

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»
Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

Директор ШПиАО
 Д.В. Денисов
(подпись) (Ф.И.О.)
« 03 » июня 2024 г.

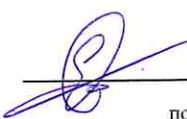
ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ВЫБОР МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВНЕДРЕНИЯ
СИСТЕМЫ ОЦЕНКИ КРЕДИТОСПОСОБНОСТИ КЛИЕНТОВ БАНКА

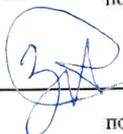
Научный руководитель: Коломыцева Анна Олеговна
к.э.н., доцент


подпись

Нормоконтролер: Огуренко Егор Владимирович


подпись

Студент группы: РИМ-220962 Зайцев Александр Васильевич


подпись

Екатеринбург
2024

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего
образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Зайцева Александра Васильевича группы РИМ-220962
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы Выбор моделей машинного обучения для внедрения системы оценки кредитоспособности клиентов банка

Утверждена распоряжением по институту от «4» декабря 2023 г. № 33.02-05/298

2. Научный руководитель

Коломыцева Анна Олеговна, доцент кафедры ИТИСУ ИРИТ-РТФ УрФУ, кандидат экономических наук

(Ф.И.О., должность, ученая степень, ученое звание)

3. Исходные данные к работе

Набор данных о заёмщиках американской компании Lending Club, выложенный на открытой площадке Kaggle

4. Перечень демонстрационных материалов Презентация, архитектура моделей машинного обучения, графики и таблицы с метриками моделей.

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	<i>1 раздел (глава)</i>	до 23.03.2024 г.	<i>выполнено</i>
2.	<i>2 раздел (глава)</i>	до 29.04.2024 г.	<i>выполнено</i>
3.	<i>3 раздел (глава)</i>	до 20.05.2024 г.	<i>выполнено</i>
4.	<i>ВКР в целом</i>	до 24.05.2024 г.	<i>выполнено</i>

Научный руководитель Коломыцева Анна Олеговна
Ф.И.О.

(подпись)

Студент задание принял к исполнению 04.12.2023
дата

(подпись)

6. Допустить Зайцева Александра Васильевича к защите выпускной квалификационной работы в экзаменационной комиссии

Директор ШПиАО

(подпись)

Д.В. Денисов

Ф.И.О.

РЕФЕРАТ

Данная дипломная работа посвящена оценке вероятности дефолта заёмщика с использованием методов машинного обучения и процессом оптимизации интерпретации предсказаний обученных моделей. Главной целью работы является исследование возможностей применения ансамблевых методов машинного обучения для оптимизации процесса и повышения точности кредитного скоринга с учетом сохранения интерпретируемости предсказания модели.

В рамках исследования были проанализированы табличные данные о заёмщиках физических лицах, предоставленные американской кредитной организацией Lending Club. Была проведена подробная предобработка данных для подготовки их к дальнейшему анализу и использованию в моделях машинного обучения.

Основной фокус работы был сосредоточен на применении трёх моделей для решения задачи бинарной классификации: логистической регрессии, случайного леса и градиентного бустинга. Логистическая регрессия была выбрана как базовое решение из-за распространенности в банковской индустрии, а случайный лес и градиентный бустинг – из-за своей способности обрабатывать большие объемы данных, выявлять сложные зависимости и более высокой точности в определенных условиях.

Для интерпретации предсказаний вышеуказанных моделей были использованы методы feature importance (для глобальных предсказаний), LIME и SHAP (для локальных предсказаний).

В ходе экспериментов алгоритм градиентного бустинга показывает более высокий коэффициент Gini на 2.5 % и более высокую оценку ROC-AUC на 1.1 % в сравнении с логистической регрессией - этот факт говорит о превосходстве данной модели по критерию качества прогноза. Модель случайного леса едва превзошла алгоритм логистической регрессии по

указанным метрикам, но всё же уступила градиентному бустингу. Вероятно, показатели метрик можно улучшить, используя другие методы для устранения дисбаланса классов и предварительной обработки данных.

Выводы и результаты, полученные в результате данной работы, не претендуют на наличие прорывных научных идей, однако они имеют прямое практическое применение в бизнесе, и именно в этом заключается главная ценность данной работы - повышение предсказательной способности модели, конвертируемое в рост ожидаемых прибылей для банков, которые станут ее использовать, а также демонстрация ее интерпретируемости, что обеспечит доверие к ней. Таким образом, несмотря на то, что модели логистической регрессии могут быть более простыми и интерпретируемыми, SHAP и LIME предлагают значительные преимущества для интерпретации более сложных и возможно более точных моделей машинного обучения в задаче предсказания дефолта заёмщика. Использование методов LIME и SHAP для ансамблевых моделей машинного обучения в перспективе может снизить кредитные риски организации и снизить её финансовые потери.

СОДЕРЖАНИЕ

РЕФЕРАТ	3
СОДЕРЖАНИЕ	5
ВВЕДЕНИЕ.....	6
Постановка проблематики работы и актуальность темы	6
Цель и задачи работы	7
1 ИЗУЧЕНИЕ И АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ	9
1.1 Особенности предметной области: определение и классификация финансовых рисков	9
1.2. Литературный обзор методов оценки кредитных рисков	12
2 ОПИСАНИЕ ПОДХОДОВ ОБРАБОТКИ ДАННЫХ И АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ФИНАНСОВОЙ ОБЛАСТИ	18
2.1. Методы предобработки и анализа табличных данных	18
2.2. Выбор и описание алгоритмов машинного обучения для решения задачи предсказания вероятности дефолта заёмщика	21
2.2.1 Логистическая регрессия	21
2.2.2. Алгоритм Random Forest Classifier	22
2.2.3. Метод XGBoost	24
2.3. Метрики моделей машинного обучения в задаче классификации	26
2.3.1. Четырехпольная таблица сопряженности.....	26
2.3.2. ROC-кривая.....	28
2.3.3 Коэффициент Джини	29
2.4. Методы интерпретации моделей машинного обучения	29
2.4.1. Метод feature importance	30
2.4.2. Метод LIME	30
2.4.3. Метод SHAP	31
3 АНАЛИЗ ИМЕЮЩИХСЯ ДАННЫХ	34
3.1. Описание датасета	34
3.2. Исследовательский анализ данных	37

3.3. Предобработка данных датасета	41
3.4. Моделирование и интерпретация предсказаний	43
3.4.1 Логистическая регрессия	43
3.4.2. XGBoost	51
3.4.3. RandomForest	58
3.5. Анализ эффективности. Перспектива использования и развития инструмента	64
ЗАКЛЮЧЕНИЕ	67
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	68

ВВЕДЕНИЕ

Постановка проблематики работы и актуальность темы

Финансовая сфера на протяжении десятилетий брала на вооружение новейшие математические методы для решения своих прикладных задач. Эта особенность во многом связана со степенью риска финансовых потерь в результате принимаемых решений отдельным специалистом. Алгоритмы кредитного скоринга стали важным инструментом для оценки кредитных рисков и принятия решений в сфере финансов. Они позволили банкам и другим финансовым учреждениям определить вероятность невозврата кредита со стороны заемщика. Однако ответственность за отдельное решение, основанное на прогнозе алгоритма скоринга, полностью всё так же лежит на конкретном сотруднике. В некоторых странах на законодательном уровне введен запрет на отказ от выдачи займа без объяснения причин (например, в США). В связи с этим, специалист, при использовании того или иного инструмента предсказания склонен выбирать именно тот метод, решение которого возможно интерпретировать, не владея специфическими знаниями в области машинного обучения и которое интуитивно можно понять.

Метод логистической регрессии из наиболее распространенных подходов в кредитном скоринге как раз из-за своей относительной простоты и вероятностной интерпретации результата. Однако метод логистической регрессии имеет ряд формальных ограничений и низкую предсказательную способность на данных с множеством сильно коррелированных признаков (эффект мультиколлинеарности). Поэтому в последние годы все большую популярность приобретают ансамблевые методы машинного обучения, которые объединяют несколько моделей для достижения лучших результатов. Преимущества ансамблевых методов заключаются в их способности обрабатывать большие объемы данных, выявлять сложные взаимосвязи между признаками и улучшать качество прогнозов относительно более

вычислительно простых моделей. Тем не менее их результат предсказания в своём базовом виде достаточно сложно проинтерпретировать. Во многом поэтому многие кредитные организации продолжают избегать использование ансамблевых моделей в задаче кредитного скоринга и во многом опираются на более консервативные подходы к анализу кредитоспособности заёмщиков.

Цель и задачи работы

Объект исследования: информационная система оценки кредитного риска и скоринга заявок на предоставление кредита клиентам банка.

Предмет исследования: применение моделей машинного обучения и методов интерпретации результатов оценки вероятности дефолта заёмщика.

Актуальность работы обусловлена необходимостью быстрого и более точного принятия решений о выдаче кредита заёмщикам при сохранении свойства доступной интерпретируемости результата в условиях возросшего объёма данных и роста объёмов кредитования.

Цель данной магистерской работы заключается в исследовании возможностей применения ансамблевых методов машинного обучения для оптимизации процесса и повышения точности кредитного скоринга с учетом сохранения интерпретируемости предсказания модели.

Задачи исследования:

1. Провести анализ предметной области, включающий в себя теоретический анализ методов оценки кредитных рисков в банковской организации;
2. Выполнить классификацию алгоритмов машинного обучения, применяющихся для решения задачи кредитного скоринга;
3. Для построения моделей машинного обучения необходимо провести подготовку и последующий анализ данных. Этот этап включает в себя заполнение пропущенных значений, обработку выбросов и шума, а также

масштабирование и преобразование данных. Для анализа данных будут использованы методы Exploratory Data Analysis;

4. Выполнить отбор моделей машинного обучения и методов интерпретации их предсказаний, применяемых в индустрии;

5. Математически формализовать предложенные алгоритмы скоринга и методы их интерпретации;

6. Описать последовательность применения методов предобработки данных о заёмщиках, обучения алгоритма машинного обучения и интерпретации предсказанных значений;

7. Провести предобработку данных о заёмщиках, представленных в табличном виде и обучить на них модель машинного обучения;

8. Провести тестирование алгоритма на тестовом наборе данных: дать оценку эффективности моделей на основе выбранной метрики. Провести тестирование методов интерпретации предсказания алгоритма на случайной выборке из используемого набора данных;

9. Сравнить численную точность оценок “классических” методов скоринга и предложенных ансамблевых методов машинного обучения. Оценить возможности интерпретации результата работы предложенных алгоритмов в сравнении с другими подходами к кредитному скорингу.

1. ИЗУЧЕНИЕ И АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Особенности предметной области: определение и классификация финансовых рисков

В данном параграфе представлено краткое описание двух на первый взгляд различных и независимых дисциплин, которые, однако, будут одинаково важны для последующего анализа. Первый раздел будет посвящен кредитному анализу и основным методам, используемым в этой области. Во втором разделе мы рассмотрим современные подходы к интерпретации сложных предсказательных моделей, одну из которых мы будем обучать в разделе по моделированию.

Наиболее емко можно охарактеризовать понятие «риск» как вероятность отклонений фактических результатов деятельности от ожидаемых в условиях неопределенности. В документах Банка России от 2004 г. и Федеральном законе «О банках и банковской деятельности» кредитные риски относятся к категории банковских рисков. Кредитные риски так же подпадают под классификацию банковских рисков в разрезе «видов деятельности». Полная классификация имеет следующий вид:

- Кредитные риски;
- Депозитные риски;
- Риски расчетов и платежей;
- Эмиссионные риски;
- Инвестиционные риски;
- Риски валютнообменных операций;
- Гарантийные риски;
- Консалтинговые риски.

Как мы видим кредитные риски лишь одна из многих групп рисков, которые берёт на себя кредитная организация, поэтому самое время дать определение этому понятию: кредитный риск — риск возникновения у

кредитной организации убытков вследствие неисполнения, несвоевременного либо неполного исполнения должником финансовых обязательств перед кредитной организацией в соответствии с условиями договора [4, С. 118]. Тем не менее, снижение кредитных рисков является наиболее важной задачей для коммерческих банков, так как кредитные операции составляют основу их деятельности.

В рамках определения уровня кредитного риска банка в первую очередь необходимо определиться с критериями кредитоспособности клиента.

Кредитоспособность клиента будь то физическое или юридическое лицо должна содержать несколько критериев. Если посмотреть на мировую практику и практику российских банков, то основными критериями кредитоспособности могут быть:

- Характер деятельности клиента банка, отрасль бизнеса, где работает заемщик — юридическое лицо;
- Возможность и способность заемщика вернуть в срок полученную ссуду;
- Возможность в срок обслуживать образовавшуюся ссудную задолженность;
- Возможность получения прибыли и возможность формирования фонда для оплаты задолженности по кредиту;
- Капитал;
- Условия осуществления кредитной сделки;
- Контроль за правильным формированием соответствующего резерва.

При оценке кредитоспособности банки, как правило, формируя мнение о клиенте-заемщике, определяют лимит кредита, который может быть выдан заемщику. В практике деятельности банков мы можем наблюдать объединение заемщиков в однотипные группы, что впоследствии предполагает формирование банком единых суждений по риск-профилю

данных клиентов. И, как итог, это приводит к формированию единых подходов для всех заемщиков, включенных в данную группу [4. С. 122–123].

Неточность в вынесении решения о принадлежности предполагаемого заёмщика к определённой группе может привести к финансовым потерям кредитной организации. В рамках минимализации кредитного риска, организации решают задачу определения ожидаемого уровня потерь при различных сценариях (expected losses). Данный показатель формулируется следующим образом:

$$EL = \sum_{i \in I} PD_i * LGD_i * EAD_i,$$

где EL – ожидаемые потери по портфелю заемщиков, I – множество кредитруемых заемщиков, PD_i – вероятность дефолта i -го заемщика, LGD_i – ожидаемая доля непогашения i -ым заемщиком при условии наступления его дефолта, EAD_i – сумма денежных средств под риском по i -му заемщику.

Расчет данных показателей помогает банкам решать классическую дилемму риска-доходности: более рискованные активы обычно приносят более высокую доходность, и наоборот. Кроме того, показатель ожидаемых потерь используется для определения банковских резервов, которые являются законодательным требованием на территории РФ [5]. При этом чем выше резервы, тем с одной стороны, банк в большей степени застрахован на случай реализации кредитного риска, с другой стороны, зарезервированные средства не выдаются заёмщикам, а значит не приносят прибыли. Это порождает ещё одну дилемму, с которой банку необходимо справляться. И наконец, заметим в формуле ожидаемых потерь компоненту вероятности дефолта заёмщика. От качества моделирования данного показателя зависит оценка ожидаемого уровня потерь. Для решения этой задачи был разработан и предложен ряд методов, которые будут детальнее описаны в следующем параграфе.

1.2. Литературный обзор методов оценки кредитных рисков

В 1958 году инженером Б.Файром и математиком Э.Айзеком была разработана первая скоринговая модель [3]. Они разработали систему кредитного скоринга FICO, который оценивается в диапазоне от 300 до 850. Рассчитывается FICO исходя из информации трех крупнейших национальных кредитных бюро: Experian, Equifax и TransUnion. Точная формула расчёта является коммерческой тайной, однако известно, какая информация и в какой мере используется при формировании кредитного балла FICO [9]:

- Кредитная история (в т. ч. история платежей по счетам) — 35%;
- Бремя задолженности (в т. ч. количество счетов с остатками, сумма задолженности по различным видам счетов, доля используемых средств от возможного кредита и др.) — 30%;
- Продолжительность кредитных историй (например, средний возраст аккаунта, возраст самого продолжительного аккаунта) — 15%;
- Типы использованных кредитов (например, возобновляемый кредит, ипотечное кредитование и др.) — 10%;
- Недавние заявки на получение кредита — 10%.

Более высокий балл, как и в большинстве других моделей, соответствует более низким рискам. Следует отметить, что определение порога, по которому отсеиваются заявки, которые не будут удовлетворены, требует дополнительных усилий. Нет строго определенной процедуры и выбор этой черты зависит от стратегии банка: какие риски готов принимать банк, насколько он стремится расширять свой кредитный портфель и т.д. Ситуация осложняется еще и тем, что в случае FICO сравнение значений скоринговых баллов во времени невозможно: если у заемщика в текущий период балл выше, чем был в предыдущий, это вовсе не значит, что кредитная операция с его участием стала менее рискованной, поскольку балл FICO зависит не только от характеристик конкретного заемщика, но и от поведенческих паттернов

больших групп населения, целой страны или отдельных регионов. Поскольку FICO не учитывает факторы, относящиеся, например, к демографическим характеристикам потенциального заемщика, решения по удовлетворению или отклонению заявки на кредит не всегда принимаются исключительно на скоринговом балле, полученном от агентств кредитной отчетности.

Значительный вклад в разработку моделирования вероятности дефолта и оценки влияния отдельных факторов внес американский экономист Э. Альтман в своей работе 1968 г. «Финансовые коэффициенты, дискриминантный анализ и прогноз корпоративного банкротства», который смог на небольшой выборке без использования сложных алгоритмов создать простую, понятную и весьма эффективную систему балльной оценки кредитоспособности заемщиков [7. С. 189–209].

В те же годы свою модель разрабатывал экономист У. Бивер. Основным отличием модели Бивера от иных подходов предсказания вероятности дефолта организации является то, что в предложенном решении основным не является интегральный показатель (если сравнивать с моделью Альтмана), а также не учитывается вес каждого отдельного коэффициента. Значения нескольких коэффициентов сравниваются с нормативными величинами, и в соответствии с ними предприятию присваивается одно из трех состояний:

- 1) Финансово устойчивое;
- 2) Вероятное банкротство в течении пяти лет;
- 3) Вероятное банкротство в течении одного года [8. С. 71–111].

