных помещениях, где допускаемые нормативные величины локального микроклимата поддерживать не представляется возможным, необходимо проводить мероприятия по защите работников от возможного перегревания и охлаждения. Это достигается разными способами: использование систем местного кондиционирования воздуха; регламентацией периодов работы в неблагоприятном локальном микроклимате и отдыха в помещении с микроклиматом, нормализующим тепловое состояние; уменьшение длительности рабочей смены и др.

Повышенный уровень шума

Излишний шум на рабочем месте приводит к снижению концентрации работника и замедляет скорость психических реакций. Шумовой фон в помещении возникает из-за работы компьютеров, 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 [30]. В таблице 29 приведены предельно допустимые уровни звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ в ходе программирования на рабочем месте математика-программиста.

Таблица 29 - Допустимые значения уровней звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ

Уровни звукового давления в октавных полосах со среднегеометрическими частотами (Гц)									Уровни звука и эквивалентные уровни звука (в дБА)
31,5	63	125	250	500	1000	2000	4000	8000	
Уровни звука (дБ)									50
86	71	61	54	49	45	42	40	38	

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.

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 [31]. Временные допустимые уровни ЭМП, создаваемых ПЭВМ на рабочих местах пользователей представлены в таблице 30.

Таблица 30 - 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.

Повышенный уровень статического электричества

Основным источником повышенного уровня статического электричества при работе за ПЭВМ является монитор. На экранах мониторов положительные заряды накапливаются под действием электронного пучка, создаваемые электронной лучевой трубкой [32].

При образовании заряда с большим электрическим потенциалом создается электрическое поле повышенной напряженности, которое вредно для человека. У людей, работающих в зоне воздействия электростатического поля, встречаются разнообразные жалобы: на раздражительность, головную боль, нарушение сна, снижение аппетита и др. При длительном пребывании человека в таком поле наблюдаются функциональные изменения в центральной нервной, сердечно-сосудистой и других системах.

Согласно ГОСТу 12.4.011-89 «Средства защиты работающих. Общие требования и классификация» к средствам защиты от повышенного уровня статического электричества относятся: заземляющие устройства, нейтрализаторы, увлажняющие устройства, антиэлектростатические вещества и экранирующие устройства [33].

Повышенное значение напряжения в электрической цепи, замыкание

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.

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.

Для обеспечения защиты от случайного прикосновения к токоведущим частям необходимо применять следующее: изоляция токопроводящих частей, защитное заземление, зануление, защитное отключение, предупредительная сигнализация и блокировки. Также 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. [34]

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

5.3. Экологическая безопасность

Объект исследования является теоретическим, но разрабатывается в компьютере. Поэтому с точки зрения влияния на окружающую среду рассмотрим влияние компьютерной техники, использованной при его разработке.

Компьютерная техника потребляет сравнительно небольшое количество электроэнергии, поэтому по затратам на электроэнергию оно не оказывает существенной опасности для окружающего мира.

Компьютеры, утратившие потребительские свойства относятся к IV классу опасности (малоопасные отходы). Обезвреживание и размещение отходов I – IV классов опасности проводятся организациями, имеющими лицензию на осуществление этой деятельности. При неправильной утилизации компьютера может значительно пострадать экология, поэтому предлагается следующий порядок утилизации:

1) Удаление всех опасных компонентов;
2) удаление всех крупных пластиковых частей. Оставшиеся после разборки части отправляют в большой измельчитель, и все дальнейшие операции автоматизированы.
3) измельченные в гранулы остатки компьютеров подвергаются сортировке. Сначала с помощью магнитов извлекаются все железные части. Затем приступают к выделению цветных металлов, которых в ПК значительно больше.

Все полученные в ходе переработки материалы могут вторично использоваться в различных производственных процессах.

5.4. Безопасность в чрезвычайных ситуациях

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 [35].

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.

Аудитория 427А обеспечена средствами пожаротушения в соответствии с нормами и имеет: пенный огнетушитель ОП-10 – 1 шт; углекислотный огнетушитель ОУ-5 – 1 шт. Также в случае эвакуации в 10 корпусе имеются аварийные маршруты и выходы.

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.

Выводы по разделу

В результате анализа рабочего помещения студента, во время выполнения выпускной квалификационной работы была получена информация о том, что помещение соответствует всем санитарным требованиям организации работы сотрудника за персональным компьютером. Были проанализированы вопросы организации рабочего пространства, снижения влияния возникающих опасных и вредных факторов на организм человека, вопросы экологической безопасности и безопасности во время чрезвычайных ситуаций.

Список литературы

1. Алешин, В.А. и Рудаева О.О. Кредитный скоринг как инструмент повышения качества банковского риск-менеджмента в современных условиях. Terra Economics, 2012. – 2(3). – с. 27-30.
2. Никаненкова, В.В., 2012. Кредитный скоринг как инструмент оценки кредитоспособности заемщиков. Вестник Адыгейского государственного университета, 5.
3. Дисбаланс классов // Анализ малых данных URL: https://dyakonov.org/2021/05/27/imbalance/ (дата обращения: 13.04.2022).
4. X. Liua, Z. Zhanga, D. Wanga, 2021. Classification of Imbalanced Credit scoring data sets Based on Ensemble Method with the Weighted-Hybrid-Sampling, arXiv.org, URL: https://arxiv.org/ftp/arxiv/papers/2102/2102.04721.pdf (дата обращения: 12.04.2022)
5. Ensemble methods: bagging, boosting and stacking // Towards Data Science URL: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (дата обращения: 12.04.2022).
6. Модели для предсказания класса объектов // Github URL: ranalytics.github.io/data-mining/023-Models-for-Class-Prediction.html (дата обращения: 10.04.2022)
7. Прокопьева, А.А., 2018. Применение информационных технологий и математического моделирования в управлении банковскими рисками, PhD thesis, СПбГУ, Санкт-Петербург.
8. Классификация данных методом k-ближайших соседей // Loginom URL: loginom.ru/blog/knn (дата обращения: 11.04.2022)
9. Классификация данных методом опорных векторов // Habr URL: habr.com/ru/post/105220/ (дата обращения: 11.04.2022)
10. Деревья решений: общие принципы // Loginom URL: https://loginom.ru/blog/decision-tree-p1 (дата обращения: 13.04.2022)
11. Композиции: бэггинг, случайный лес // Habr URL: habr.com/ru/company/ods/blog/324402/ (дата обращения: 12.04.2022)

12. Градиентный бустинг // Habr URL: habr.com/ru/company/ods/blog/327250/ (дата обращения: 12.04.2022)
13. AdaBoost from Scratch // Towards Data Science URL: https://towardsdatascience.com/adaboost-from-scratch-37a936da3d50 (дата обращения: 12.04.22)
14. Григорян Д. А. Алгоритм обоснования операции на основе анализа данных групп пациентов: выпускная квалификационная работа бакалавра / Д. А. Григорян; Санкт-Петербургский государственный университет, Кафедра технологии программирования; науч. рук. В. Ю. Добрынин – Санкт-Петербург, 2017.
15. Метрики в задачах машинного обучения // Habr URL: habr.com/ru/company/ods/blog/328372/ (дата обращения: 13.04.2022)
16. Коэффициент Джини. Из экономики в машинное обучение // URL: habr.com/ru/company/ods/blog/350440/ (дата обращения: 13.04.2022)
17. Understanding Random Forest // Towards Data Science URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (дата обращения: 13.04.2022)
18. Дисперсионный анализ // ПОИВС URL: http://poivs.tsput.ru/ru/Math/ProbabilityAndStatistics/MathStatistics/DispersionAnalysis (дата обращения: 13.04.2022).
19. Ранжирование признаков с помощью Recursive Feature Elimination в Scikit-Learn // Хабр URL: https://habr.com/ru/company/otus/blog/528676/ (дата обращения: 13.04.2022).
20. Корреляционный анализ // Statistica URL: http://statistica.ru/glossary/general/korrelyatsionnyy-analiz/ (дата обращения: 13.04.2022).
21. Automatic Hyperparameter Tuning with Sklearn Using Grid and Random Search // Towards Data Science URL: https://towardsdatascience.com/automatic-hyperparameter-tuning-with-sklearn-gridsearchcv-and-randomizedsearchcv-e94f53a518ee (дата обращения: 12.04.2022)

22. Шеров Ш. Модель кредитного скоринга: магистерская диссертация / Ш. Шеров; Национальный исследовательский Томский политехнический университет, Инженерная школа ядерных технологий, Отделение экспериментальной физики, науч. рук. М. Е. Семенов – Томск, 2021.
23. Трудовой кодекс Российской Федерации [Текст]: от 30.12.2001 № 197-ФЗ (ред. от 25.02.2022)
24. ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования.
25. 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.
26. ГОСТ 12.0.003-2015 Опасные и вредные производственные факторы. Классификация.
27. СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*.
28. СанПиН 2.2.2/2.4.1340-03 Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.
29. СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.
30. СН 2.2.4/2.1.8.562-96 Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки.
31. СанПиН 2.2.4.3359-16 Санитарно-эпидемиологические требования к физическим факторам на рабочих местах.
32. ГОСТ Р 53734.1-2014 «Электростатические явления»
33. ГОСТ 12.4.011-89 Средства защиты работающих. Общие требования и классификация.
34. ГОСТ 12.1.019-2017 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты.
35. СНиП 2.01.02-85. Противопожарные нормы.