С течением времени алгоритм логистической регрессии стал индустриальным стандартом в финансовой сфере из-за своей точности, простоты интерпретации и низких требований к вычислительным мощностям. Во многом все последующие исследования по адаптации определённого алгоритма машинного обучения для решения задачи кредитного скоринга на том или ином наборе данных отталкивались от сравнения с результатами предсказания алгоритма логистической регрессии. Наиболее полным

исследованием по тематике оценки прогностической способности алгоритмов машинного обучения в задаче кредитного скоринга является работа С.Лессманна, Б.Баесенса, Х.Сео, Л.Томас [16]. Исследователи сравнили 41 алгоритм машинного обучения на 8 различных наборах данных по кредитному скорингу. Авторы выбрали 6 метрик для оценки моделей и на их основе проранжировали выбранные алгоритмы по точности прогноза. В целом, показатели эффективности делятся на три типа. Те, которые оценивают дискриминационную способность классификатора (например, AUC); те, которые оценивают точность вероятностных предсказаний классификатора (например, Brier Score) и те, которые оценивают правильность категориальных предсказаний классификатора (например, percentage correctly classified (PCC)). Сами модели были разделены так же на 3 категории: индивидуальные классификаторы (например, logistic regression, k-nearest neighbor, SVM, ANN), гомогенные ансамблевые классификаторы (Random forest, Stochastic gradient boosting, decision tree) и гетерогенные ансамблевые классификаторы (Stacking, NCES with bootstrap sampling, k-Means clustering) [16. С. 7–10]. Гомогенные ансамбли отличаются от гетерогенных тем, что создают базовые модели используя один и тот же алгоритм, а гетерогенных ансамблях используются разные базовые классификаторы.

Далее исследователи сравнили лучший по оценкам алгоритм из каждой категории (ANN, RF, NCES-Bag) с логистической регрессией. На основании результатов они пришли к выводу, что логистическая регрессия предсказывает значительно менее точно, чем другие классификаторы. Показатели RF и ANN статистически значимо не отличаются друг от друга. Алгоритм случайного леса показал, что его использование может существенно снизить стоимость ошибок в оценочной карте (в среднем на 5.7%) в сравнении с логистической регрессией. Улучшения в несколько процентов являются значимыми с управленческой точки зрения, особенно если учесть большое количество кредитных операций в финансовой сфере. В заключение авторы статьи

рекомендовали использовать модель случайного леса в качестве нового эталона в задаче кредитного скоринга с которым можно будет сравнить новые модели и архитектуры [16. С. 22–25].

В указанной выше статье сравнивались в том числе алгоритмы классификации на основе искусственных нейронных сетей, однако на момент 2015 г. модели ANN ещё уступали классическим ансамблевым алгоритмам. В следующие годы вышел ряд исследований, в которых отмечается более высокая точность прогнозирования нейронных сетей по сравнению с ансамблевыми моделями в задаче кредитного скоринга [10, 21, 22]. Однако из-за невозможности интерпретации результатов предсказаний искусственных нейронных сетей, их применение в финансовой индустрии в настоящее время сильно ограничено.

Отдельно отмечу исследования, посвященные разработке алгоритмов интерпретации предсказаний моделей машинного обучения. Метод LIME (Local Interpretable Model-agnostic Explanations) был разработан группой исследователей и описан в работе 2016 г. под названием «Why Should I Trust You?»: Explaining the Predictions of Any Classifier" [18. С. 1135–1144]. Выдвинутый подход позволяет анализировать предсказания сложных моделей, с помощью их локальной аппроксимации более простыми и интерпретируемыми моделями.

Результаты LIME являются локальными, то есть они объясняют предсказание модели для конкретного примера или наблюдения, а не для всего набора данных. В случае линейной модели коэффициенты могут быть положительными или отрицательными, указывая на то, как признак влияет на предсказание (положительно или отрицательно). Величина коэффициентов указывает на степень важности признаков: большие абсолютные значения коэффициентов свидетельствуют о большем вкладе признака в предсказание [18. С. 1137–1138].

Год спустя был разработан и описан метод SHAP (SHapley Additive exPlanations) для интерпретации предсказаний в работе «A Unified Approach to Interpreting Model Predictions» [17]. SHAP — это метод объясняющий вклад каждого признака в предсказание конкретного наблюдения, в этом он схож с LIME. Однако, SHAP основывается на концепции значений Шепли — методе из кооперативной теории игр, который распределяет «выигрыш» (или влияние) между всеми участниками игры (в контексте SHAP — между признаками модели). В получение данного прироста каждый из факторов вносит определенный вклад, который мы пытаемся оценить.

Это среднее и является значением Шепли для данного признака. Полный перебор всех возможных коалиций для каждого признака является вычислительно затратным, особенно при большом количестве признаков. Поэтому при решении задач часто производится лишь оценка вектора Шэпли, а не его полный вывод.

Помимо рассмотрения влияния отдельных переменных, исследователям важно понимать общие закономерности и значимость анализируемых характеристик. Здесь стоит упомянуть подход feature importance в анализе важности переменных - его можно применить к любой предсказывающей модели.

Данный подход позволяет проанализировать одновременно все факторы, не требует переобучения модели, и опять же интерпретируем. Особенностью метода является то, что он даёт интерпретацию глобальной важности признаков (всей выборки), но не локальных предсказаний [11, С. 244]. Этот факт усложняет применение данного метода в сфере банковского скоринга. Тем не менее, в целом данный подход является весьма удобным и вполне эффективным для первичного анализа влияние факторов.

В данной главе мы рассмотрели основные аспекты и особенности предметной области, а именно определение и классификацию финансовых рисков. Мы подробно изучили сущность кредитных рисков. Это позволило

нам лучше понять, какие факторы могут влиять на финансовую устойчивость организации и какие меры можно предпринять для минимизации потенциальных рисков.

В параграфе 1.2 мы провели обзор научной литературы по методам оценки кредитных рисков и интерпретации предсказаний моделей машинного обучения. Мы рассмотрели, как классические подходы, такие как рейтинговые системы и скоринговые модели, так и более современные методы, основанные на машинном обучении. Этот обзор позволил нам увидеть широкий спектр подходов к оценке кредитных рисков и понять, какие методы могут быть наиболее эффективными в различных ситуациях.

В целом, данная глава создает фундамент для дальнейшего изучения темы оценки и управления кредитными рисками. Она помогает лучше понять контекст проблемы и оценить потенциальные подходы к ее решению.

2. ОПИСАНИЕ ПОДХОДОВ ОБРАБОТКИ ДАННЫХ И АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ФИНАНСОВОЙ ОБЛАСТИ

2.1. Методы предобработки и анализа табличных данных

Методы предобработки и анализа табличных данных оказывают существенное влияние на процесс оценки и прогнозирования кредитных рисков в рамках данной темы. Этот процесс включает в себя предварительную обработку данных, агрегацию, нормализацию, обработку пропущенных значений и отбор значимых признаков. В рамках данной работы можно применить следующие методики обработки данных [1. С. 26–31]:

– Предварительная обработка данных — на этом этапе осуществляется оценка качества данных, устранение дубликатов, обработка выбросов и аномалий. Также может быть выполнена стандартизация данных для обеспечения их сопоставимости [14. С. 40-45].

– Агрегация данных — если данные о кредитных рисках представлены в разных источниках или форматах, необходимо провести агрегацию данных для их объединения в единую таблицу. Это позволяет создать унифицированный набор данных для последующего анализа.

– Нормализация данных — нормализация данных представляет собой важный этап, особенно при применении методов машинного обучения. Она позволяет привести данные к единому масштабу и избежать проблем с несбалансированными весами признаков [14. С. 46–48].

– Обработка пропущенных значений — в реальных данных часто встречаются пропущенные значения. Для их заполнения можно применять различные подходы, включая заполнение средними значениями, интерполяцию или применение алгоритмов машинного обучения [14. С. 59–63].

– Отбор значимых признаков — важно отобрать наиболее значимые признаки, которые влияют на оценку кредитных рисков. Для этого можно

использовать различные методы, такие как анализ важности признаков (например, с использованием алгоритма случайного леса) или корреляционный анализ [14. С. 163–190].

Методика анализа данных включает в себя последовательность этапов и методов, которые обеспечивают выявление закономерностей, взаимосвязей и паттернов в данных, а также позволяют сформулировать гипотезы, выводы и прогнозы относительно кредитных рисков. В рамках данной работы, следующие методологические подходы могут быть использованы:

- Исследовательский анализ данных (Exploratory Data Analysis) включает в себя изучение и визуализацию данных с целью выявления основных характеристик и паттернов. Это позволяет получить глобальное понимание данных и их распределения, а также выявить потенциальные аномалии и выбросы.

- Статистический анализ данных включает применение статистических методов для оценки степени связи между различными переменными и их влияния на кредитные риски. Это может включать расчет коэффициента корреляции, проведение статистических тестов и построение регрессионных моделей для прогнозирования.

Методы машинного обучения представляют собой мощный инструмент для анализа и прогнозирования кредитных рисков. Они включают в себя алгоритмы классификации, регрессии и кластеризации, которые могут использоваться для создания моделей оценки кредитных рисков на основе исторических данных.

Визуализация данных представляет собой эффективный способ представления сложных данных в понятной форме. Использование различных графических инструментов позволяет наглядно представить результаты анализа и сформулировать выводы и гипотезы.

Отбор информативных признаков представляет собой критически важный этап в оценке кредитных рисков. Это включает определение

соответствующих финансовых, экономических и демографических показателей, которые могут влиять на вероятность невыполнения кредитных обязательств. При отборе признаков следует учитывать их доступность, актуальность и значимость для аналитических целей. Некоторые из распространенных методов отбора признаков включают:

– Корреляционный анализ: позволяет определить степень взаимосвязи между различными показателями и целевой переменной (кредитным риском). Признаки с высокой корреляцией с целевой переменной могут быть выбраны для включения в модель.

– Методы отбора на основе важности признаков: Эти методы, такие как случайный лес, оценивают важность каждого признака в предсказании целевой переменной. Признаки с высокой важностью могут быть выбраны для включения в модель. При отборе показателей важно учитывать допустимость и применимость этих показателей для анализа кредитных рисков, а также обеспечить их достоверность и актуальность.

Генерация признаков представляет собой существенный этап в анализе и моделировании кредитных рисков. Это позволяет формировать новые переменные, которые могут включать дополнительную информацию и повышать эффективность и точность предсказаний моделей в оценке кредитных рисков. В контексте оценки кредитных рисков с использованием методов машинного обучения, существуют различные подходы к созданию признаков, включая анализ предметной области и применение генетического программирования — перед началом создания новых признаков полезно провести анализ предметной области кредитных рисков. Это позволяет понять основные факторы, влияющие на кредитные риски, и выявить потенциально значимые переменные. Например, такие как доход заемщика, семейное положение, история кредитных платежей и другие, которые могут быть использованы для создания новых признаков.

Генетическое программирование представляет собой эволюционный алгоритм, используемый для автоматического создания программ или моделей, способных решать задачи. В контексте создания признаков для оценки кредитных рисков, генетическое программирование может быть применено для формирования новых математических выражений или функций, основываясь на существующих переменных.

2.2. Выбор и описание алгоритмов машинного обучения для решения задачи предсказания вероятности дефолта заёмщика

2.2.1 Логистическая регрессия

В качестве эталона и общепризнанного стандарта в банковской сфере был использован метод логистической регрессии для задачи классификации заёмщиков. В данной работе рассматривается задача о построении логистической регрессии с бинарной зависимой переменной. Логит-преобразование позволяет получать значение величины, принимающей значение от нуля до единицы, и интерпретируется как функция распределения некоторой случайной величины.

Значение прогнозируемого отклика, близкое к 0, говорит о малой вероятности наступления интересующего нас исхода, а значение, близкое к 1, говорит в пользу высокой вероятности наступления этого события.

В общем виде алгоритм выглядит так:

1) Мы имеем величины x_1, \dots, x_m являющиеся предикторами (характеристиками) заёмщиков;

2) Имеется множественная регрессия

$$z = \beta_0 + \beta_1 x_1 + \beta_m x_m + \varepsilon ;$$

3) Применяется регрессионная логит-функция вида:

$$F(z) = \frac{1}{1 + e^{-z}},$$

предсказывающая непрерывную переменную со значениями на отрезке от нуля до единицы.

Следовательно, в основе логистической регрессионной модели прогнозируется не само значение переменной отклика, а вероятность того, что переменная отклика принимает значение 1.

Так же для решения задачи планируется использовать ансамблевые методы машинного обучения, в частности алгоритм Random Forest Classifier и XGBoost. Выбор этих алгоритмов обусловлен точностью их прогнозов в конкретной задаче классификации (описано в главе 1) и распространенностью в рабочей практики в финансовой индустрии.

2.2.2. Алгоритм Random Forest Classifier

Опишу общий алгоритм обучения Random Forest Classifier. На входе: обучающая выборка из N образцов, размерность пространства признаков равна M , параметр m - количество признаков для обучения.

Наиболее распространённый способ построения деревьев ансамбля — бэггинг (*bootstrap aggregation*) — производится следующим образом [15. С. 587–601]:

- 1) Генерируем случайную выборку повторную подвыборку размером N из обучающей выборки;
- 2) Строим решающее дерево, классифицирующее образцы данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать набор признаков, на основе которых производится разбиение (не из всех M признаков, а лишь из m случайно выбранных), используя критерий Джини или критерий прироста информации;
- 3) Дерево строится до полного исчерпания подвыборки;

- 4) Повторение шагов а-в для каждого дерева в ансамбле;
- 5) Классификация объектов проводится путём голосования — каждое дерево комитета относит классифицируемый объект к одному из классов, а побеждает класс, за который проголосовало наибольшее число деревьев.

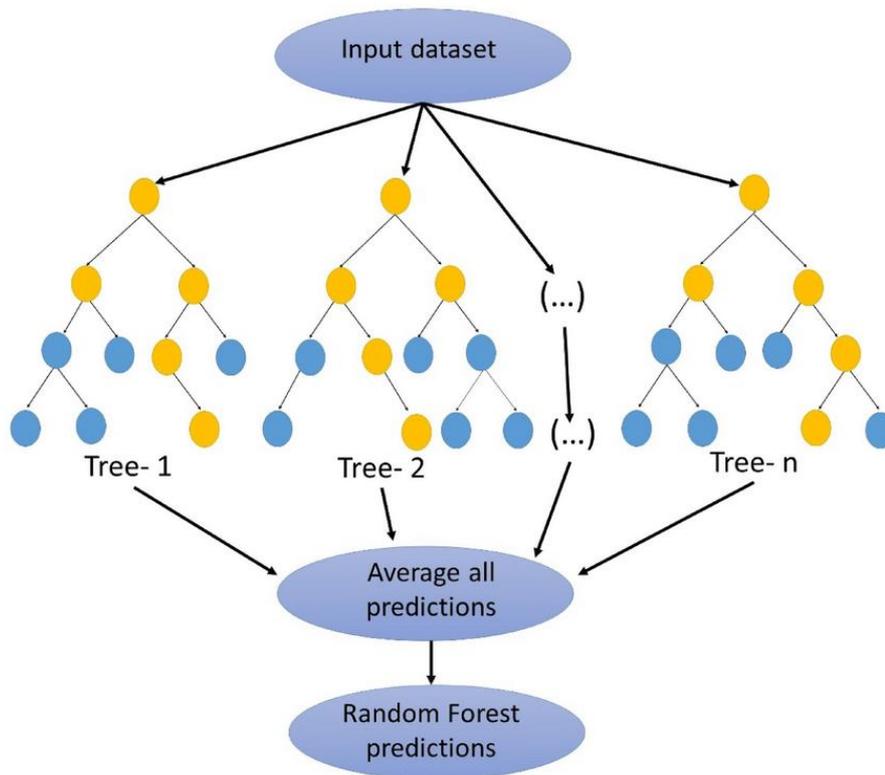


Рисунок 1 – Иллюстрированная схема алгоритма Random Forest

2.2.3. Метод XGBoost

XGBoost (eXtreme Gradient Boosting) — это метод, основанный на алгоритме градиентного бустинга, который использует деревья решений как базовые модели. Общая схема представлена на рисунке 2. Общий упрощенный алгоритм обучения XGBoost предполагает следующие шаги [13. С. 1189–1232]:

- 1) Инициализация: начинаем с константного прогноза, который минимизирует функцию потерь.

$$F_0(x) = \operatorname{argmin}_f \sum_{i=1}^n L(y_i, f),$$

где L - функция потерь, y_i - истинные значения, f - прогнозные значения.

2) Для каждой итерации (до заданного числа деревьев или до тех пор, пока улучшения не станут незначительными):

3) Вычисляем градиенты функции потерь по отношению к прогнозным значениям текущей модели;

4) Построение дерева решений, которое приближает эти градиенты. Это делается путем выбора точек разделения, которые максимизируют снижение функции потерь;

5) Обновляем модель, добавляя взвешенное дерево решений. Вес выбирается так, чтобы минимизировать функцию потерь.

6) Итоговая модель представляет собой сумму всех деревьев решений:

$$F(x) = \sum \eta_t * \phi_t(x)$$

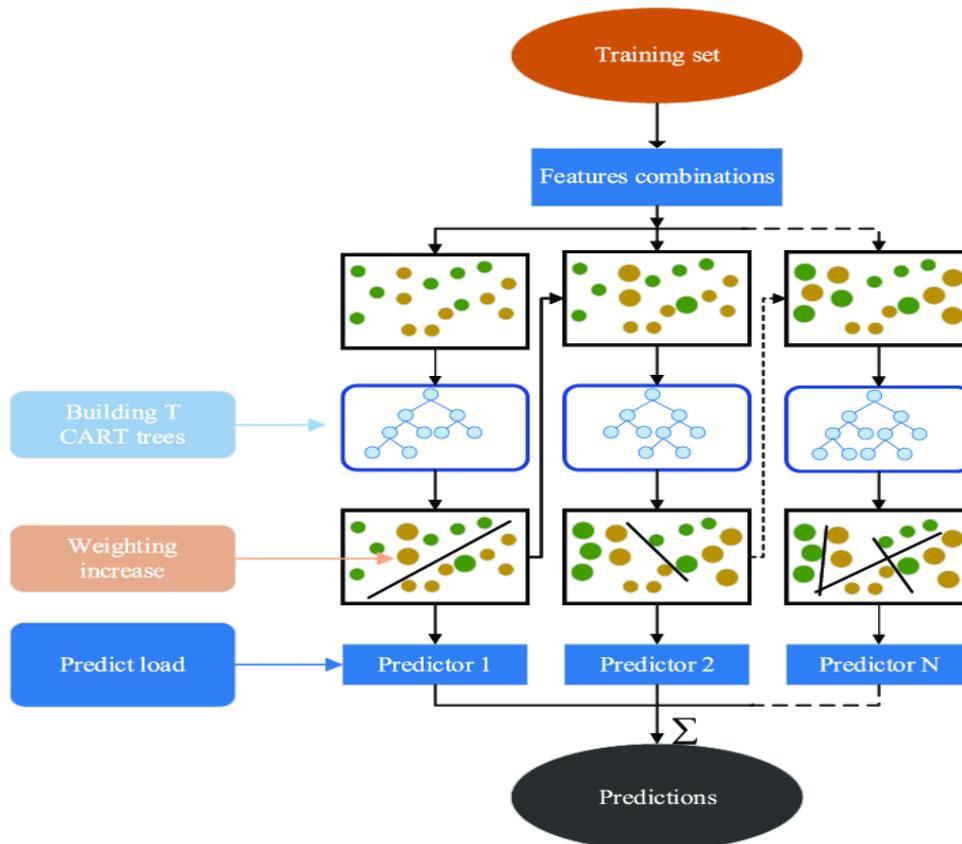


Рисунок 2 – Иллюстрированная схема алгоритма XGBoost

2.3. Метрики моделей машинного обучения в задаче классификации

2.3.1. Четырехпольная таблица сопряженности

ROC-кривая показывает зависимость количества корректно классифицированных положительных исходов от количества неверно классифицированных отрицательных исходов. В терминологии ROC-анализа первые называются *истинно положительным множеством*, вторые — *ложно отрицательным множеством*. При этом предполагается, что у классификатора (исследуемой независимой переменной) имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение значений вероятностей на два вышеупомянутых класса. Этот параметр называют порогом, или точкой отсечения (*cut-off value*). В зависимости от него будут получаться различные величины ошибок I и II рода. Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности (*confusion matrix*), называемой матрицей классов или четырехпольной таблицей (таблица 1), которая строится на основе результатов классификации с использованием регрессионной модели и принадлежностью наблюдений к двум вышеописанным классам.

Таблица 1 – Структура таблицы сопряженности

Фактическое состояние объектов	Принятое решение	
	Положительное	Отрицательное
Положительное	True Positives	False Positives
Отрицательное	False Negatives	True Negatives

В бинарной классификации каждое предсказание может иметь четыре исхода:

– True Positives — верно классифицированные положительные примеры (истинно положительных случаи),

– True Negatives — верно классифицированные отрицательные примеры (истинно отрицательные случаи),

– False Positives — отрицательные примеры, классифицированные как положительные (ошибка II рода). Это ложное обнаружение, так как при отсутствии события ошибочно выносится решение о его присутствии (ложноположительных случаи),

– False Negatives — положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый “ложный пропуск” — когда интересующее нас событие ошибочно не обнаруживается (ложноотрицательные примеры).

С помощью матрицы классов можно оценить точность модели, которая вычисляется по формуле:

$$\text{Overall success} = \frac{TP + TN}{TP + TN + FP + FN}.$$

ROC-анализ представляет собой графическую методику оценивания эффективности модели с помощью двух показателей – специфичности и чувствительности.

Чувствительность (sensitivity) является отношением числа истинно положительных наблюдений к числу фактически положительных наблюдений:

$$\text{Sensitivity} = \frac{TP}{FN + TP}.$$

Специфичность (specificity) определяется как отношение числа истинно отрицательных наблюдений к числу фактически отрицательных наблюдений:

$$Specificity = \frac{TN}{FP+TN}.$$

В логистической регрессионной модели прогноз положительного события дается при его вероятности ≥ 0.5 и отрицательного – при его вероятности < 0.5 . Таким образом, порог по разделению двух классов (cutoff) равен 0.5.

2.3.2. ROC-кривая

ROC-анализ (Receiver Operator Characteristic) — это особый метод статистического анализа, пригодный осуществления бинарной классификации, т.е. выявляющих наличие некоего состояния (условно обозначаемое «1») либо его отсутствие (условно обозначаемое «0»). Соответственно, два варианта выявляемого состояния должны быть *взаимоисключающими* — либо клиент получает кредит, либо нет. Таким образом, ROC-анализ оперирует двумя классами событий — с *положительными исходами* и с *отрицательными исходами*.

Построим ROC-кривую, которая отражает качество модели. Диаграмма строится в координатах (1-Specificity) и Sensitivity. ROC-кривая является параметрической кривой. Специфичность и чувствительность — это независимые характеристики и зависят от порога отсечения. Порог отсечения, как и вероятность, варьируется от 0 до 1 с шагом (например. 0.01). Меняя порог отсечения, получаем набор точек. Оптимальный параметр должен максимизировать чувствительность и специфичность. Следовательно,

оптимальный параметр соответствует точке на графике, которая максимально приближена к верхнему левому углу пространства построения.

2.3.3 Коэффициент Джини

Популярный индекс неравенства Джини был впервые предложен в 1912 году, но стал известен после публикации 1914 года, указывающей на связь с кривой Лоренца. Индекс Джини может быть описан несколькими математическими представлениями - каждое из них может иметь свою интерпретацию и, естественно, приводит к различным формулам оценки. Среди этих формул наиболее популярным является геометрический подход, основанный на функции Лоренца, где индекс Джини определяется как удвоенная площадь между этой функцией и диагональю, называемой линией равных долей. Однако в литературе посвященной оценке моделей машинного обучения, в том числе при решении задачи кредитного скоринга принято использовать следующую формулу [20. С. 171–178]:

$$Gini = 2 * AUC - 1.$$

Именно её мы и будем использовать для оценки моделей.

2.4. Методы интерпретации моделей машинного обучения

В результате обучения вышеописанных моделей машинного обучения и получения предсказаний целевого признака появляется задача интерпретации результата. Говоря об интерпретации моделей случайного леса и бустинга стоит начать с анализа весов факторов, получаемых с помощью подхода *feature importance* — того насколько каждый фактор относительно других уменьшает функцию потерь, или другими словами улучшает разделение между исследуемыми классами. Однако, во введении была обоснована

важность интерпретируемости конкретных предсказаний модели в задаче кредитного скоринга заёмщиков и в качестве решения были предложены методы LIME и SHAP - они позволят оценить влияние признаков на итоговое предсказание на нескольких примерах.

2.4.1. Метод feature importance

Тем не менее в работе будет использован и классический метод feature importance. Его алгоритм выглядит следующим образом [12]:

- 1) Оценивается первоначальная ошибка обученной модели;
- 2) Для каждого фактора:
 - а) Генерируется матрица с помощью перестановки значений факторов;
 - б) Оценивается новая ошибка;
 - в) Сравниваются ошибки до и после перестановки (вычисляется коэффициент Fl);
- 3) Факторы сортируются по убыванию Fl .

2.4.2. Метод LIME

Принципы метода LIME отличаются, их можно описать следующим образом:

- 1) Для каждого предсказания N раз генерируются наблюдения с помощью перестановок (permute);
- 2) Для всех наблюдений из предыдущего пункта делаются новые предсказания;
- 3) Рассчитываются расстояния от новых точек до первоначальных;
- 4) Расстояния трансформируются в меру близости;
- 5) Выбирается M признаков, которые лучше всего описывают результаты модели для генерированных наблюдений [17. С. 1137–1141].

На этих данных обучается простая модель, чьи веса затем используются для объяснения влияния факторов на сложную модель.

2.4.3. Метод SHAP

Во многом схожий с LIME метод SHAP имеет такой математический алгоритм [16. С. 4–6]:

На входе: число итераций M , наблюдение x_i , матрица признаков X , обученная модель f .

1) Для всех $m = 1, \dots, M$:

- а) Случайным образом выбирается наблюдение z из данных X ;
- б) Случайным образом o переставляем значения факторов;
- в) Упорядочиваем наблюдения z и x :

$$x_o = (x_1, \dots, x_j, \dots, x_p); z_o = (z_1, \dots, z_j, \dots, z_p)$$

г) Создаем два новых наблюдения:

$$x_{+j} = (x_1, \dots, x_{j-1}, x_j, z_{j+1}, \dots, z_p),$$

$$x_{-j} = (x_1, \dots, x_{j-1}, z_j, z_{j+1}, \dots, z_p)$$

д) Рассчитываем вклад j -го фактора для o -го наблюдения:

$$\phi_j^m = f(x_{+j}) - f(x_{-j})$$

2) Усредняем результат по всем итерациям и получаем вклад признака:

$$\phi_j(x) = \frac{1}{M} \sum_{m=1}^M \phi_j^m$$

Это среднее и является значением Шепли для данного признака. Полный перебор всех возможных коалиций для каждого признака является вычислительно затратным, особенно при большом количестве признаков. Поэтому при решении задач часто производится лишь оценка вектора Шэпли, а не его полный вывод. Пример визуализации значений вектора Шэпли примененные к признакам данных приведен на рисунке 3.

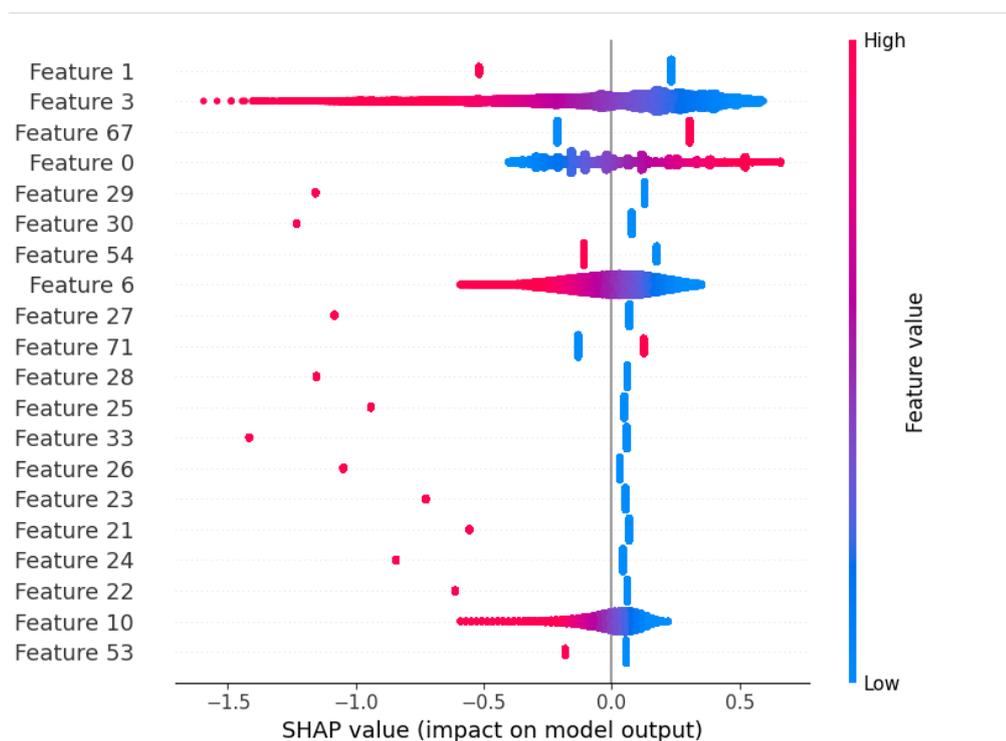


Рисунок 3 – Значения вектора Шэпли примененные к признакам данных

В данной главе мы подробно рассмотрели различные методы обработки данных, алгоритмы машинного обучения и метрики точности, которые используются в финансовой области для предсказания вероятности дефолта заемщика.

В параграфе 2.1 мы обсудили различные методы предобработки и анализа табличных данных, которые являются критически важными для подготовки данных к последующему анализу.

В параграфе 2.2 мы проанализировали различные алгоритмы машинного обучения, включая логистическую регрессию, Random Forest Classifier и XGBoost. Каждый из этих алгоритмов имеет свои преимущества и недостатки, и выбор конкретного алгоритма зависит от специфики задачи и данных.

В параграфе 2.3 мы рассмотрели различные метрики точности, включая confusion matrix, ROC-кривую и коэффициент Джини. Эти метрики помогают оценить эффективность моделей машинного обучения и сравнивать их между собой.

В параграфе 2.4 мы обсудили различные методы интерпретации предсказаний моделей машинного обучения, включая методы feature importance, LIME и SHAP. Эти методы позволяют понять, какие факторы влияют на локальные и глобальные предсказания модели.

В целом, данная глава представляет собой обзор методологии подходов и методов, которые используются в финансовой области для обработки данных и создания моделей машинного обучения для задачи оценки вероятности дефолта заёмщика. Эти методы будут использованы в дальнейшем исследовании в третьей главе.

3 АНАЛИЗ ИМЕЮЩИХСЯ ДАННЫХ

3.1. Описание датасета

В этой работе для решения задачи кредитного скоринга я буду использовать известный и общедоступный датасет «All Lending Club loan data» в котором собраны данные выданных займах физическим лицам американской финансовой компанией Lending Club за период с 2007 по 2018 гг. [6]

LendingClub использует бизнес-модель «Peer-to-peer». Заемщики могут оформить необеспеченные персональные кредиты на сумму от 1 до 40 тыс. долларов. Стандартный срок кредитования составляет от трёх до пяти лет. Инвесторы могли искать и просматривать объявления о займах на сайте Lending Club и выбирать займы, в которые они хотели вложить деньги, основываясь на информации о заемщике, сумме займа, классе займа и цели займа. Инвесторы зарабатывали на процентах по этим займам. LendingClub зарабатывал деньги, взимая с заемщиков комиссию за выдачу кредита, а с инвесторов - комиссию за обслуживание.

Приведенные данные содержат информацию о заявителях на получение кредита и о том, допустили они «дефолт» или нет. Цель - выявить закономерности, указывающие на вероятность дефолта, что может быть использовано для принятия таких мер, как отказ в выдаче кредита, уменьшение суммы кредита, предоставление кредита (рискованным заявителям) под более высокую процентную ставку и т. д.

Когда человек подает заявку на получение кредита, компания может принять два типа решений:

Кредит принят (Loan accepted) — если компания одобряет заем, возможны 3 сценария, описанные ниже:

1) Полностью выплачен (Fully paid): Заявитель полностью выплатил кредит (основную сумму и процентную ставку).

2) Действующий (Current): Заявитель находится в процессе выплаты платежей, т. е. срок займа еще не закончился. Такие кандидаты не обозначаются как "дефолтные".

3) Списанный (Charged-off): заявитель не выплачивает взносы в срок в течение длительного периода времени, т. е. допустил дефолт по кредиту.

Кредит отклонен (Loan rejected): Компания отказала в займе (потому что кандидат не соответствует их требованиям и т. д.). Поскольку заем был отклонен, история транзакций этих заявителей в компании отсутствует, поэтому эти данные недоступны в компании (и, следовательно, в этом наборе данных).

Далее следует описать переменные датасета - их всего 151, но я выделю и опишу основные. Именно эти признаки выделяют предшествующие исследователи данного датасета [19].

Список признаков:

– Loan_amnt: Указанная сумма кредита, на которую подал заявку заемщик;

– Term: Срок или количество платежей по кредиту. Значение указывается в месяцах и может составлять 36 или 60;

– Int_rate: Процентная ставка по кредиту;

– Installment: Ежемесячный платеж, который должен внести заемщик в случае выдачи кредита;

– Grade: Присвоенный кредитной организацией класс кредита;

– Sub_grade: Присвоенный кредитной организацией подкласс кредита;

– Emp_title: Название должности, указанное заемщиком при подаче заявки на кредит;

– Emp_length: Продолжительность трудовой деятельности в годах. Возможные значения - от 0 до 10, где 0 означает менее одного года стажа, а 10 — десять или более лет;

- Home_ownership: Статус владения жильем, предоставленный заемщиком при регистрации или полученный из кредитного отчета. Имеющиеся значения статуса: аренда, собственное жилье, ипотека, другое, жилье отсутствует;
- Annual_inc: Заявленный годовой доход, предоставленный заемщиком при регистрации;
- Verification_status: Признак указывает, был ли доход проверен кредитной организацией или был ли проверен источник дохода;
- Issue_d: Месяц, в котором был выдан заем;
- Loan_status: Текущий статус займа;
- Purpose: Цель или категория, указанная заемщиком для запроса займа;
- Title: Обозначение кредита, указанное заемщиком;
- Dti: Первые три цифры почтового индекса, указанного заемщиком в заявке на кредит;
- Addr_state: Штат, указанный заемщиком в заявке на кредит;
- Earliest_cr_line: Месяц, когда была открыта самая ранняя кредитная линия заемщика;
- Open_acc: Количество открытых кредитных линий в кредитном досье заемщика;
- Pub_rec: Количество “негативных” публичных записей о заемщике;
- Revol_bal: Общий остаток по возобновляемым кредитам;
- Revol_util: Коэффициент использования возобновляемой линии, или сумма кредита, которую заемщик использует по отношению ко всем доступным возобновляемым кредитам;
- Total_acc: Общее количество кредитных линий, имеющих в кредитном досье заемщика;
- Initial_list_status: Статус первоначального листинга кредита;
- Application_type: Признак указывает, является ли кредит индивидуальной заявкой или совместной заявкой с двумя созаемщиками;

- Mort_acc: Количество ипотечных счетов;
- Pub_rec: Количество банкротств в публичных записях;
- Address: Адрес заёмщика, указанный в анкете.