Приложение А

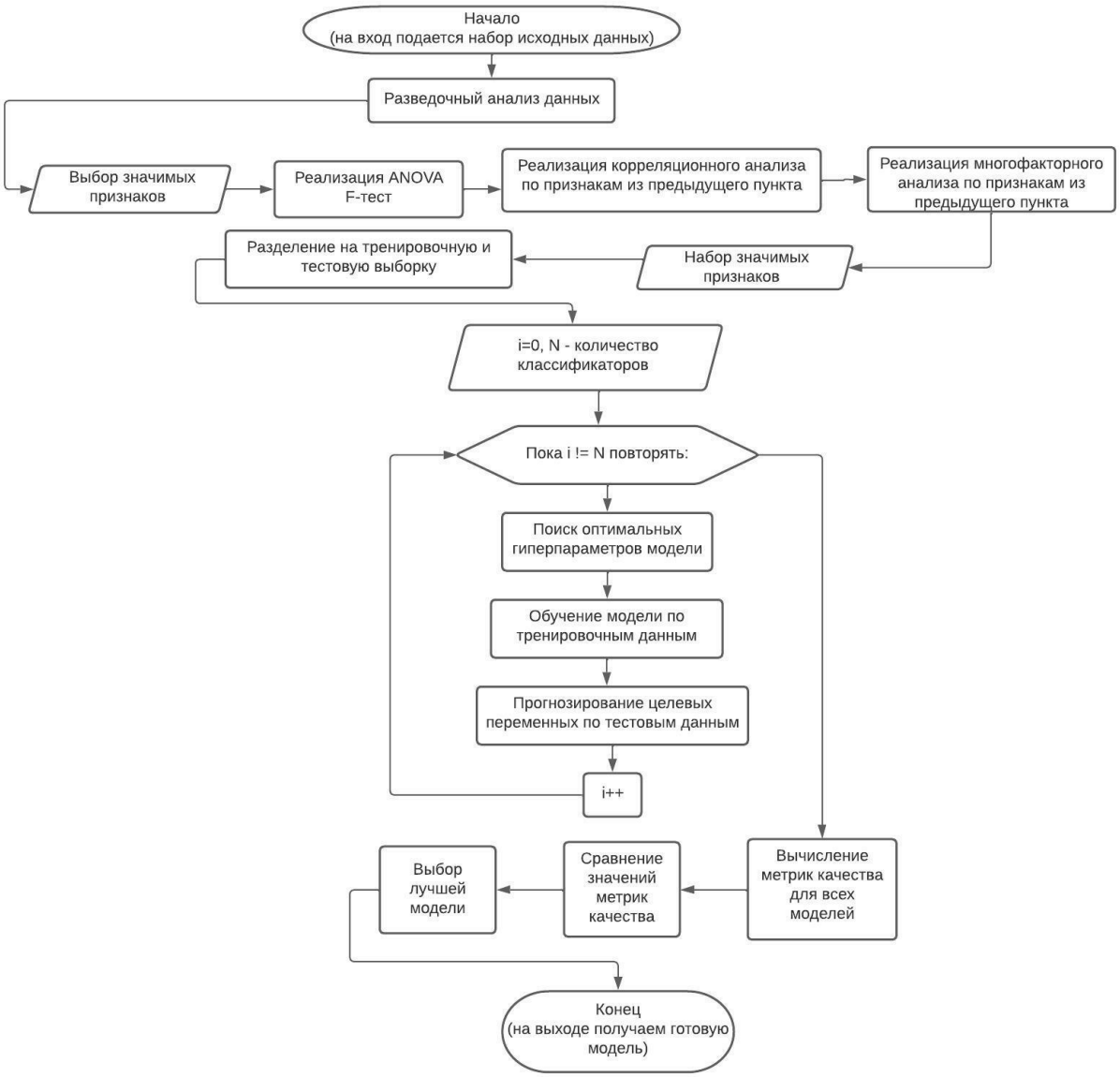

Рисунок 1 – Алгоритм с данными без балансировки

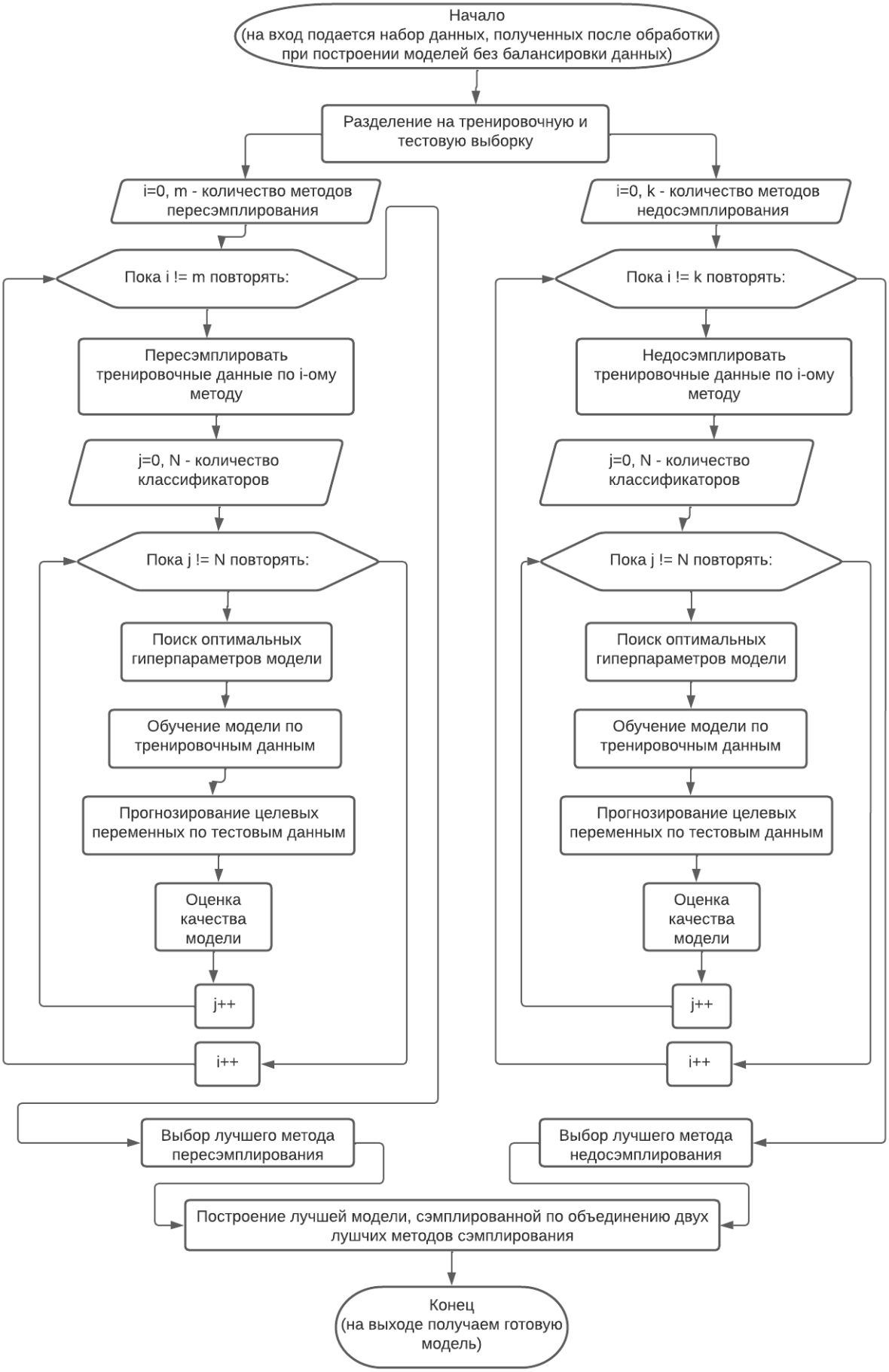

Рисунок 2 – Алгоритм с данными с балансировкой

Приложение Б

Таблица 3 – Описание признаков

1.	vintage_year,	Int	Год открытия счета
2.	monthly_installment,	Float	Ежемесячный взнос
3.	loan_balance,	Float	Остаток кредита
4.	num_bankrupt_iva,	Float	Количество случаев несостоятельности
5.	time_since_bankrupt,	Float	Время с момента последней просрочки платежа
6.	time_since_ccj,	float	Время с момента обращения в суд
7.	ccj_amount,	Float	Сумма обращений в суд
8.	num_bankrupt,	Float	Число банкротов
9.	num_iva,	Float	Число индивидуального добровольного соглашения
10.	min_months_since_bankrupt,	Float	Мин количесвто месяцев с момента банкротства
11.	pl_flag,	Int	-
12.	region,	String	Регион
13.	ltv,	Float	Отношение суммы кредита к рыночной стоимости залога
14.	arrears_months,	Float	просроченные месяцы
15.	repayment_type,	String	Тип погашения
16.	arrears_status,	Int	Статус просроченной задолженности
17.	arrears_segment,	Int	Сегмент просроченной задолженности
18.	mob,	Int	Количество месяцев, прошедших с даты выдачи кредита
19.	remaining_mat,	Int	Оставшийся срок погашения
20.	loan_term,	Int	Срок кредита
21.	live_status,	Int	Статус закрытия заема
22.	repaid_status,	Int	Статус погашения
23.	month,	int	месяц
24.	worst_arrears_status,	Int	Наихудший статус просроченной задолженности
25.	avg_mia_6m,	Float	Ср. месяцев просрочки
26.	max_arrears_bal_6m,	Float	Максимальный остаток по счету просроченной задолженности за 6 последние месяцев
27.	max_mia_6m,	Float	Максимум месяцев просрочки
28.	avg_bal_6m,	float	Средний остаток по счету за 6 последние месяцев
29.	avg_bureau_score_6m,	float	Средняя оценка бюро за 6 последних месяцев
30.	default_flag,	int	Флаг дефолта заемщика
31.	woe_max_arrears_12m,	float	Оценка максимального количества месяцев просроченной задолженности за последние 12 месяцев
32.	woe_bureau_score,	float	Оценка заемщика внешним бюро
33.	woe_cc_util,	float	Оценка использования кредитной карты
34.	woe_num_ccj,	float	Оценка количества обращений в суд
35.	woe_emp_length,	Int	Оценка стажа работы
36.	woe_months_since_recent_cc_delinq,	int	Оценка количества месяцев с момента последней просрочки по кредитной карте
37.	woe_annual_income	int	Оценка годового дохода

Приложение В

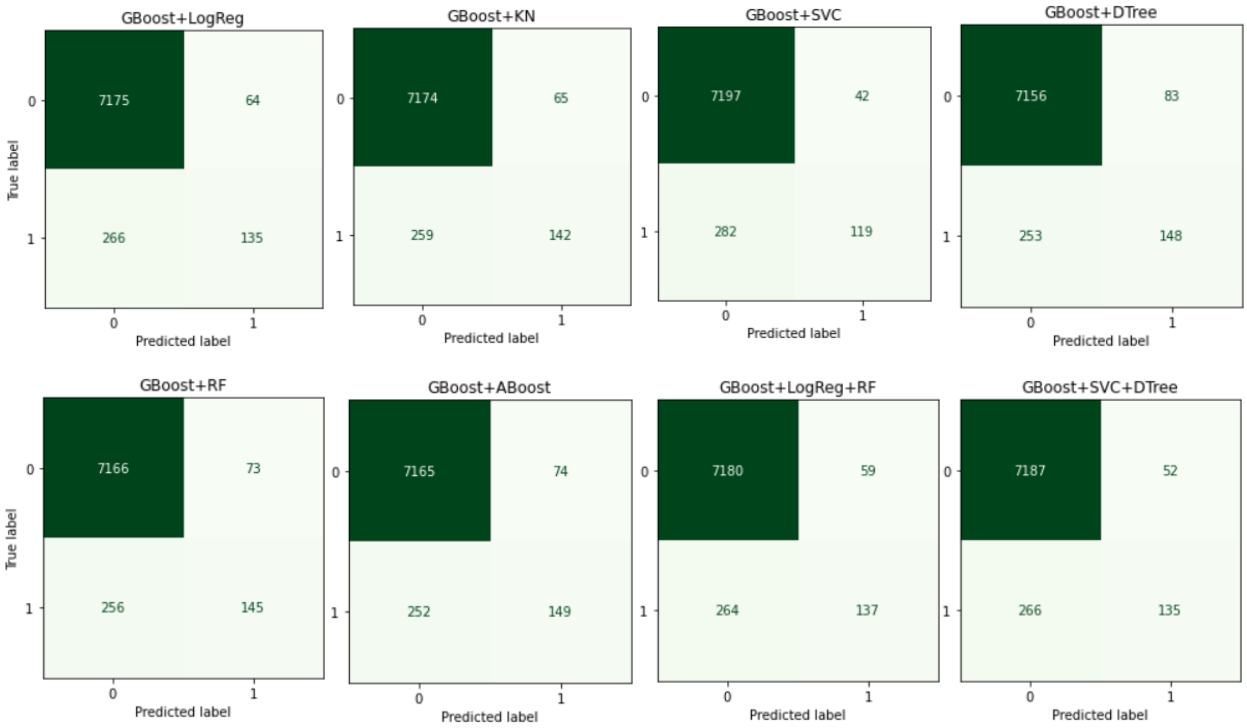

Рисунок 1 - Матрица ошибок моделей голосования по несбалансированным данным

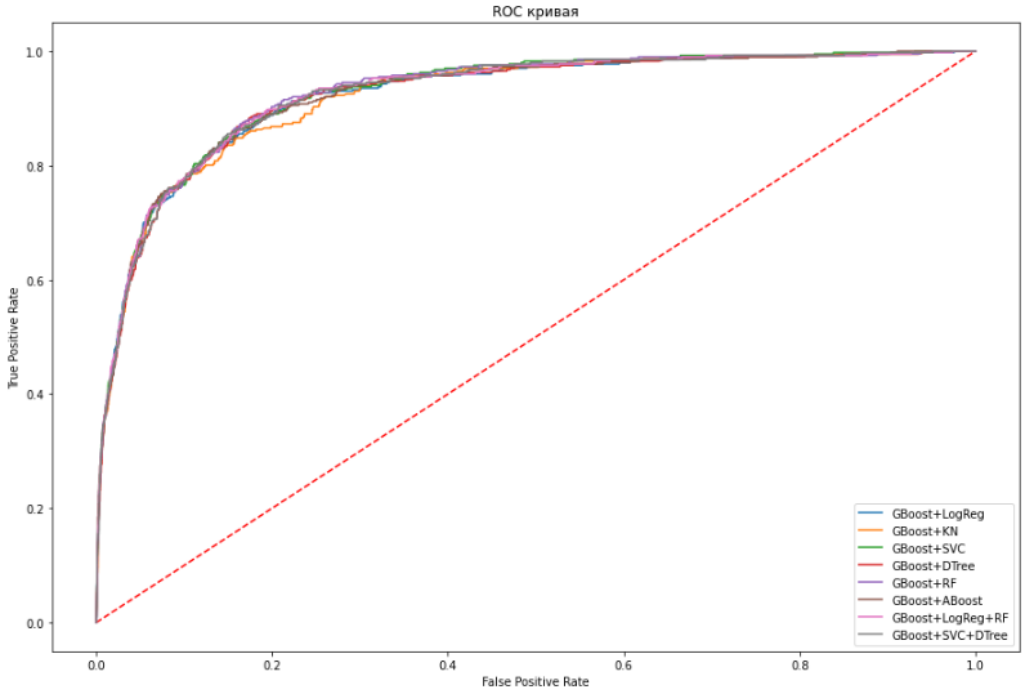

Рисунок 2 - ROC-кривая моделей голосования по несбалансированным данным

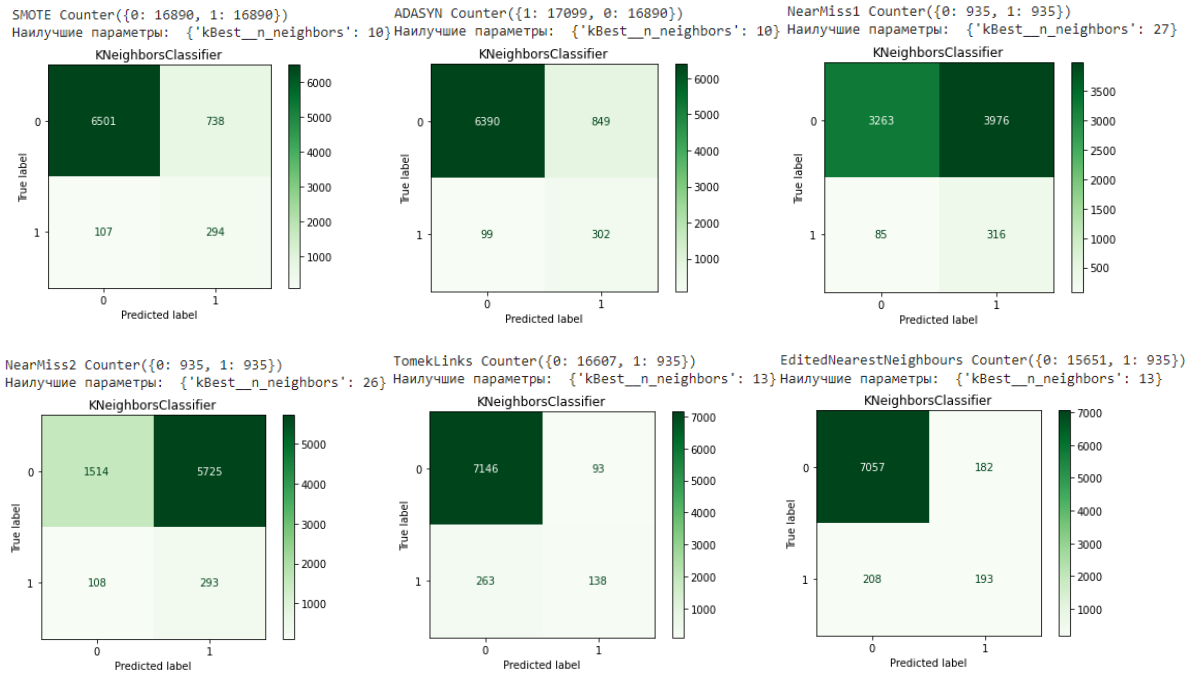

Рисунок 15 – Матрица ошибок KNeighboursClassifier при разных методах сэмплирования