3.2. Исследовательский анализ данных

Для проведения анализа кредитных рисков на основе датасета мы можем использовать данные об истории кредитных заявок и платежах, а также другую информацию о клиентах, предоставленную компанией Lending Club.

В начале исследования был выполнен анализ датасета для выявления пропущенных значений. Была проанализирована структура датасета и определены переменные с пропущенными значениями.

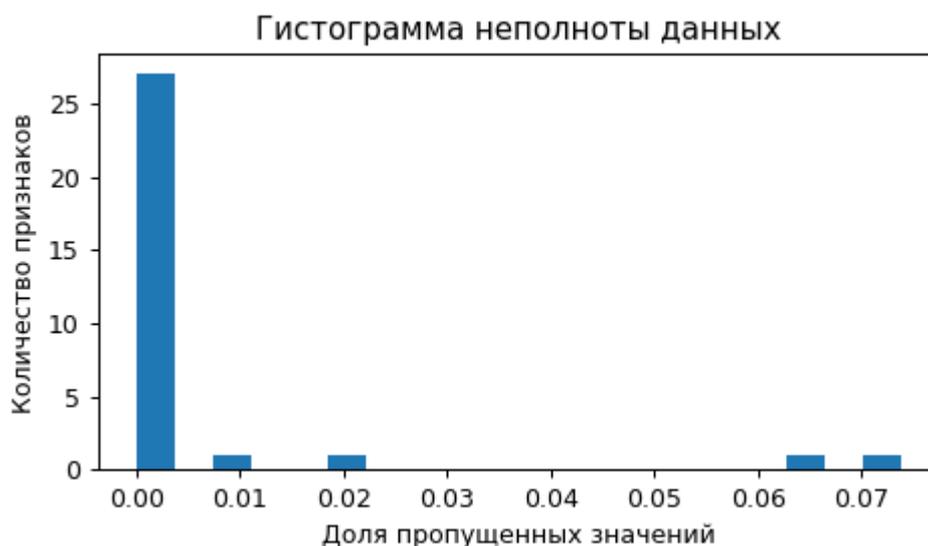


Рисунок 4 – Доля пропущенных значений среди признаков

Как видно из гистограммы большинство признаков не имеют пропущенных значений и лишь четыре из них имеют доли пропусков от 1 до 7% — это приемлемый показатель, который позволяет нам работать с ними и применять различные стратегии заполнения пропущенных значений. При поиске дубликатов в наборе данных не было обнаружено повторяющихся записей.

Для начала взглянем на распределение целевого признака `Loan_status`:

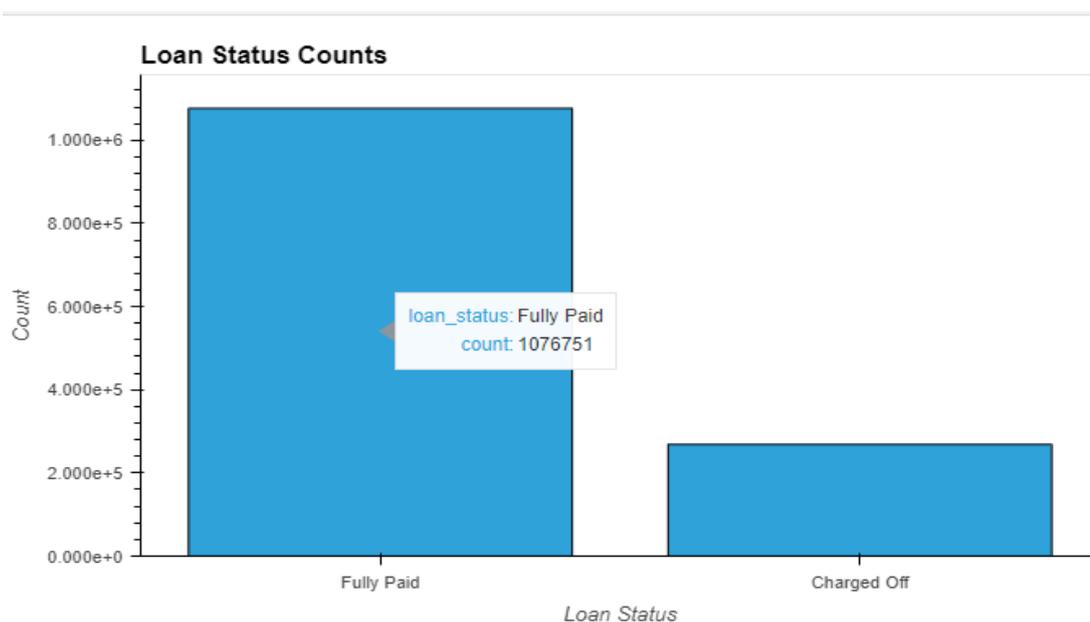


Рисунок 5 – Распределение признака `Loan_status`

Видно, что целевой признак распределен неравномерно, в пропорции 80% (`Fully Paid`) на 20% (`Charged Off`), т.е наблюдается дисбаланс классов.

Если взглянуть на эту пропорцию в разрезе подклассов кредита, то дисбаланс сохраняется в самых многочисленных категориях, однако, чем ниже класс кредита, тем выше доля наблюдений с меткой «`Charged Off`».

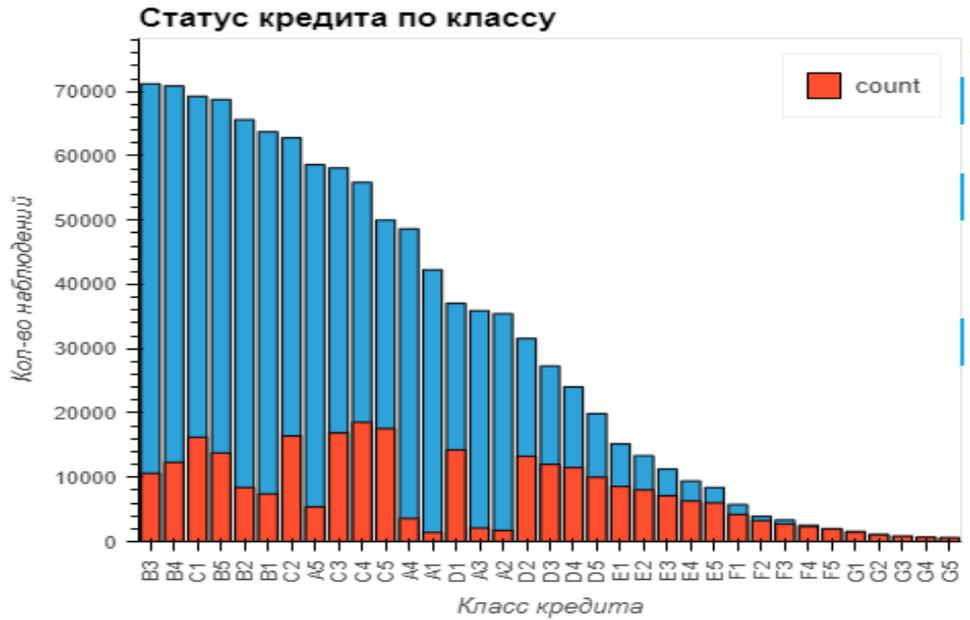


Рисунок 6 – Распределение признака loan_status в разрезе классов

Аналогичная картина наблюдается на рисунке 7: с увеличением процентной ставки доля выплаченных кредитов уменьшается.

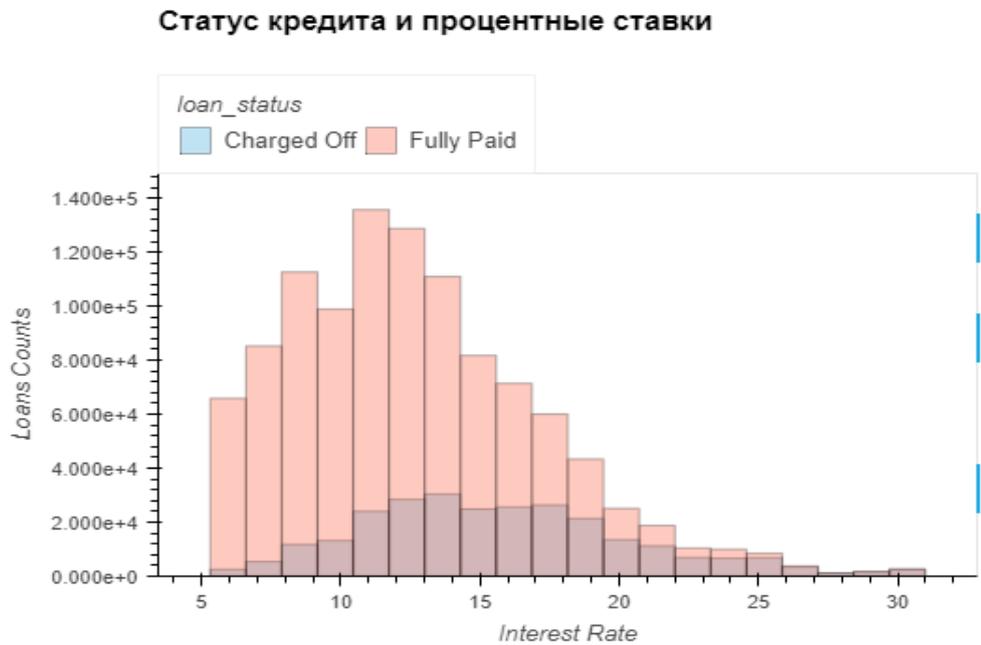


Рисунок 7 – Распределение loan_status в разрезе процентных ставок

На рисунке 8 представлено распределение стажа работы заёмщиков: выделяется количество заёмщиков со стажем более 10 лет, однако пропорция выплаченных и невыплаченных кредитов в каждой группе примерно на одном уровне.

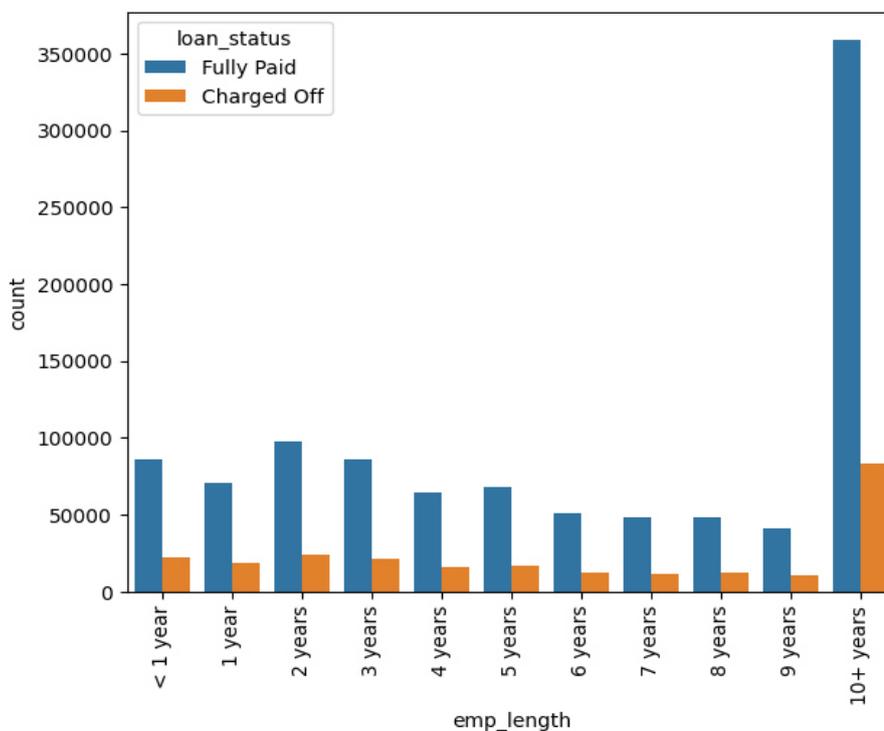


Рисунок 8 – Статус кредита в разрезе стажа работы заёмщика

Исследовательский интерес представляет информация о профессии заёмщика (рисунок 9): наиболее часто с запросом на займ обращались учителя, менеджеры, частные предприниматели и квалифицированные медработники.

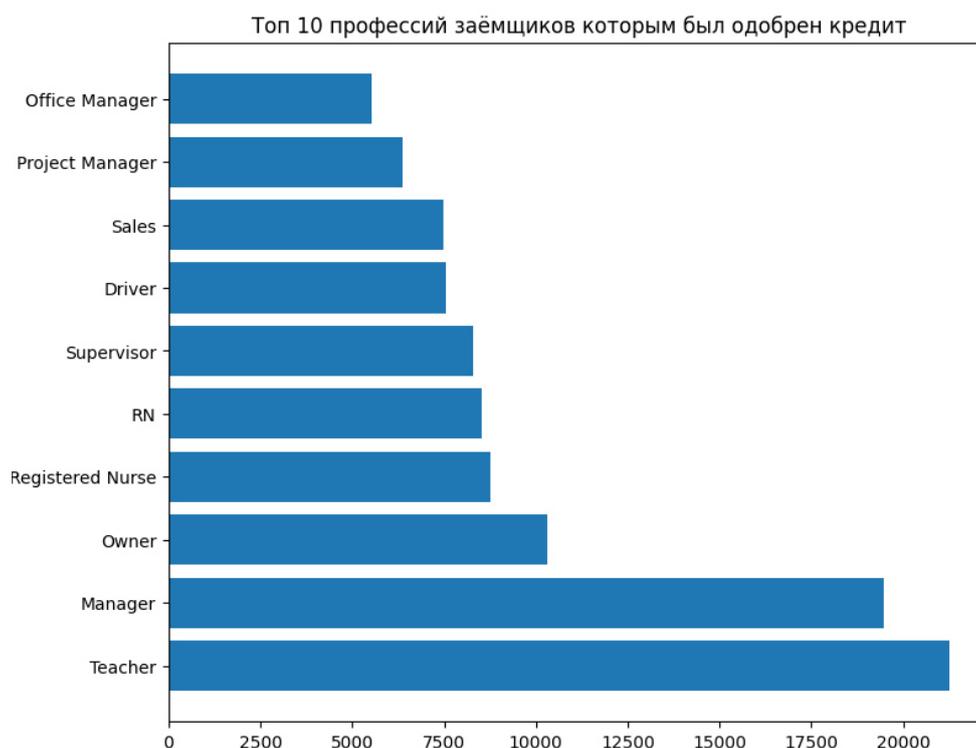


Рисунок 9 – Топ профессий заёмщиков

3.3. Предобработка данных датасета

Для построения модели прогнозирования дефолта заемщика на основе данных организации Lending Club следовало изначально провести предобработку данных. Предобработка данных – это процесс очистки и подготовки данных для анализа и построения моделей машинного обучения. В данном проекте была проведена предобработка данных, чтобы убрать выбросы и заполнить пропущенные значения. В процессе предобработки использовались как числовые, так и категориальные признаки.

В качестве стратегии борьбы с пропущенными значениями было выбрано полное удаление подобных записей с использованием метода «dropna» из библиотеки Pandas. Использование данной стратегии в нашем случае имеет несколько преимуществ: 1) Так как максимальный процент пропущенных значений признаков датасета не превышает 6%, а размер доступной выборки составляет более одного миллиона строк, то мы незначительно урезаем объём нашей выборки; 2) Замена значений на среднее,

медиану или используя параметры «bfill» и «ffill» метода fillna из библиотеки Pandas вносит смещение в выборку, то есть удаляя записи с пропущенными значениями, мы не избегаем этого смещения в данных.

Из датасета были удалены признаки Grade, Emp_title, Address, Issue_d, так как они либо дублируют информацию (Grade, Address), либо неинформативны (Emp_title), либо являются “leakage features” (Issue_d).

Далее были проведены действия по преобразованию категориальных данных, которые не могут быть прямо использованы в модели машинного обучения. Для этого был использован метод One-Hot Encoding, который позволил преобразовать категориальные признаки в бинарные, которые можно использовать в модели. Каждый уникальный категориальный признак был преобразован в отдельный столбец, где значение 1 указывает, что данный признак присутствует, а значение 0 – что он отсутствует. Это позволило модели работать с категориальными данными без необходимости задания порядка между ними.

Для обработки выбросов в числовых переменных был использован метод межквартильного размаха. Этот метод позволяет определить границы выбросов на основе распределения значений в переменной. Первоначально были вычислены первый и третий квартили (Q1 и Q3) для каждой числовой переменной. Затем был рассчитан межквартильный размах (IQR) как разница между Q3 и Q1. На основе IQR были определены верхняя и нижняя границы выбросов. Значения, находящиеся за пределами этих границ, были заменены на граничные значения, чтобы минимизировать их влияние на анализ. Это было достигнуто с использованием функции «clip» из библиотеки Pandas. Для каждой переменной были определены соответствующие граничные значения, значения вне этих границ были заменены на граничные значения.