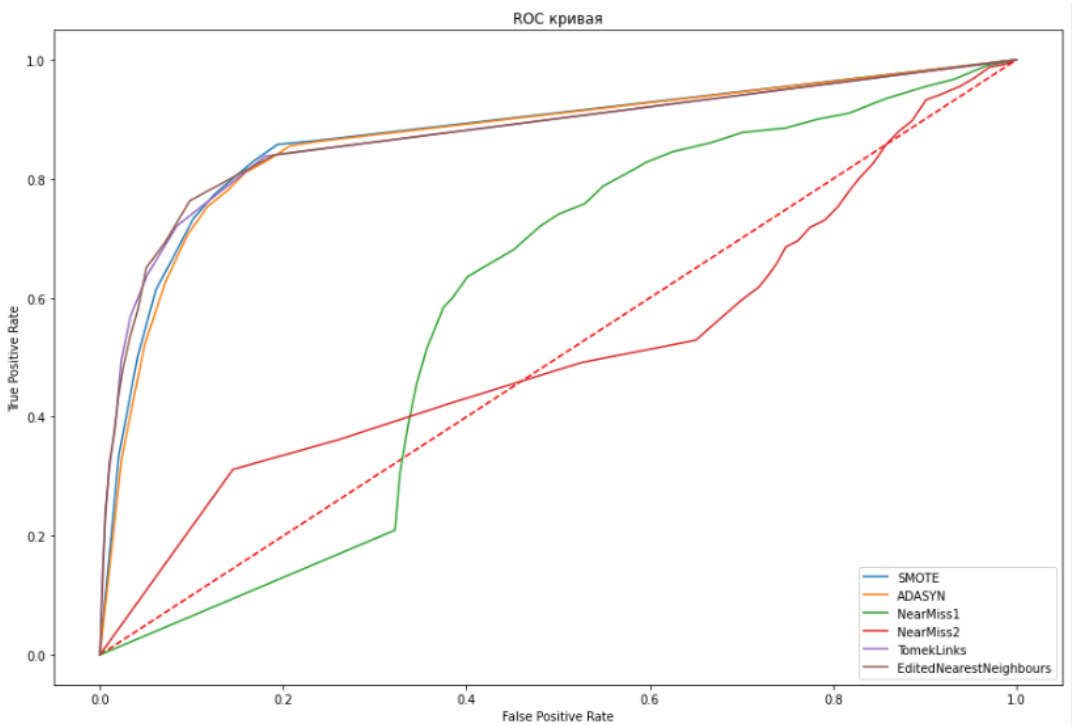

Рисунок 16 - ROC-кривая KNeighboursClassifier при разных методах сэмплирования

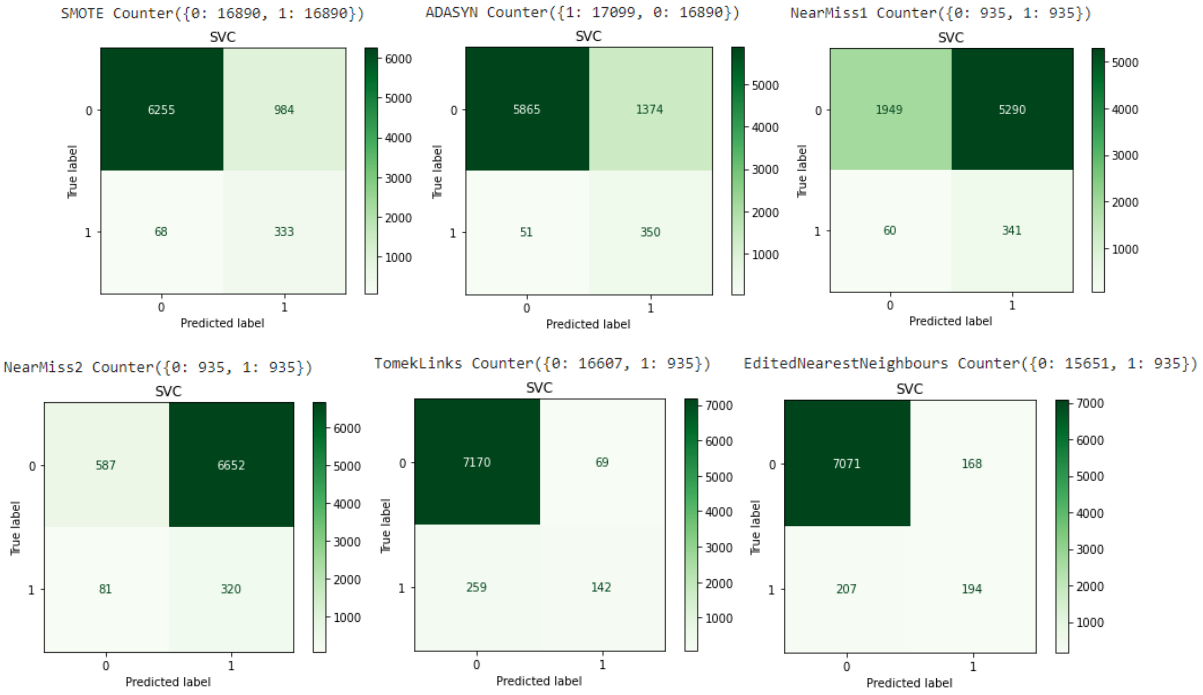

Рисунок 17 - Матрица ошибок SVC при разных методах сэмплирования

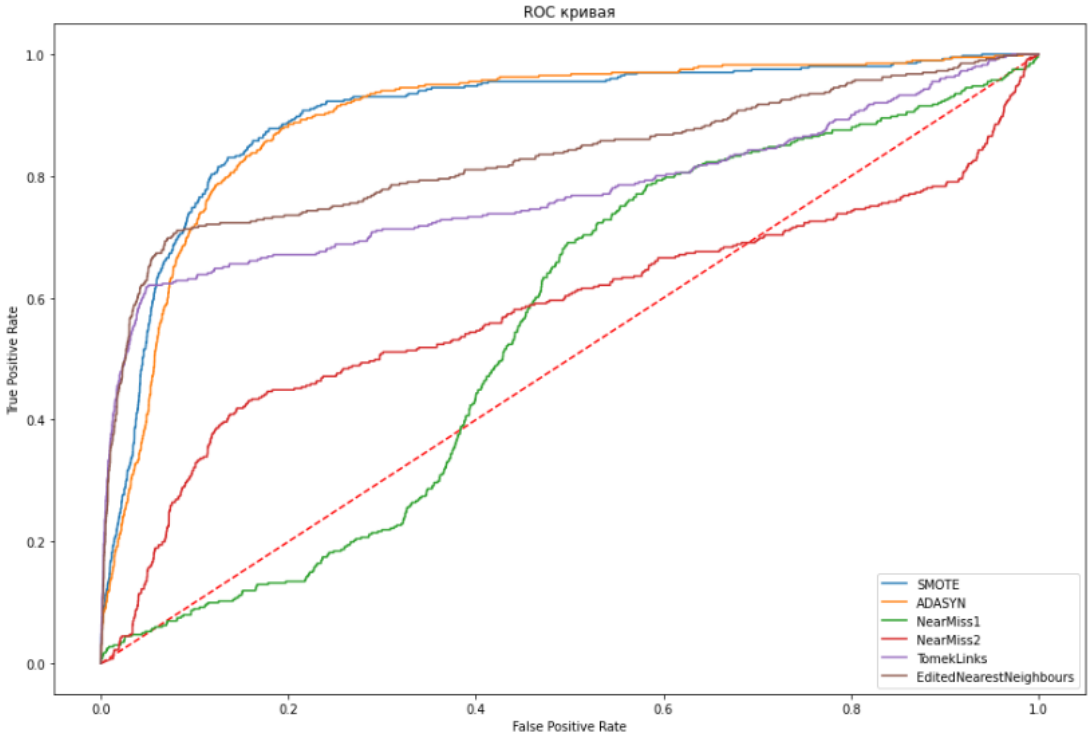

Рисунок 18 - ROC-кривая SVC при разных методах сэмплирования