Во время работы с данными была обнаружена проблема несбалансированных классов в задаче бинарной классификации. В данных было намного больше объектов с положительным классом (клиенты,

вернувшие кредит вовремя) по сравнению с отрицательным классом (клиенты, не вернувшие кредит вовремя). Так как это может привести к некорректной работе алгоритмов машинного обучения, было принято решение использовать методы `RandomOverSampler` и `RandomUnderSampler` из модулей `over_sampling` и `under_sampling` библиотеки `Imblearn` соответственно. Эти методы позволяют увеличить или уменьшить преобладающий класс, чтобы достичь баланса между классами.

3.4. Моделирование и интерпретация предсказаний

В рамках данной дипломной работы была реализована модель логистической регрессии, случайного леса и градиентного бустинга для решения задачи классификации. Для начала процесса моделирования было выполнено выделение целевой переменной и 25 признаков из исходного датасета. Затем данные были предобработаны путем преобразования категориальных признаков с помощью метода `One-Hot Encoding`. Для обучения и оценки модели были разделены данные на обучающую и тестовую выборки с помощью функции `train_test_split` из модуля `model_selection` библиотеки `Sklearn`. Оптимальные гиперпараметры для моделей были подобраны с помощью метода `RandomizedSearchCV` из модуля `model_selection` библиотеки `Sklearn`.

3.4.1 Логистическая регрессия

В ходе обучения модели логистической регрессии были найдены оптимальные гиперпараметры:

- 1) Решатель (`solver`) — линейный (`liblinear`);
- 2) Максимальное количество итераций, необходимое для сходимости решателей — 500;
- 3) Обратное значение силы регуляризации — 0.01.

Остальные гиперпараметры были оставлены по умолчанию класса LogisticRegression открытой библиотеки Scikit-learn.

В результате обучения модели были получены следующие результаты на обучающей и тестовой выборках, включающие 95-процентный доверительный интервал для оценок (таблица 2).

Таблица 2 – Метрики модели логистической регрессии

Метрика	Обучающая выборка	Тестовая выборка
Gini	0.415	0.421 ± 0.004
ROC-AUC	0.72	0.711 ± 0.004
F1	0.655	0.641 ± 0.004
Precision	0.655	0.659 ± 0.004
Recall	0.66	0.625 ± 0.004

Коэффициент Джини на тестовой выборке, который измеряет степень неравенства предсказаний, равен 0.42. Это относительно низкое значение, что указывает на то, что модель может быть улучшена. Площадь под ROC-кривой составляет 0.71. Это достаточно хороший результат, который указывает на то, что модель может отличать классы довольно надежно. Значение ROC-AUC в диапазоне 0.7-0.8 обычно считается "удовлетворительным". F1-мера, которая является гармоническим средним между точностью и полнотой, равна 0.65. Это довольно неплохой результат, который указывает на то, что модель достаточно хороша в предсказании положительного класса. Точность (Precision) модели составляет 0.66. Это означает, что из всех объектов, которые модель предсказала как положительные, 66% действительно являются положительными. Полнота (Recall) модели составляет 0.625. Это означает, что модель правильно идентифицировала около 63% из всех реальных положительных объектов.

Таблица сопряженности отражает разбиение наших предсказаний указывает на большое количество ложноположительных и ложноотрицательных предсказаний, примерно в пропорции 33% от выборки (Рисунок 10).

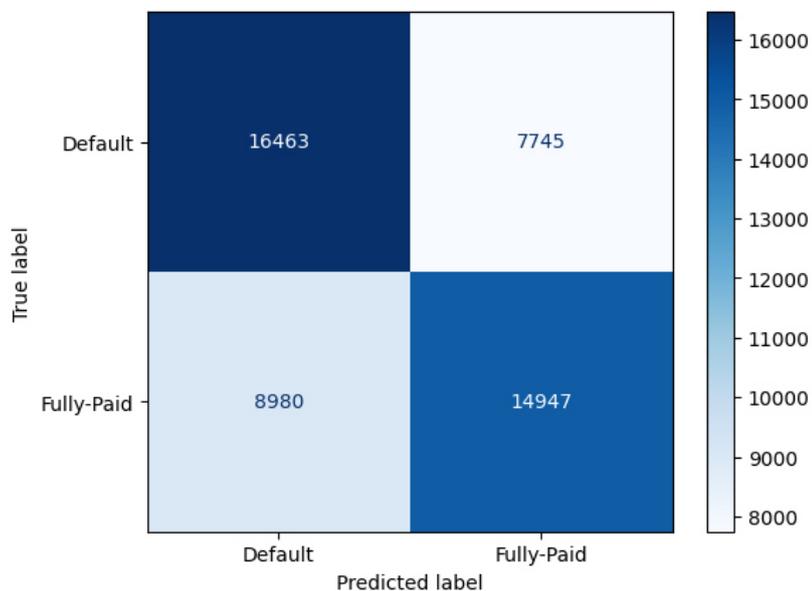


Рисунок 10 – Таблица сопряженности для логистической регрессии

График площади под ROC-кривой иллюстрирует вероятность в 71%, что модели удастся успешно разделить классы (Рисунок 11).

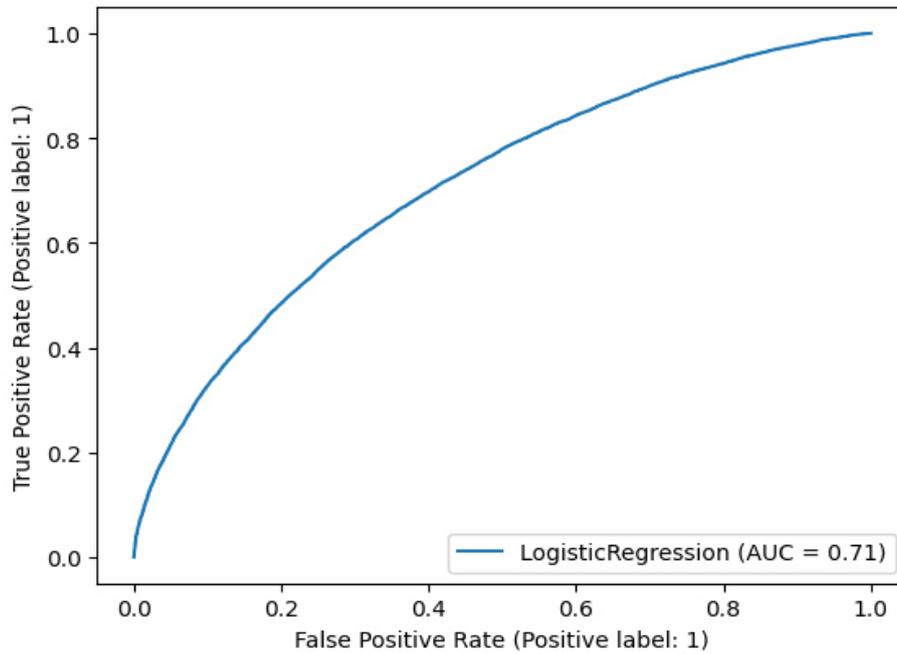


Рисунок 11 – График ROC-кривой для логистической регрессии

Интерпретацию влияния признаков модели логистической регрессии можно начать с анализа *feature importance* (Рисунок 12). Как отмечалось ранее *feature importance* это метод для глобальной интерпретации весовых коэффициентов модели. В данном случае лидируют признаки `add_state` и `purpose` преобразованные с помощью методов *One-Hot-Encoding*.

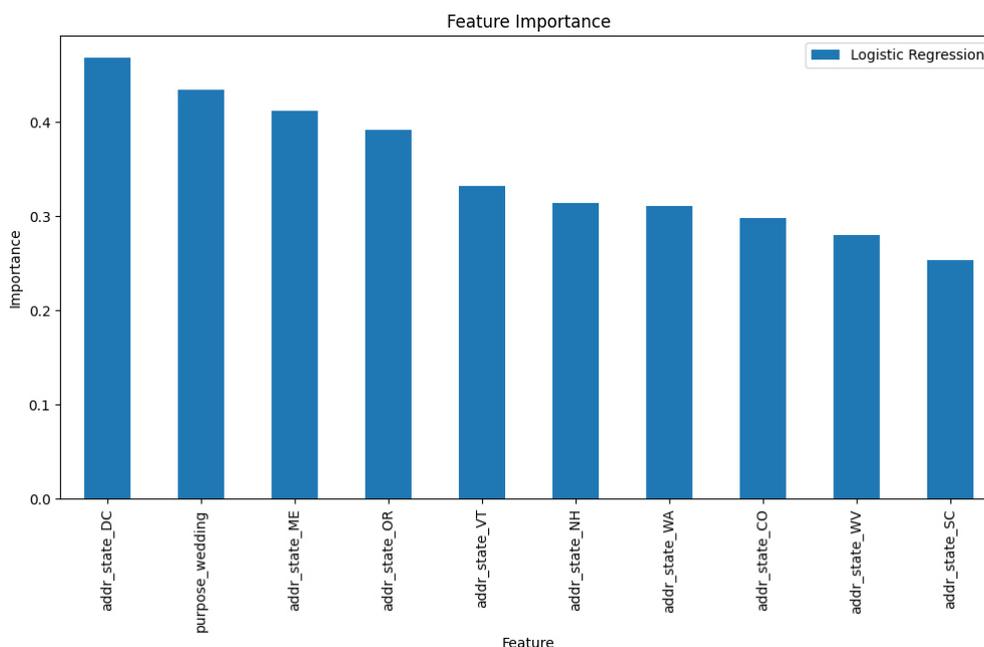


Рисунок 12 – График feature importance для логистической регрессии

Возвращаясь к задаче интерпретации отдельных прогнозных значений модели, а не ее общей структуры, следует рассмотреть метод LIME.

Рассмотрим трех разных заемщиков, чьи результаты отображены и визуализированы с помощью метода `show_in_notebook` из библиотеки `Lime` для Python.

В первом случае (A) заемщик имеет относительно низкую вероятность дефолта — недостаточную чтобы отклонить заявку. Интерпретатор LIME объясняет следующим образом (Рисунок 13):

- Сумма займа менее 8400 долларов (`loan_amnt` — 8000);
- Низкая процентная ставка по займу (`int_rate` — 8.9%);
- Относительно большой ежемесячный платеж по кредиту (`installment` — 254);
- Принадлежность к группе почтовых адресов (`dti`)

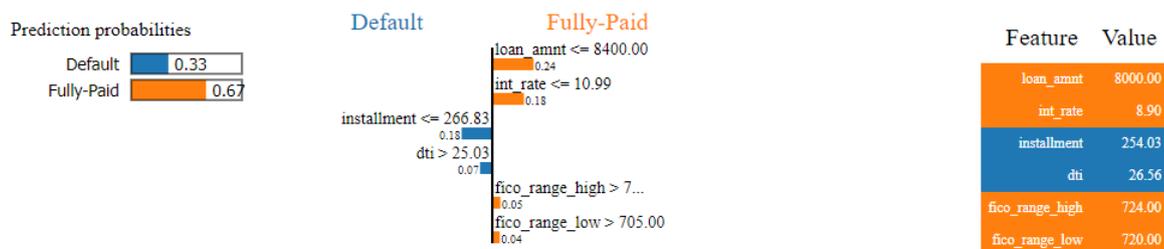


Рисунок 13 – Визуализация предсказания для заёмщика А методом LIME

Во втором случае (В) заемщик имеет высокую вероятность дефолта, но достаточную чтобы отклонить заявку. Интерпретатор LIME объясняет следующим образом (Рисунок 14):

- Сумма займа более 20000 долларов (loan_amnt — 30000);
- Относительно высокая процентная ставка по займу (int_rate — 14%);
- Принадлежность к группе с высоким дефолт-рейтом;

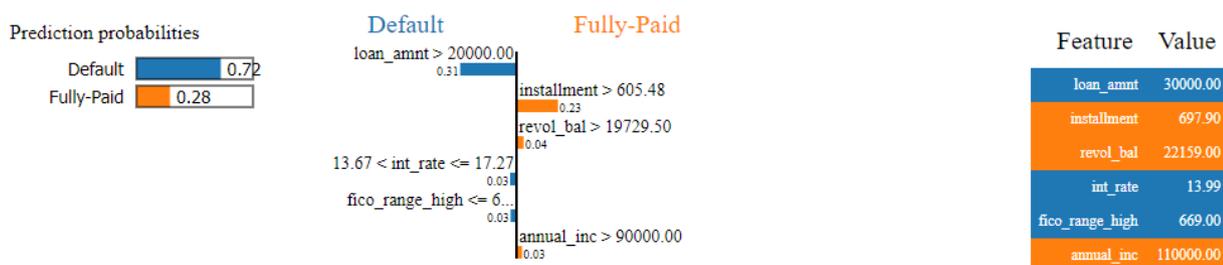


Рисунок 14 – Визуализация предсказания для заёмщика В методом LIME

В третьем случае (С) модель не уверена относительно своего предсказания, возвращая примерно одинаковые значения вероятностей для обоих целевых классов. Интерпретатор LIME объясняет это следующим образом (Рисунок 15):

- Богатая кредитная история, количество кредитных линий более 32 (total_acc - 33);
- Запрашиваемая сумма кредита достаточно велика (loan_amt — 19200);
- Умеренная процентная ставка по кредиту (int_rate — 12%)
- Принадлежность к группе с высоким дефолт-рейтом.

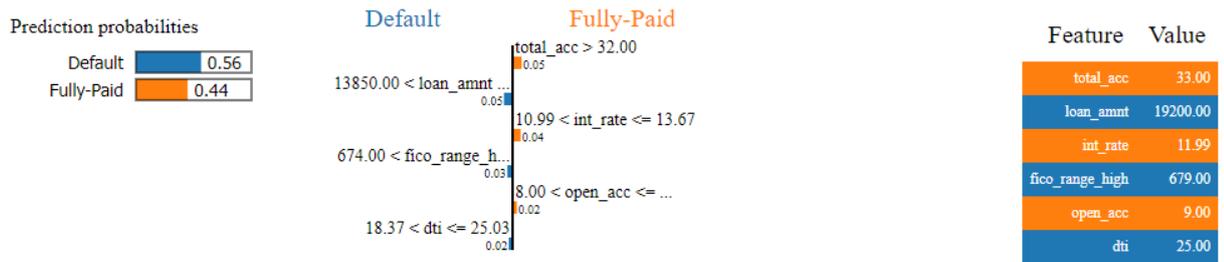


Рисунок 15 – Визуализация предсказания для заёмщика С методом LIME

С помощью подхода SHAP, мы также можем получить ответ на вопрос «почему» был сделан именно такой прогноз для отдельного наблюдения.

Рассмотрим в качестве примера тех же трёх заемщиков. Из графика для заёмщика А (Рисунок 16) видно, что существенный положительный вклад в предсказание был внесен факторами int_rate, loan_amnt и fico_range (суммарно 1.54), тогда как по показатель размера ежемесячного платежа (installment) снижает балл сразу на 0.43. Итоговый балл для данного заемщика составил 0.8, что говорит о умеренной «уверенности» модели относительно качества данного заемщика.

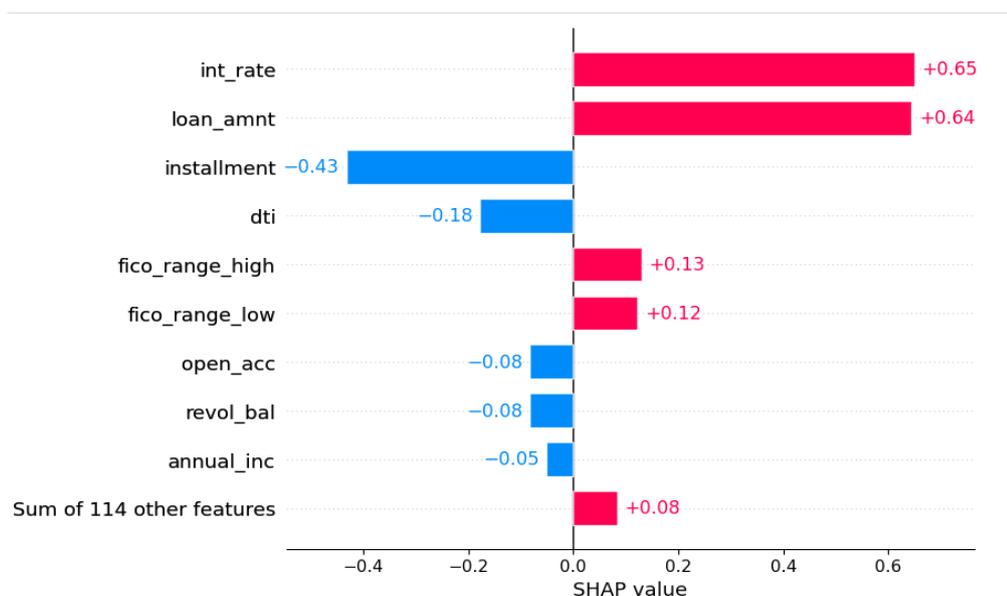


Рисунок 16 – Визуализация предсказания для заёмщика А методом SHAP

Из графика для заёмщика В (Рисунок 17) видно, что сильный отрицательный вклад в предсказание был внесен фактором `loan_amnt` (-1.34), тогда как самый крупный положительный признак (показатель размера ежемесячного платежа “`installment`”) поднимает балл лишь на 0.57. Итоговый балл для данного заемщика составил -0.81, что говорит о низкой «уверенности» модели относительно качества данного заемщика.

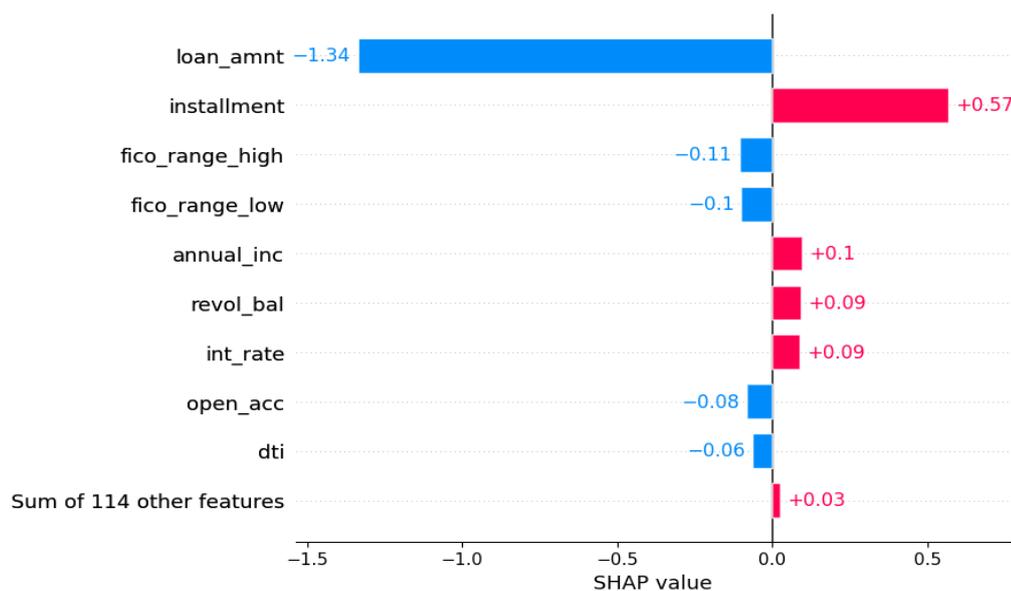


Рисунок 17 – Визуализация предсказания для заёмщика В методом SHAP

Интересная ситуация сложилась на графике для заёмщика С (Рисунок 18). Видно, что признаки примерно уравнивают предсказание. Итоговый балл для данного заемщика составил -0.39, что говорит о некоторой неопределенности модели относительно качества данного заемщика.

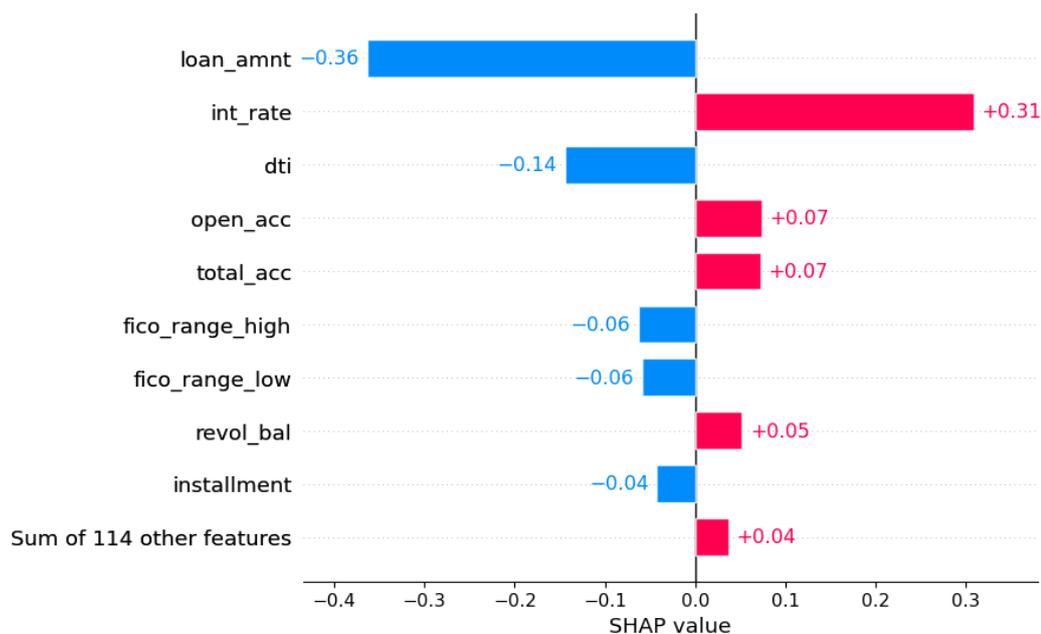


Рисунок 18 – Визуализация предсказания для заёмщика С методом SHAP

3.4.2. XGBoost

В ходе обучения модели градиентного бустинга были найдены оптимальные гиперпараметры:

- 1) Максимальная глубина дерева — 5;
- 2) Количество деревьев — 200;
- 3) Значение регуляризации L2 на весах (reg_lambda) — 5;
- 4) Значение регуляризации L1 на весах (reg_alpha) — 0;
- 5) Уменьшение размера шага при обновлении весов признаков (learning rate) — 0.1;
- е) Минимальное уменьшение потерь для дальнейшего разбиения дерева (gamma) — 1;
- ж) Глобальное смещение (base_score) — 0.2.

Остальные гиперпараметры были оставлены по умолчанию класса XGBClassifier открытой библиотеки XGBoost.

В результате обучения модели были получены следующие результаты на обучающей и тестовой выборках, включающие 95-процентный доверительный интервал для оценок (таблица 3).

Таблица 3 – Метрики модели бустинга

Метрика	Обучающая выборка	Тестовая выборка
Gini	0.445	0.446 ± 0.004
ROC-AUC	0.726	0.723 ± 0.004
F1	0.685	0.66 ± 0.004
Precision	0.685	0.67 ± 0.004
Recall	0.68	0.66 ± 0.004

Коэффициент Джини составляет 0.446. Это немного выше, чем у предыдущей модели, что указывает на более равномерное распределение предсказаний между классами. Площадь под ROC-кривой составляет 0.72. Это немного выше, чем у предыдущей модели, что так же указывает на то, что модель лучше различает классы. F1-мера составляет 0.66. Это немного выше, чем у предыдущей модели, что указывает на то, что модель XGBoost лучше справляется с предсказанием положительного класса. Так же на 1 п.п. подросли Precision (составляет 0.67) и Recall (0.66) из всех реальных положительных объектов.

Модель XGBoost показывает немного лучшие результаты по сравнению с моделью логистической регрессии. Она демонстрирует более высокую точность и полноту, что ведет к улучшению F1-меры. Кроме того, более высокое значение коэффициента Джини и ROC-AUC указывают на более равномерное распределение предсказаний и лучшую способность модели различать классы. Это делает модель XGBoost более предпочтительной для этой задачи.

Таблица сопряженности для XGBoost (Рисунок 19) отражает примерно ту же пропорцию ложноположительных и ложноотрицательных предсказаний что и в случае с логистической регрессией.

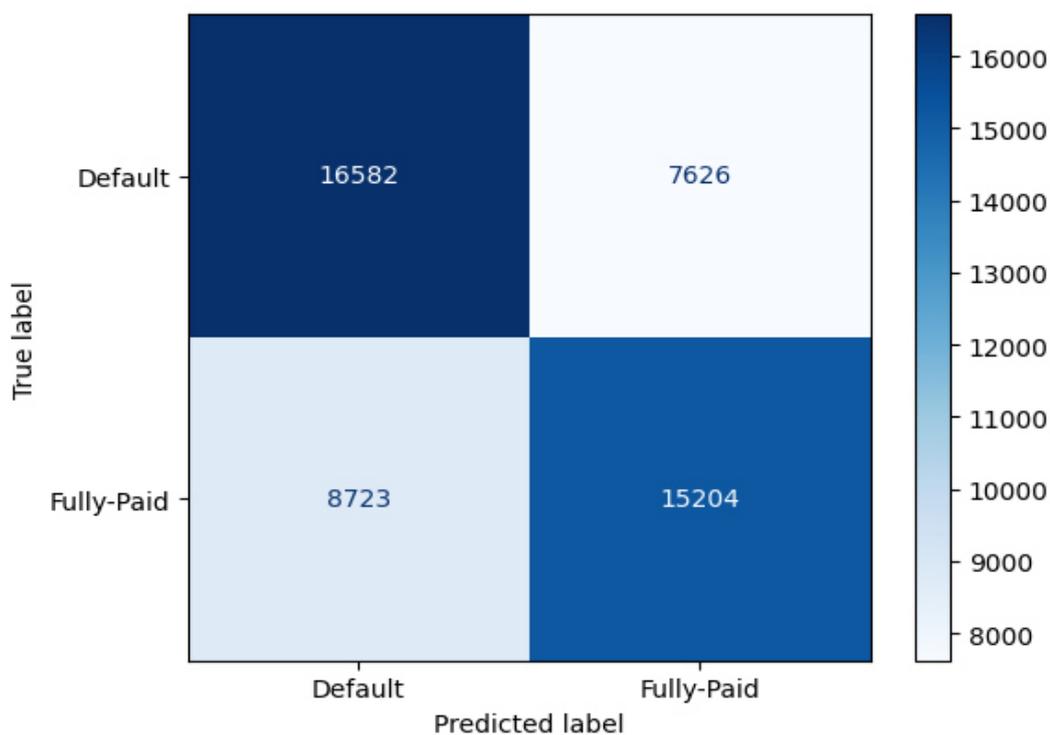


Рисунок 19 – Таблица сопряженности для бустинга

График площади под ROC-кривой иллюстрирует вероятность в 72%, что модели удастся успешно разделить классы (Рисунок 20).

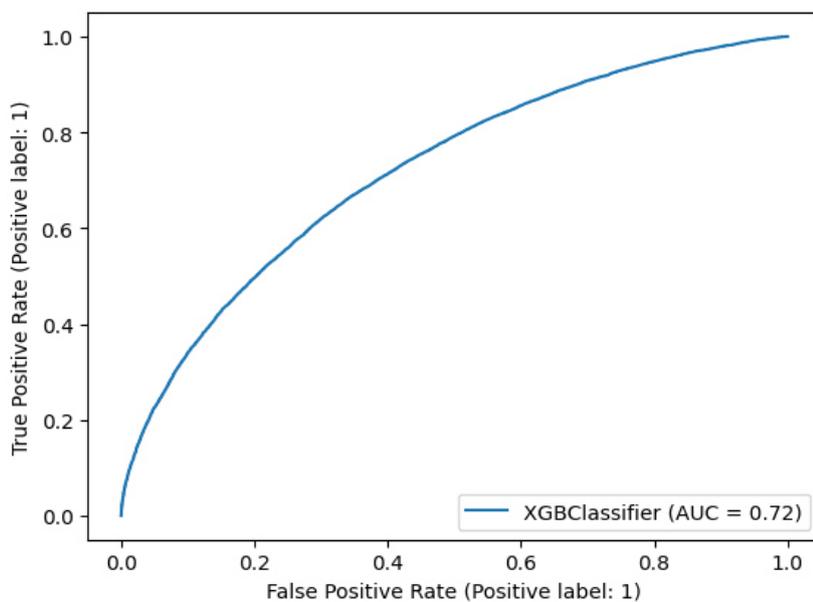


Рисунок 20 – График ROC-кривой для бустинга

Интерпретацию влияния признаков модели градиентного бустинга можно начать с анализа feature importance (Рисунок 21). В данном случае лидируют `int_rate`, `mort_acc` и признаки `sub_grade` и `home_ownership` преобразованные с помощью методов One-Hot-Encoding.

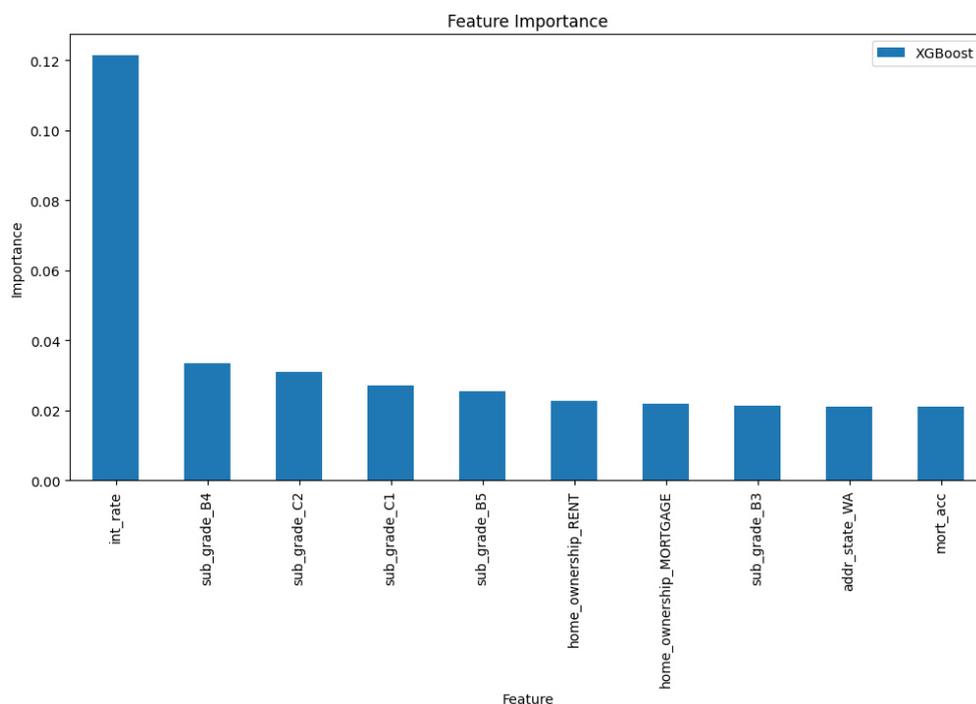


Рисунок 21 – График feature importance для бустинга

Проведем интерпретацию на тех же примерах заёмщиков, которые использовались выше для модели логистической регрессии. Рассмотрим метод LIME.

В первом случае А (Рисунок 22) заемщик так же имеет относительно низкую вероятность дефолта. Только в случае XBoost, модель более уверена в выплате суммы кредита заёмщиком. Вклад признаков в предсказание аналогичен примеру с логистической регрессией.

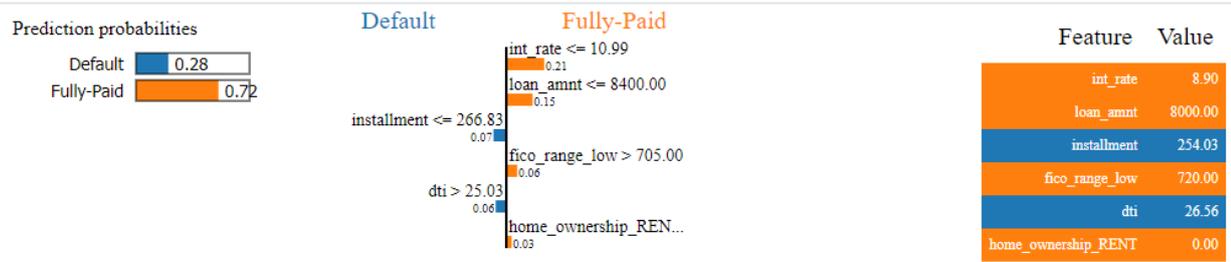


Рисунок 22 – Визуализация предсказания для заёмщика А методом LIME

Во втором случае В (Рисунок 23) заемщик также имеет высокую вероятность дефолта, но в данном случае эта вероятность уже ниже. Вклад признаков в предсказание аналогичен примеру с логистической регрессией, однако важность признаков loan_amnt и installment снижена.

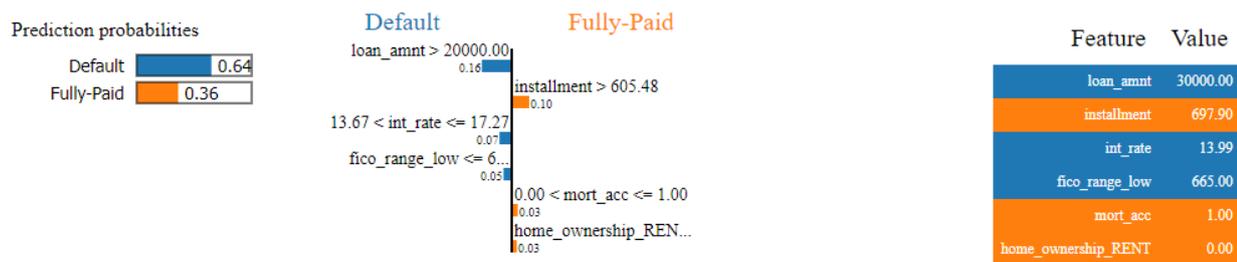


Рисунок 23 – Визуализация предсказания для заёмщика В методом LIME

В третьем случае С (Рисунок 24) модель так же не уверена относительно своего предсказания, возвращая примерно одинаковые значения вероятностей для обоих целевых классов. Отмечу, что незначительно изменились веса признаков в предсказание.