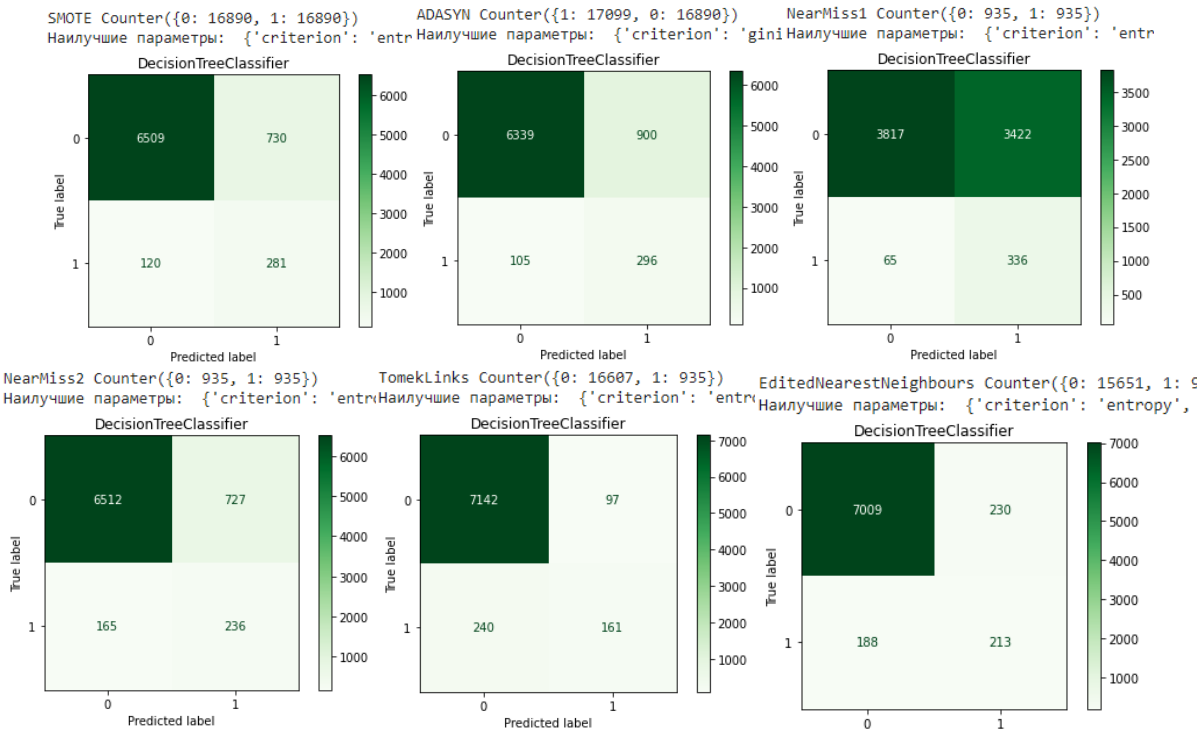

Рисунок 19 - Матрица ошибок деревьев решений при разных методах сэмплирования

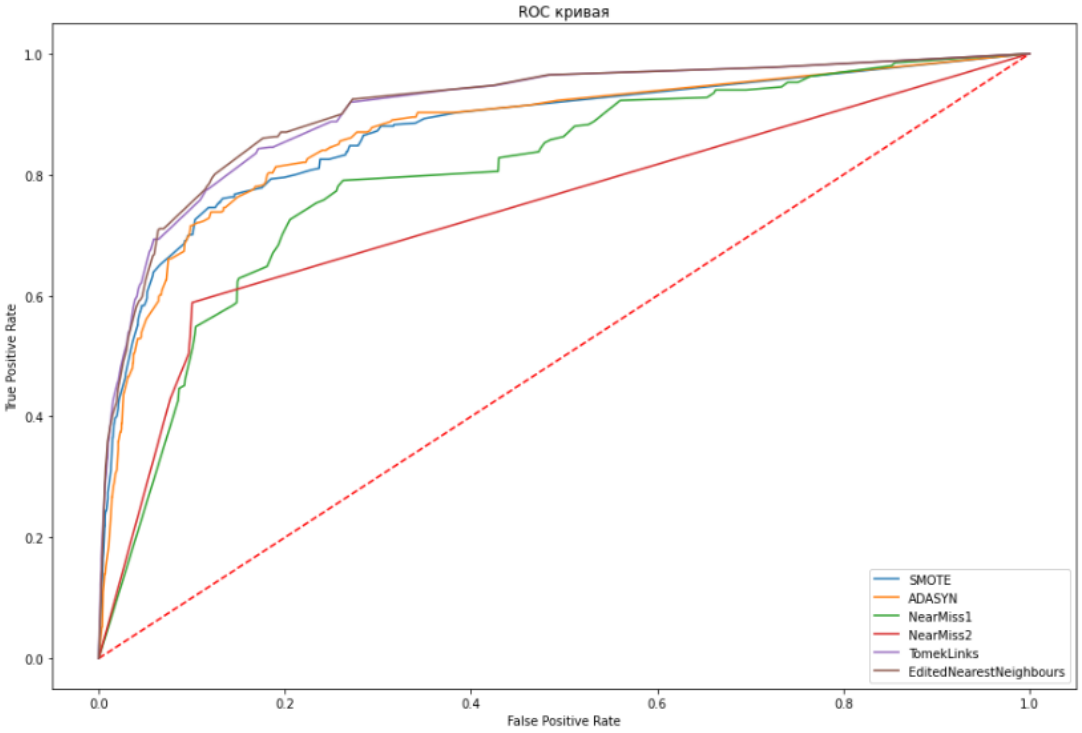

Рисунок 20 - ROC-кривые деревьев решений при разных методах сэмплирования

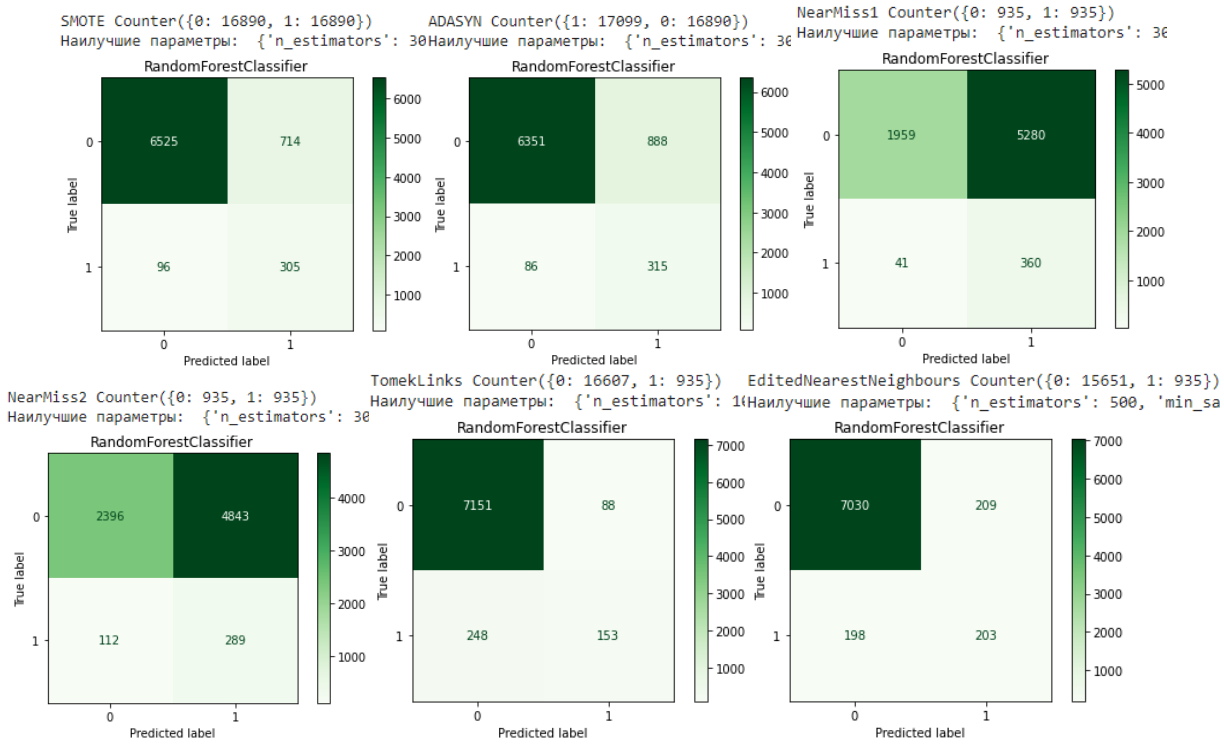

Рисунок 21 - Матрица ошибок случайного леса при разных методах сэмплирования

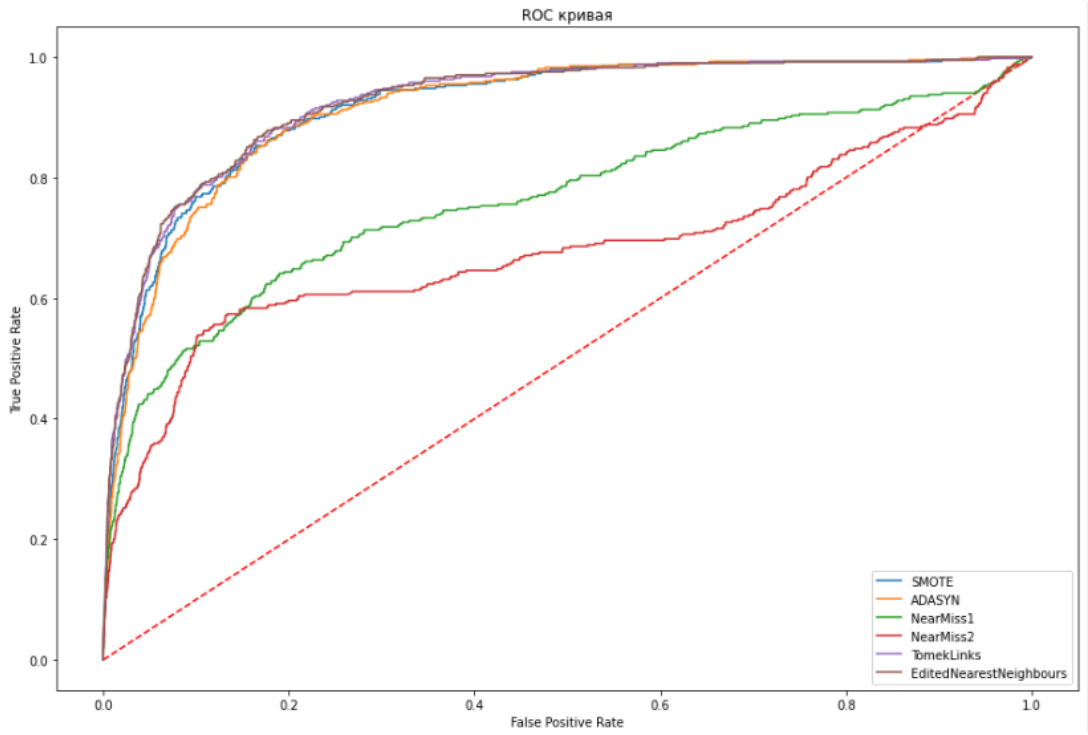

Рисунок 22 - ROC-кривые случайного леса при разных методах сэмплирования

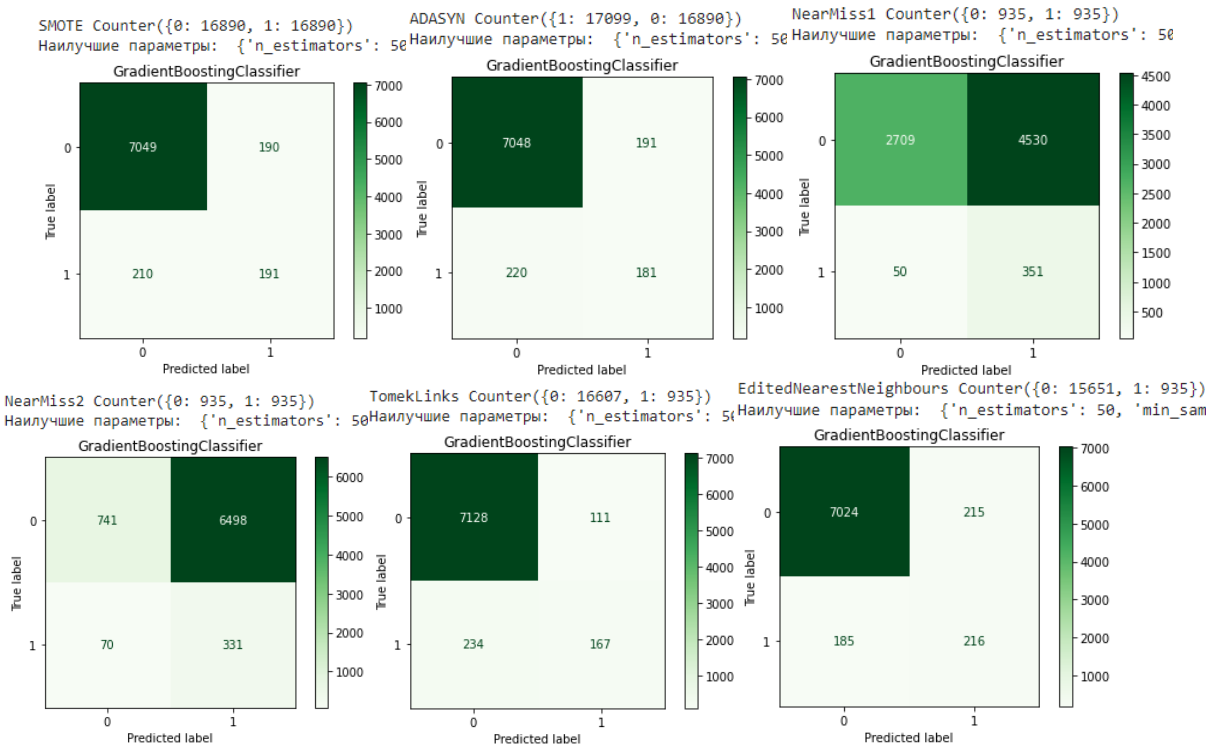

Рисунок 23 - Матрица ошибок градиентного бустинга при разных методах сэмплирования