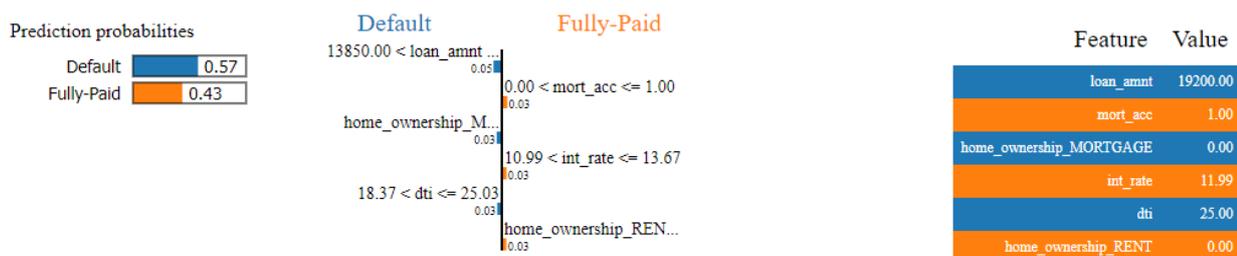


Рисунок 24 – Визуализация предсказания для заёмщика С методом LIME

Проведу аналогичное сравнение с интерпретацией предсканий модели XGBoost с использованием метода SHAP на тех же примерах.

Из графика для заёмщика А (Рисунок 25) видно, что существенный положительный вклад в предсказание был внесен теми же факторами `int_rate`, `loan_amnt` и `fico_range` (суммарно 1.14), тогда как показатель размера ежемесячного платежа (`installment`) снижает балл на 0.17. Итоговый балл для данного заемщика составил 0.88, что говорит модели XGBoost более уверена в возврате кредита заёмщиком А, нежели модель логистической регрессии.

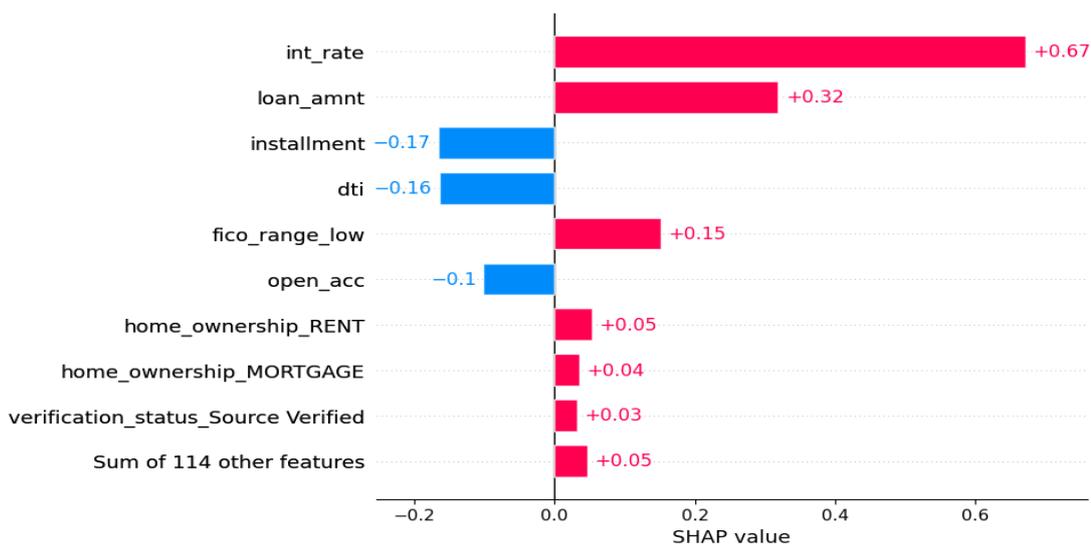


Рисунок 25 – Визуализация предсказания для заёмщика А методом SHAP

Из графика для заёмщика В (Рисунок 26) видно, что сильный отрицательный вклад в предсказание в этот раз был внесен несколькими факторами: `loan_amnt` (-0.48), `int_rate` (-0.26), `fico_range_low` (-0.18) и `open_acc` (-0.13) тогда как самый крупный положительный признак `installment` поднимает балл лишь на 0.1. Итоговый балл для данного заемщика составил -0.68, что говорит о низкой «уверенности» модели относительно качества

данного заемщика, а так же о том, что модель XGB оценивает вероятность возврата кредита заёмщиком В чуть выше, чем логистическая регрессия.

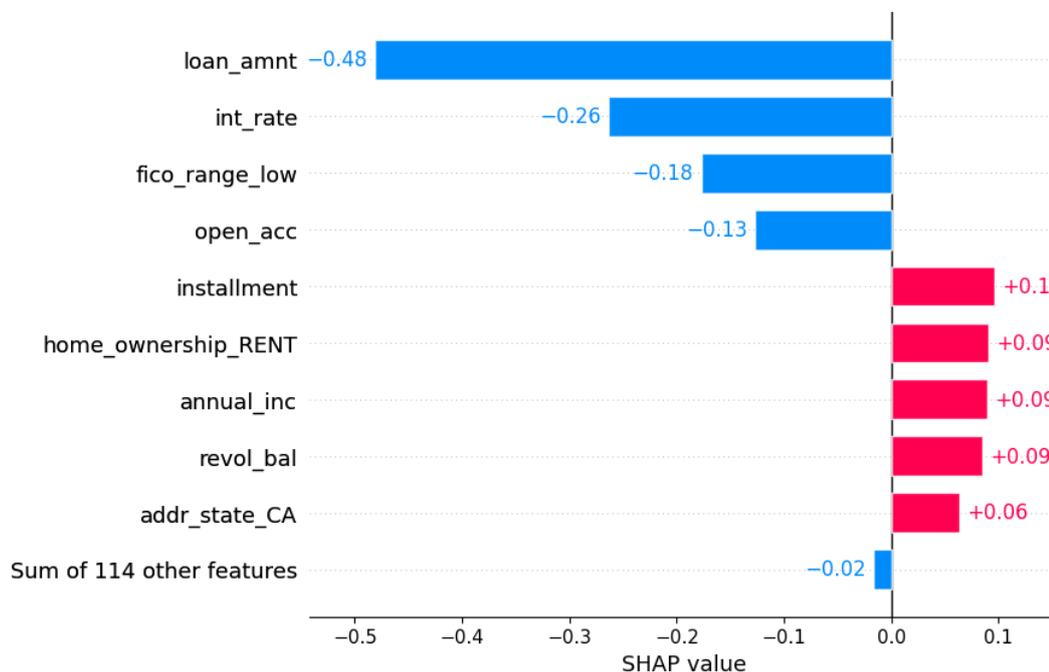


Рисунок 26 – Визуализация предсказания для заёмщика В методом SHAP

Интересная ситуация сложилась на графике для заёмщика С (Рисунок 27). Итоговый балл для данного заемщика составил -0.39, что говорит о низкой «уверенности» модели относительно качества данного заемщика. Отмечу, что SHAP в случае предсказания моделью логистической регрессии давал более неопределённое предсказание.

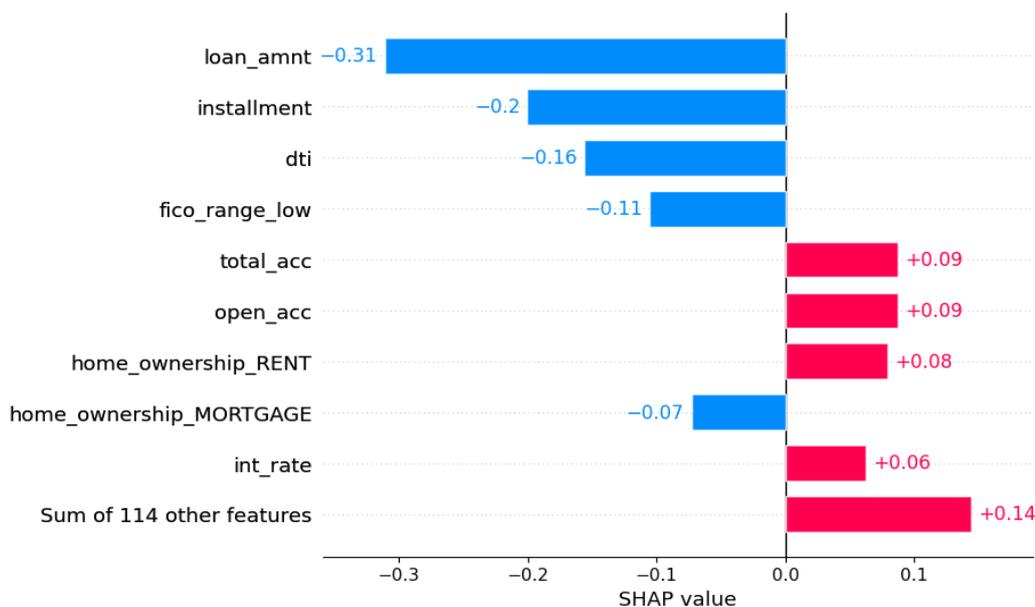


Рисунок 27 – Визуализация предсказания для заёмщика С методом SHAP

3.4.3. RandomForest

В ходе обучения модели случайного леса были найдены оптимальные гиперпараметры:

- 1) Максимальная глубина дерева — 20;
- 2) Минимальное количество образцов, необходимое для нахождения в узле листа — 2;
- 3) Количество деревьев — 25.

Остальные гиперпараметры были оставлены по умолчанию класса RandomForestClassifier открытой библиотеки Scikit-learn.

В результате обучения модели были получены следующие результаты на обучающей и тестовой выборках, включающие доверительный интервал для оценок (таблица 4).

Таблица 4 – Метрики модели случайного леса

Метрика	Обучающая выборка	Тестовая выборка
Gini	0.425	0.432 ± 0.004

ROC-AUC	0.81	0.71 ± 0.004
F1	0.67	0.65 ± 0.004
Precision	0.69	0.67 ± 0.004
Recall	0.66	0.64 ± 0.004

Продолжение таблицы 4

Коэффициент Джини составляет 0.432. Это немного ниже, чем у модели XGBoost, но выше, чем у модели логистической регрессии, что указывает на относительно равномерное распределение предсказаний между классами. Площадь под ROC-кривой составляет 0.71. Это такое же значение, как у модели логистической регрессии, и немного ниже, чем у модели XGBoost. F1-мера составляет 0.65. Это такое же значение, как у модели логистической регрессии, и немного ниже, чем у модели XGBoost. Precision модели составляет 0.67. Это такое же значение, как у модели XGBoost, и выше, чем у модели логистической регрессии. Recall модели составляет 0.64 — здесь ситуация аналогична случаю с метрикой Precision.

Модель RandomForest показывает смешанные результаты по сравнению с моделями логистической регрессии и XGBoost. Она имеет средний коэффициент Джини и ROC-AUC, такую же точность, как у модели XGBoost, и такую же F1-меру, как у модели логистической регрессии. Однако ее полнота ниже, чем у других моделей. Это может указывать на то, что модель RandomForest может быть менее эффективной при идентификации реальных положительных объектов.

Таблица сопряженности для RandomForest (Рисунок 28) отражает примерно ту же пропорцию ложноотрицательных предсказаний что и в случае с логистической регрессией и градиентным бустингом. Однако выросла доля ложноположительных предсказаний по сравнению с другими моделями - на это указывает и падение метрики Recall.

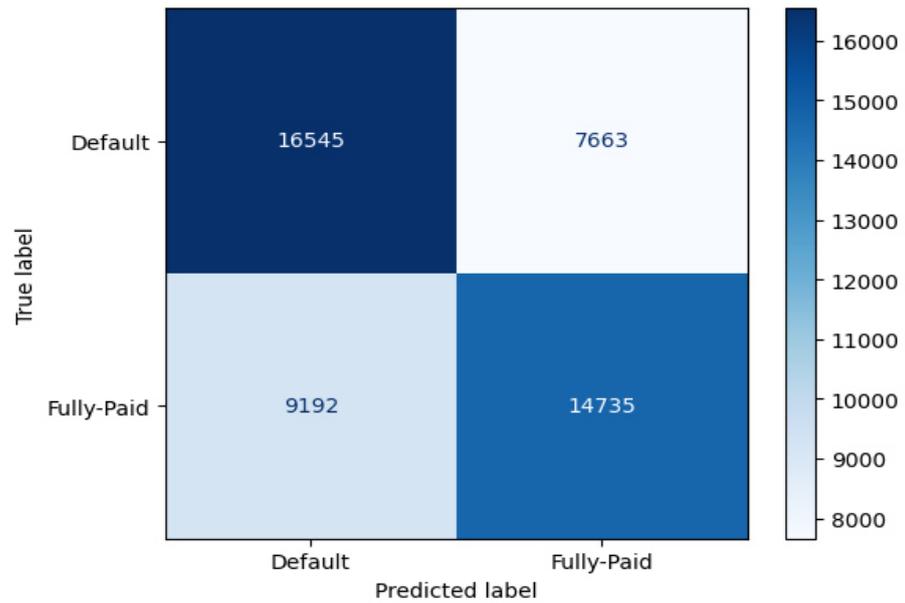


Рисунок 28 – Таблица сопряженности для случайного леса

График площади под ROC-кривой иллюстрирует вероятность в 71%, что модели удастся успешно разделить классы (Рисунок 29). Этот показатель не отличается от результатов линейной регрессии и уступает результатам градиентного бустинга.

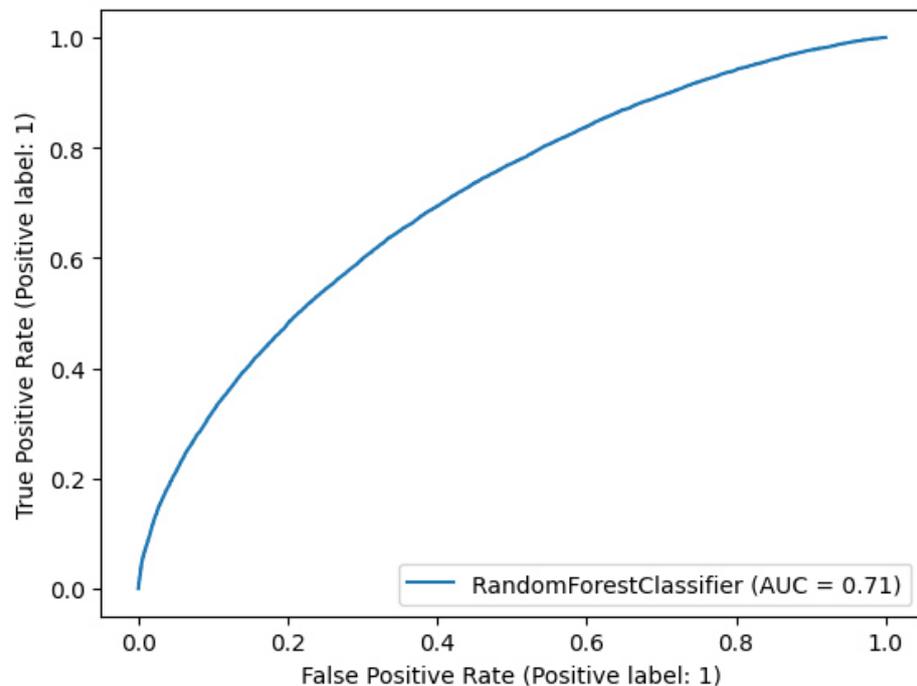


Рисунок 29 – График ROC-кривой для случайного леса

Интерпретацию влияния признаков модели случайного леса можно начать с анализа feature importance (Рисунок 30). В данном случае лидируют `int_rate`, `dti` и `instalment`. Отмечу что преобразованные признаки с помощью методов One-Hot-Encoding не попали в выборку самых значимых.

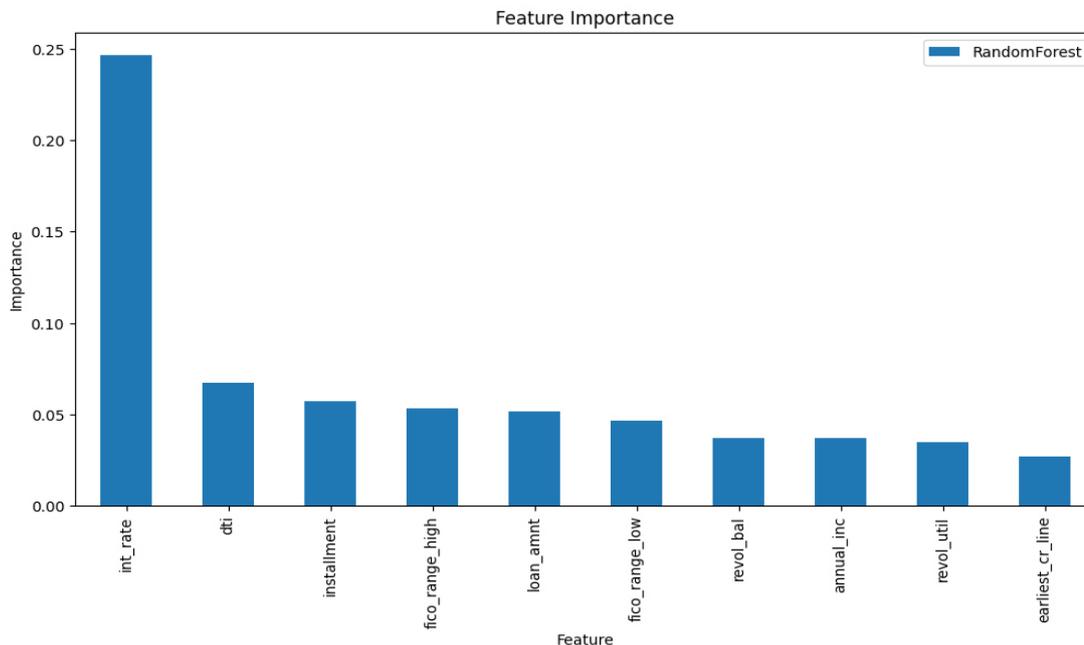


Рисунок 30 – График feature importance для случайного леса

Проведем интерпретацию на тех же примерах заёмщиков, которые использовались выше для двух других моделей. Рассмотрим метод LIME.