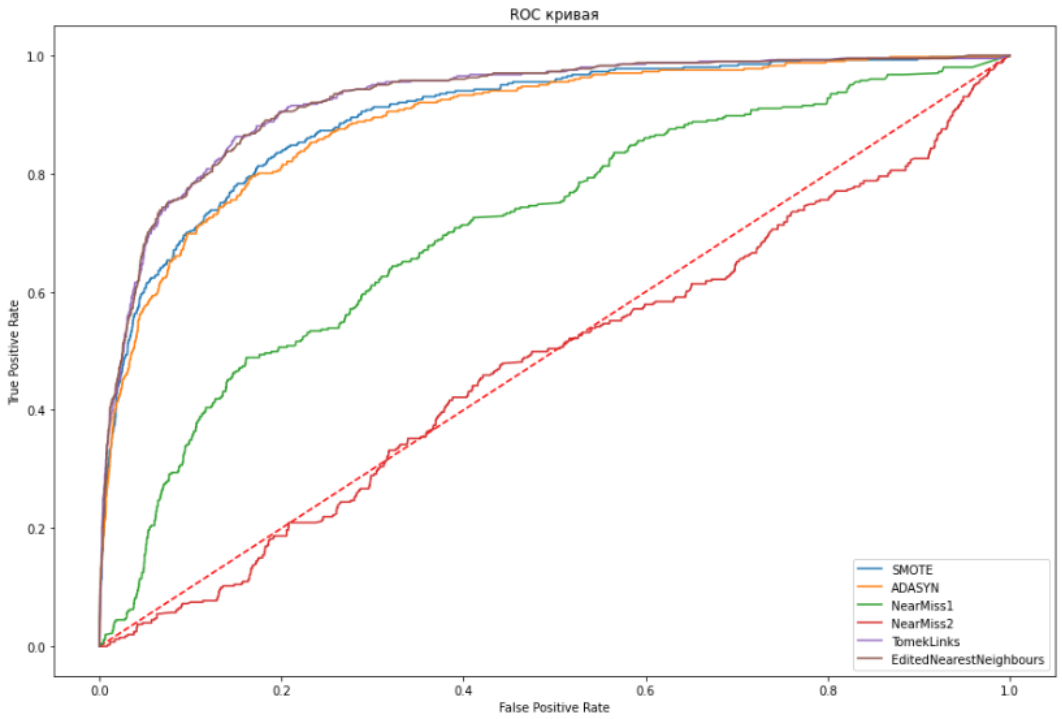

Рисунок 24 - ROC-кривые градиентного бустинга при разных методах сэмплирования

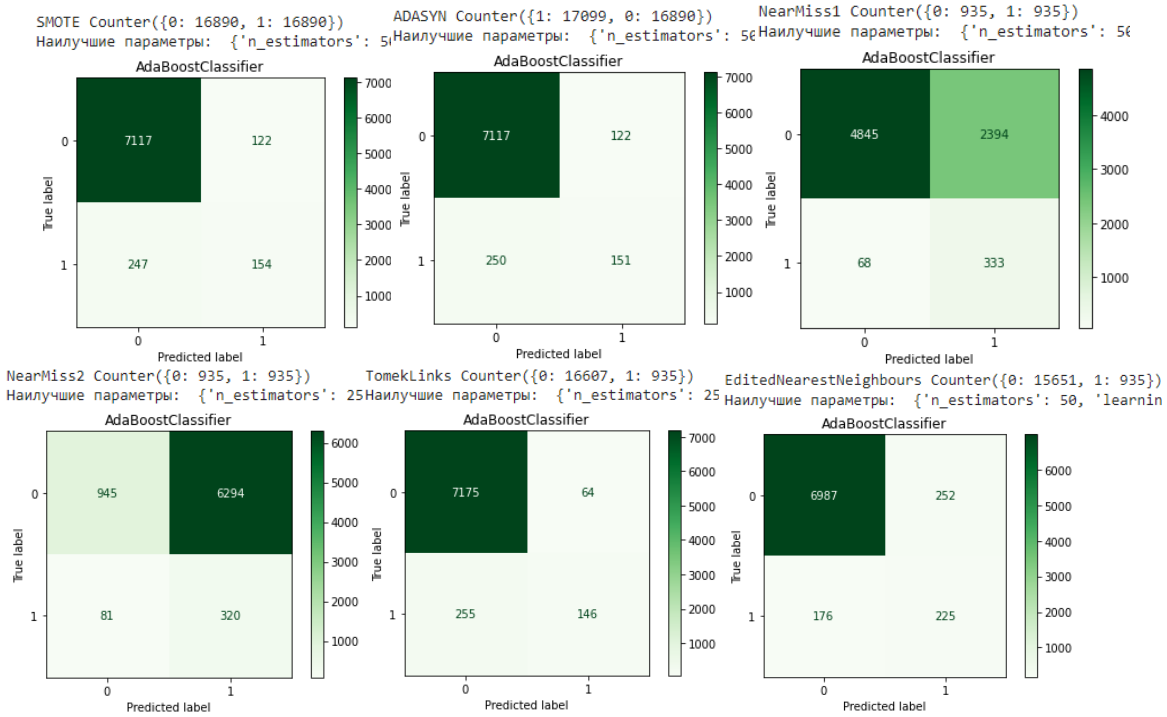

Рисунок 25 - Матрица ошибок адаптивного бустинга при разных методах сэмплирования

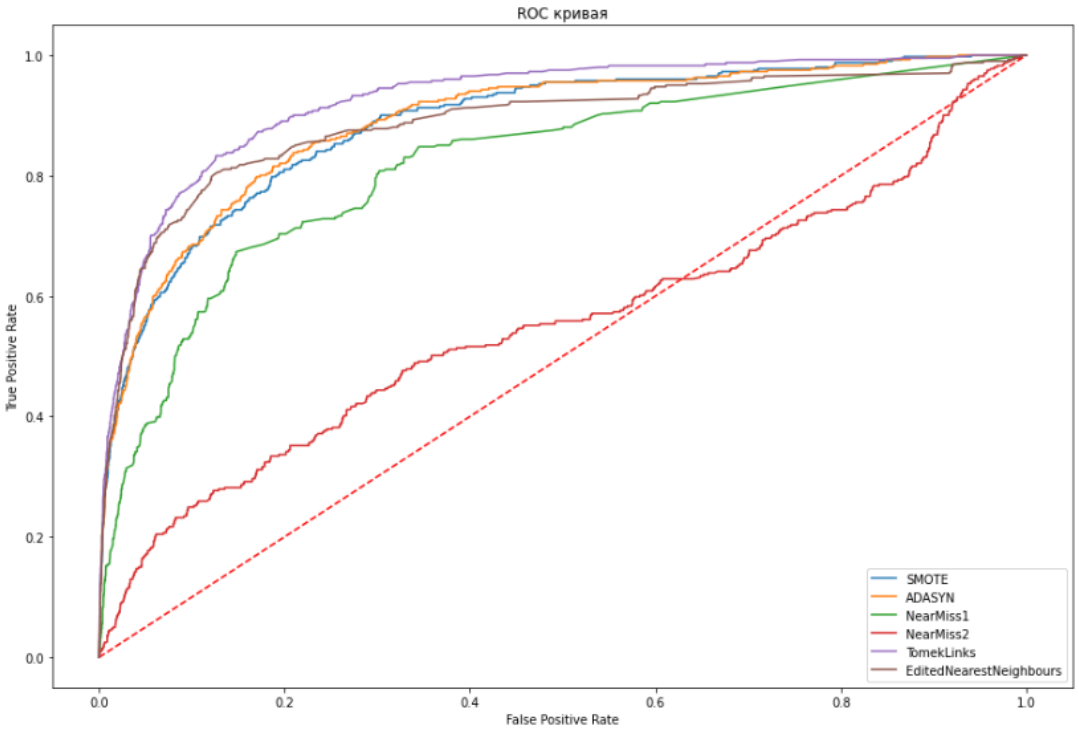

Рисунок 26 - ROC-кривые адаптивного бустинга при разных методах сэмплирования

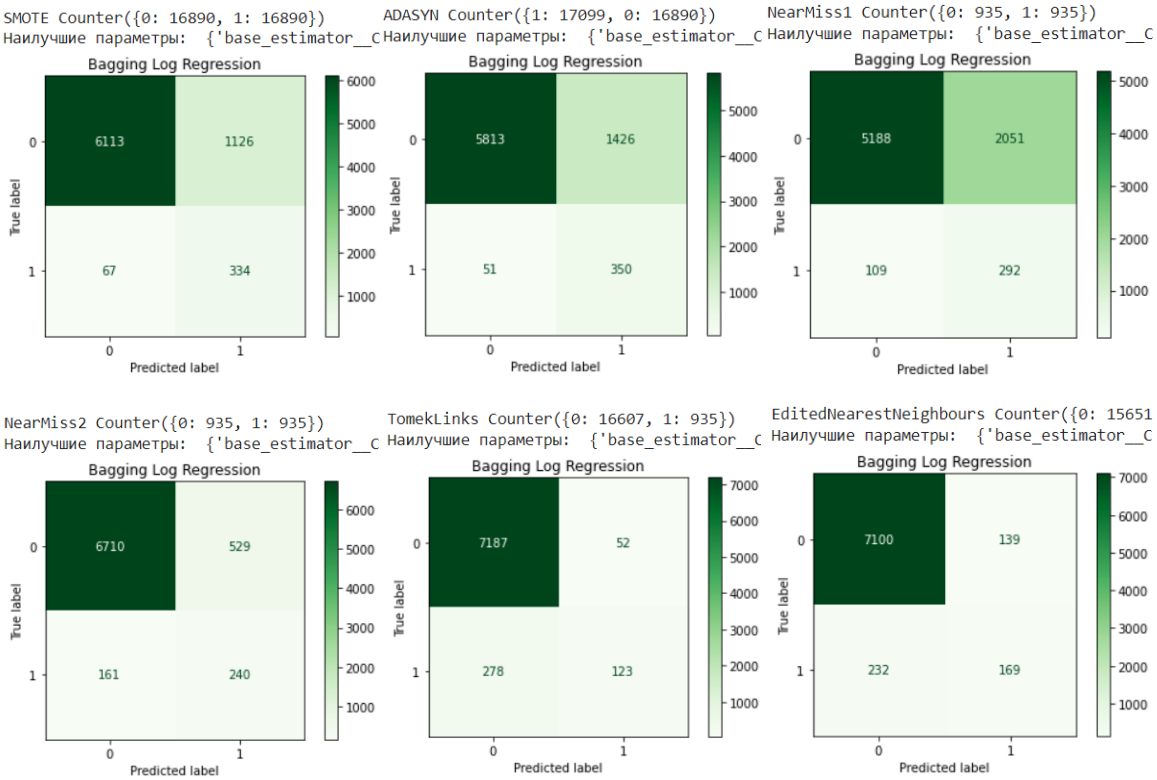

Рисунок 27 - Матрица ошибок бэггинга логистической регрессии при разных методах сэмплирования

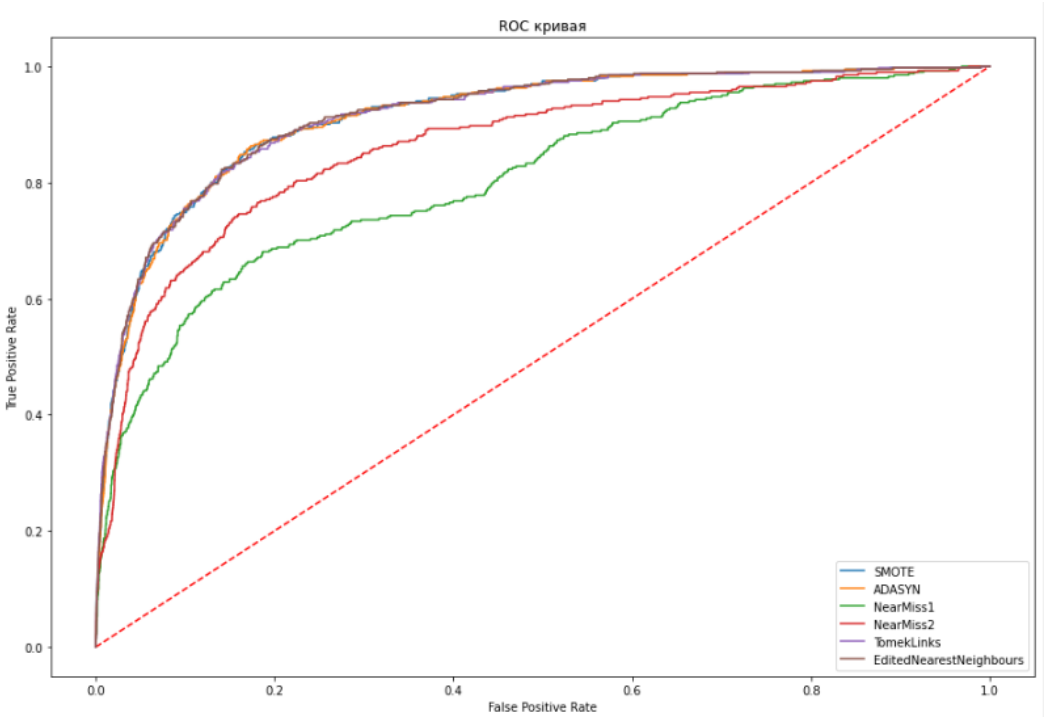

Рисунок 28 - ROC-кривые бэггинга логистической регрессии при разных методах сэмплирования

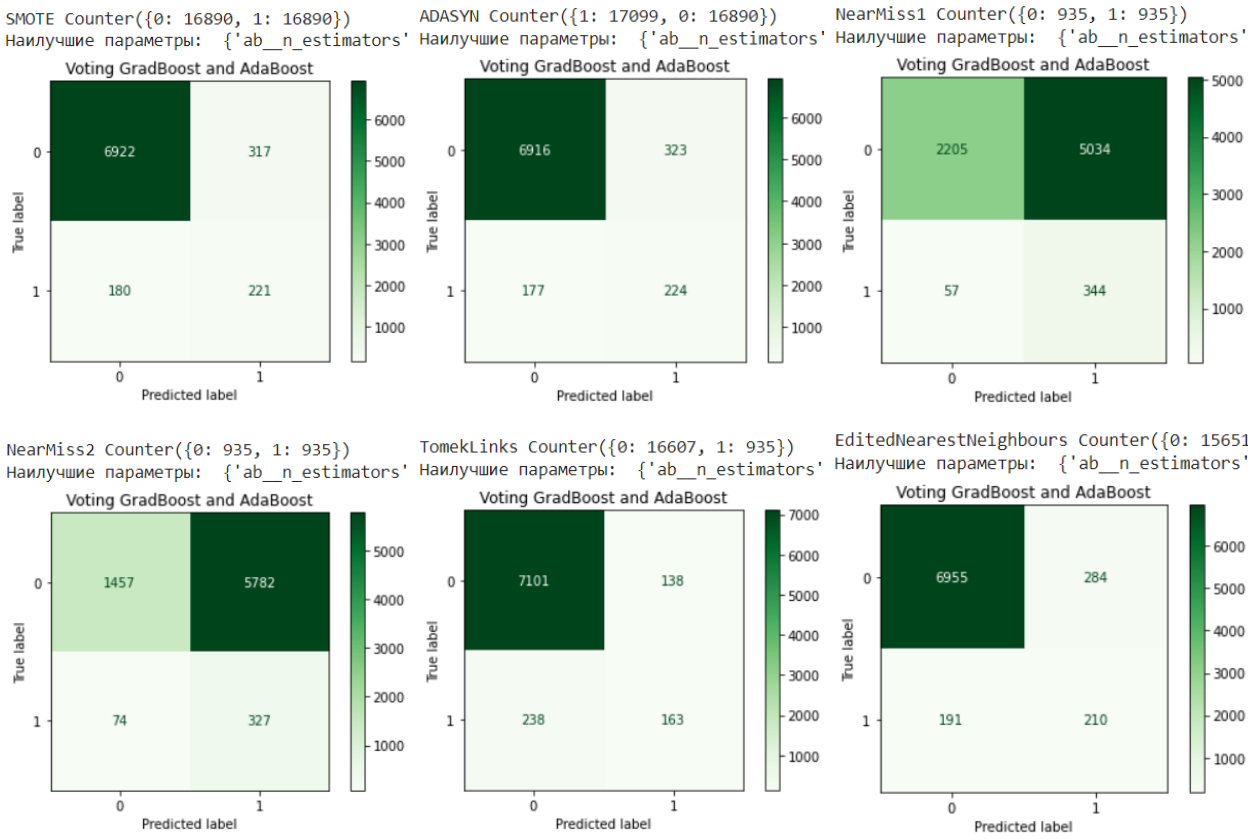

Рисунок 29 - Матрица объединения голосованием методов градиентного и адаптивного бустинга при разных методах сэмплирования

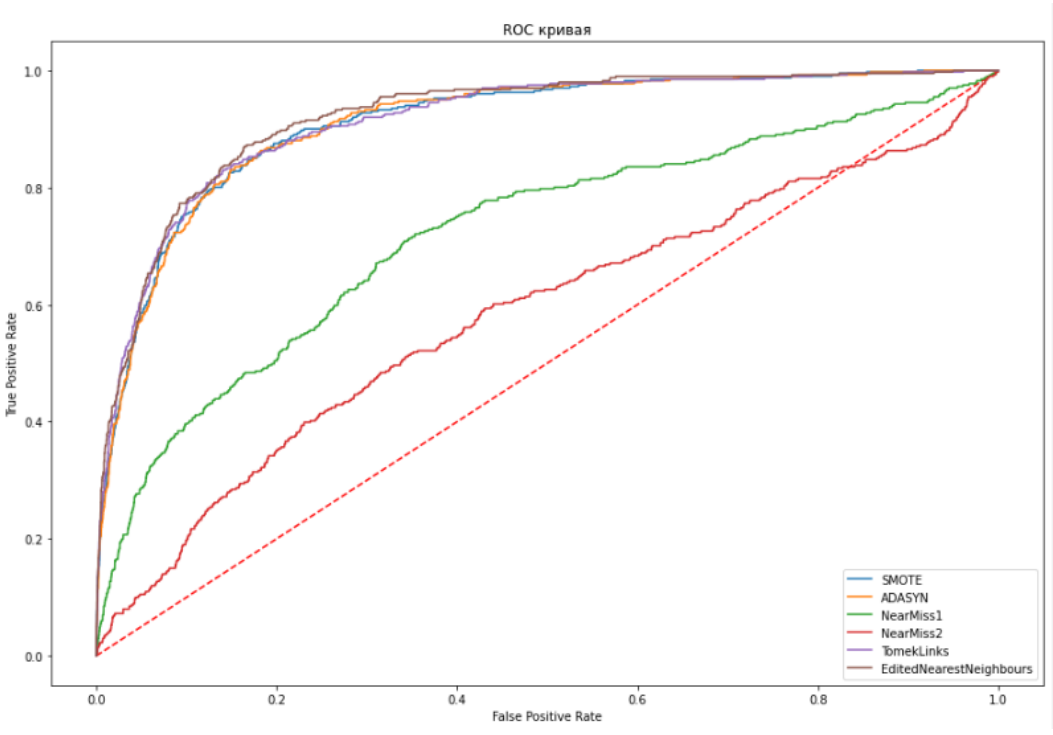

Рисунок 30 - ROC-кривые объединения голосованием методов градиентного и адаптивного бустинга при разных методах сэмплирования

Приложение Г

1) Ссылка на листинг программы с несбалансированными данными: https://colab.research.google.com/drive/1wnbh8mAmkS5HSEAfVYoEYVNpGfx_94wQ?usp=sharing
2) Ссылка на листинг программы со сбалансированными данными: https://colab.research.google.com/drive/1Q43cpwbU2hHsytkAqBDLMaC_6SmqvYzR?usp=sharing