В первом случае А заёмщик также имеет относительно низкую вероятность дефолта (Рисунок 31). Примечательно, что в качестве положительного признака появляется подкласс кредита (`sub_grade_A5` — 1), а `instalment` из отрицательного признака стал положительным.

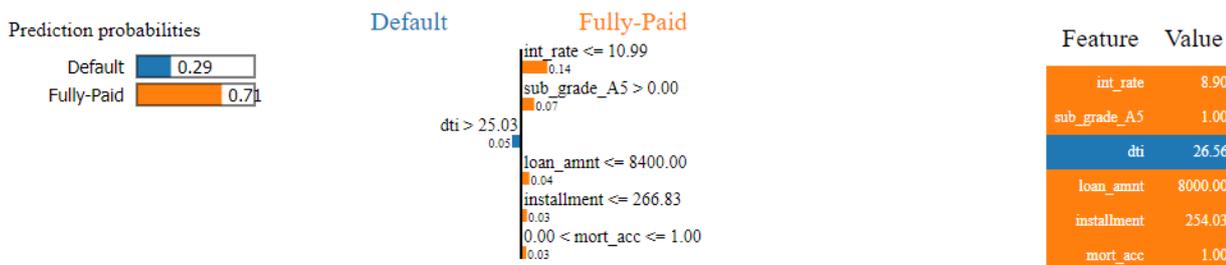


Рисунок 31 – Визуализация предсказания для заёмщика А методом LIME

Во втором случае В модель Random Forest уже менее уверена в дефолте заёмщика, нежели логистическая регрессия или XGB (Рисунок 32). На наличие неопределенности подталкивают обозначенные признаки и их веса: первые три признака с отрицательными значениями обозначают, что кредит не относится к подгруппам А3, А5, В1, но они скорее являются косвенными характеристиками заёмщика.

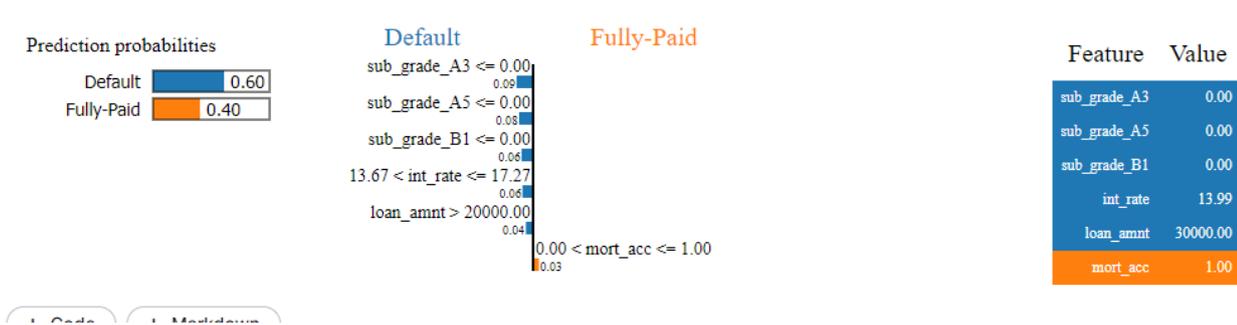


Рисунок 32 – Визуализация предсказания для заёмщика В методом LIME

В третьем случае С модель также не уверена относительно своего предсказания, возвращая примерно одинаковые значения вероятностей для обоих целевых классов и также в качестве самых весомых признаков фигурируют метки отсутствия подкласса кредита А или В (Рисунок 33).

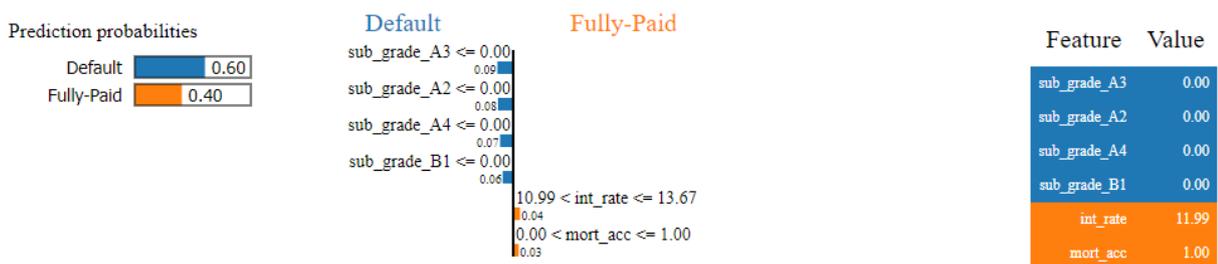


Рисунок 33 – Визуализация предсказания для заёмщика С методом LIME

Так же для визуализации локальных предсказаний модели можно использовать метод force_plot из библиотеки SHAP.

Из графика для заёмщика А (Рисунок 34) видно, что существенный положительный вклад в предсказание был внесен теми же факторами `int_rate`, `sub_grade_A5` и `fico_range_high`, тогда как среди негативных факторов со значительным влиянием на прогноз значится только признак `dti`. Итоговый балл для данного заемщика выше `base_value`, что говорит модель RandomForest скорее уверена в возврате кредита заёмщиком А.



Рисунок 34 – Визуализация предсказания для заёмщика А методом SHAP

Из графика для заёмщика В (Рисунок 35) видно, что сильный отрицательный вклад в предсказание в этот раз был внесен несколькими факторами: `loan_amnt`, `dti`, `fico_range` и `int_rate` тогда как самыми крупными положительными признаками являются `mort_acc` и `annual_inc`. Итоговый балл для данного заемщика ниже `base_value`, что говорит о низкой «уверенности» модели относительно качества данного заемщика.

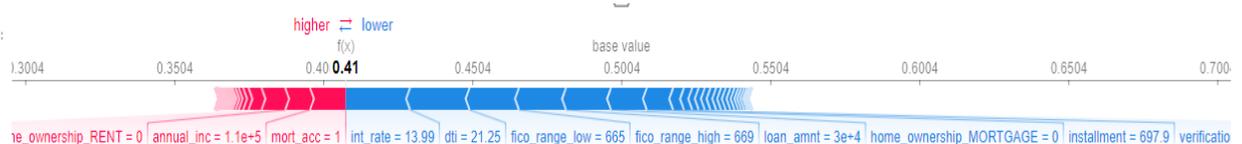


Рисунок 35 – Визуализация предсказания для заёмщика В методом SHAP

Интересная ситуация сложилась на графике для заёмщика С (Рисунок 36). Итоговый балл для данного заемщика составил 0.47 и очень близок к значению `base_value`, что говорит о неопределенности модели относительно качества данного заемщика. Отмечу, что SHAP в случае заёмщика С и модели логистической регрессии так же давал неопределённое предсказание.

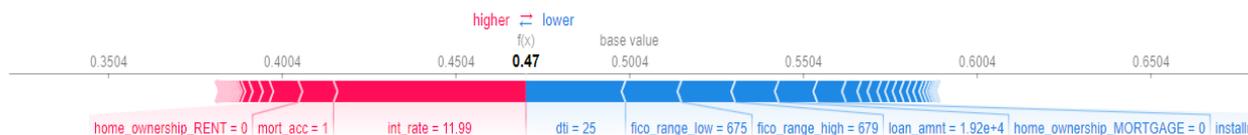


Рисунок 36 – Визуализация предсказания для заёмщика С методом SHAP

3.5. Анализ эффективности. Перспектива использования и развития инструмента

Таблица 5 – Сравнение качества моделей на тестовой выборке

Метрика	Модель		
	Логистическая регрессия	Случайный лес	Градиентный бустинг
Gini	0.421 ± 0.004	0.431 ± 0.004	0.446 ± 0.004
ROC-AUC	0.711 ± 0.004	0.721 ± 0.004	0.723 ± 0.004

В данном блоке мы сравним результаты прогнозирования по трём моделям. Как мы видим из таблицы 5, алгоритм градиентного бустинга показывает более высокий коэффициент Gini на 2.5 % и более высокую оценку ROC-AUC на 1.1 % в сравнении с логистической регрессией - этот факт говорит о превосходстве данной модели по критерию качества прогноза. Так же отмечу что модель случайного леса едва превзошла алгоритм логистической регрессии по указанным метрикам и всё же уступила градиентному бустингу. Вероятно, показатели метрик можно улучшить, используя другие методы для устранения дисбаланса классов и предварительной обработки данных. Так же можно произвести более тонкую настройку гиперпараметров не задействованных при использовании метода RandomizedSearchCV.

Продемонстрированные примеры интерпретации предсказаний ансамблевых моделей машинного обучения для задачи предсказания

вероятности дефолта заёмщика с использованием методов LIME и SHAP позволяют сделать предположения о наличии ряда преимуществ этих методов в сравнении с возможностью модели логистической регрессии:

– Детализация предсказаний — как и в случае модели логистической регрессии, SHAP и LIME позволяют интерпретировать отдельные прогнозы, а не только общую структуру модели. Это означает, что вы можете понять, какие конкретные признаки и их значения влияют на решение о предоставлении кредита для конкретного клиента;

– Сохранение интерпретируемости при усложнении скоринговой модели — ансамблевые методы машинного обучения часто являются "черными ящиками" для представителей бизнес-подразделений кредитной организации, и так как представители бизнеса как правило напрямую взаимодействуют с предполагаемыми заёмщиками, то использование результата предсказания модели градиентного бустинга в «сыром виде» для вынесения взвешенного решения о выдаче кредита не представляется возможным. SHAP и LIME позволяют «взглянуть внутрь» этих моделей и понять, какие признаки влияют на прогнозы в отдельном и общем случае. Это свойство может позволить использовать ансамблевые модели в промышленной среде в соответствии с законодательными требованиями местных регуляторов (например, в США, Российской Федерации и других странах);

– Улучшение моделей - интерпретация предсказаний с использованием LIME и SHAP может помочь кредитной организации улучшить их модели. Например, они могут обнаружить, что некоторые признаки вносят неожиданный вклад в прогнозы, что может указывать на проблемы с данными или моделью.

Отдельно отмечу превосходство ансамблевых методов для построения скоринговой системы в точности прогноза не только на технических метриках (Precision, Recall, F1-мера), но и на метриках, имеющих приложение к бизнесу

(Gini, ROC-AUC). Однако, это утверждение верно в случае, если модель хорошо специфицирована под имеющиеся данные.

Таким образом, несмотря на то, что модели логистической регрессии могут быть более простыми и интерпретируемыми, SHAP и LIME предлагают значительные преимущества для интерпретации более сложных и возможно более точных моделей машинного обучения в задаче предсказания дефолта заёмщика. Использование методов LIME и SHAP для ансамблевых моделей машинного обучения в перспективе может снизить кредитные риски организации и снизить её финансовые потери.

Помимо этого, графики интерпретации локального предсказания модели, полученные с помощью методов LIME и SHAP, можно встроить в инструменты работника фронт-офиса банка. Это может быть особенно полезно в качестве рекомендательного инструмента при вынесении решения по выдаче кредита заемщику. Например, можно также совместно выводить графики методов LIME, SHAP и весовых коэффициентов логистической регрессии в этом инструменте, чтобы сотрудник располагал исчерпывающей информацией по заемщику. Однако, работники должны быть обучены пониманию и интерпретации этих графиков. Они должны понимать, что представляют собой эти графики, как интерпретировать их и как использовать эту информацию при принятии решений.

ЗАКЛЮЧЕНИЕ

В рамках работы мы увидели, что, как отмечалось в разделе с обзором литературы, ансамблевые методы прогнозирования вероятности дефолта имеют более высокую предсказательную силу, чем логистическая регрессия. При этом важно отметить, что прирост в предсказательной способности за счет построения более сложной модели не повлек за собой существенных проблем с интерпретацией. Более того, мы успешно рассмотрели различные варианты объяснений как модели в целом, так и локальных предсказаний с помощью популярных на текущий момент методик.

Конечно, выводы и результаты, полученные в результате данной работы, не претендуют на наличие прорывных научных идей, однако они имеют прямое практическое применение в бизнесе, и именно в этом заключается главная ценность данной работы - повышение предсказательной способности модели, конвертируемое в рост ожидаемых прибылей для банков, которые станут ее использовать, а также демонстрация ее интерпретируемости, что обеспечит доверие к ней.

В части следующих шагов, можно выделить два основных направления. Во-первых, в работе не было предложено последовательного подхода к интерпретации сложной модели. Поэтому в будущем было бы полезно иметь конкретный бизнес-процесс, связанный с объяснением предсказанных моделью значений. Данная идея не входила в цели работы, однако служит одной из отправных точек для дальнейшего анализа по заданной теме.

Вероятно, появление инструментов объяснения предсказаний сложных моделей вряд ли вытеснит в краткосрочной перспективе классические методы, однако данная работа позволяет сделать шаг в сторону автоматизированного подхода, основанного на современных и более эффективных прогнозных моделях, что в конечном итоге должно положительно отразиться на стабильности кредитных организаций.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Акимов А.А., Валитов Д.Р., Кубряк А.И. Предварительная обработка данных для машинного обучения / Акимов А.А., Валитов Д.Р., Кубряк А.И. // Научное обозрение. Технические науки. – 2022. – №. 2. – С. 26–31.
2. Кочеткова В.В., Ефремова К.Д. Обзор методов кредитного скоринга [Электронный ресурс] // *Juvenis scientia*. – 2017: <https://cyberleninka.ru/article/n/obzor-metodov-kreditnogo-skoringa> (дата обращения: 01.03.2024)
3. Официальный сайт компании FICO. [Электронный ресурс]: <https://www.fico.com/en/history> (дата обращения: 01.03.2024)
4. Юзвович Л.И. Финансовые и банковские риски: учебник / Л.И. Юзвович, Ю.Э. Слепухина, Ю.А. Долгих, В.А. Татьянников, Е.В. Стрельников, Р.Ю. Луговцов, М.Н. Клименко; под ред. Л.И. Юзвович, Ю.Э. Слепухиной; Министерство науки и высшего образования Российской Федерации, Уральский государственный экономический университет. – Екатеринбург: Изд-во Урал. ун-та, 2020. – 336 с.: ил
5. Федеральный закон «О банках и банковской деятельности» от 02.12.1990 N 395-1 Статья №25. [Электронный ресурс]: https://www.consultant.ru/document/cons_doc_LAW_5842/5890a39bd69d0484d1fc62b3ce51bb043c25e4a3 (дата обращения: 06.03.2024)
6. All Lending Club loan data. [Электронный ресурс]: <https://www.kaggle.com/datasets/wordsforthewise/lending-club> (дата обращения: 20.03.2024)
7. Altman I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy // *Journal of Finance*. – 1968. – P. 189–209.
8. Beaver W. Financial Ratios as Predictors of Failure // *Journal of Accounting Research*. – 1966. – Vol. 4. – P. 71–111.

9. Demyanyk Y. Your Credit Score Is a Ranking, Not a Score [Электронный ресурс] // Economic Commentary. – 2010. – №2010-16: <https://www.clevelandfed.org/publications/economic-commentary/2010/ec-201016-your-credit-score-is-a-ranking-not-a-score> (дата обращения: 01.03.2024)
10. Ilter D. Credit scoring by artificial neural networks based cross-entropy and fuzzy relations. / D. Ilter, O. Kocadagli // Sigma Journal of Engineering and Natural Sciences. – 2019. – 37(3). – P. 855–870.
11. Främling K. Feature Importance versus Feature Influence and What It Signifies for Explainable AI // Springer Nature Switzerland. – 2023. – P. 241–259.
12. Fisher A. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously / A. Fisher, C. Rudin, F. Dominici // Journal of Machine Learning Research. – 2019. – Т. 20. – №. 177. – P. 1–81.
13. Friedman J. Greedy function approximation: A gradient boosting machine // The Annals of Statistics. – №. 5. – P. 1189–1232.
14. García S. Data Preprocessing in Data Mining / S. García, J. Luengo, F. Herrera // Springer International Publishing. – 2015. – 320 p.
15. Hastie T. The Elements Of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman // Springer Series in Statistics. – 2009. – 745 p.
16. Lessmann S. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research / S. Lessmann, B. Baesens, H. Seow, L. Thomas // European Journal of Operational Research. – 2015. [Электронный ресурс]: https://www.researchgate.net/publication/276280838_Benchmarking_state-of-the-art_classification_algorithms_for_credit_scoring_An_update_of_research (дата обращения: 96.03.2024)
17. Lundberg S.M. A unified approach to interpreting model predictions / S.M. Lundberg, S.I. Lee // Advances in neural information processing systems. – 2017. – Т. 30. – P. 1–10.

18. Ribeiro M. T. «Why should i trust you?» Explaining the predictions of any classifier / M.T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – 2016. – P. 1137–1141.
19. Sayah F. Lending Club Loan: Defaulters prediction [Электронный ресурс]: <https://www.kaggle.com/code/faessayah/lending-club-loan-defaulters-prediction> (дата обращения: 28.03.2024)
20. Schechtman E. The relationship between Gini terminology and the ROC curve / E. Schechtman, G. Schechtman // METRON, Springer, Sapienza Università di Roma. – vol. 77(3). – 2019. – P. 171–178.
21. Weng C. Hybrid Machine Learning Model for Credit Approval / C. Weng, C.A. Huang // Applied Artificial Intelligence. – 35:15. – 2021. – P. 1439–1465.
22. Zhang R. Optimizing hyper-parameters of neural networks with swarm intelligence: A novel framework for credit scoring. / R. Zhang, Z. Qiu // PLOS ONE. – 15(6). – 2020. [Электронный ресурс]: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234254> (дата обращения: 06.03.2024).