Санкт-Петербургский государственный

университет

**Жак Роман Викторович**

**Выпускная квалификационная работа**

Математические методы оценки кредитного риска при различных типах пропусков в данных

Направление 38.04.01 «Экономика»

Основная образовательная программа магистратуры ВМ.5629.2016 «Математические методы в экономике»

Научный руководитель:

докт. экон. наук, доцент

Кудрявцев А.А.

Рецензент:

начальник отдела,

Отдел систем анализа и прогнозирования,

Публичное акционерное общество «Газпром нефть»

Черницын И. Г.

Санкт-Петербург

2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 3

ГЛАВА 1. ЗАКОНОМЕРНОСТИ И ОСОБЕННОСТИ РАБОТЫ С ПРОПУЩЕННЫМИ ЗНАЧЕНИЯМИ ПРИ ОЦЕНКЕ КРЕДИТНОГО РИСКА 6

1.1. Основные подходы к оценке кредитного риска при неполных данных 9

1.2. Математические и инструментальные методы оценки кредитного риска 15

1.3. Методы работы с пропущенными значениями и их анализ 19

1.3.1. Базовые методы обработки пропущенных значений 19

1.3.2. Продвинутые методы обработки пропущенных значений 20

1.3.3. Программное обеспечение и особенности базисной выборки 24

Выводы 26

ГЛАВА 2. ПРОГНОЗИРОВАНИЕ ИСХОДОВ ПРИ РАЗЛИЧНЫХ РЕЖИМАХ ПРОПУСКОВ 27

2.1. Описание распределений пропусков 30

2.2. Исследование режимов пропусков 35

2.3. Квантильная регрессия 41

Выводы 44

ГЛАВА 3. ОПРЕДЕЛЕНИЕ РЕЖИМА ПРОПУСКОВ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО КОМПЛЕКСА R 45

3.1. Общее определение матрицы-индикатора 45

3.2. Определение режима пропусков с использованием временных рядов 49

3.3. Определение режима пропусков в рамках многомерного анализа 57

Выводы 60

ЗАКЛЮЧЕНИЕ 61

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 63

# **ВВЕДЕНИЕ**

Кредитный риск занимает центральное место в общей системе финансовых рисков банка, оказывая существенное влияние на устойчивость и эффективность его работы, и как следствие - уровень прибыли. Определение кредитного риска во многом зависит от целей, которые преследует банковская организация. В общем случае, кредитный риск можно рассматривать как вероятность неуплаты заемщиком долга или начисленных процентов, то есть вероятность его дефолта.

При оценке кредитного риска используют сформированные определенным образом выборки данных, в которых часто наблюдаются пропуски, возникающих по различным причинам. Наличие пропусков, точно так же как и их удаление, может привести к снижению точности прогнозных значений при оценке кредитного риска, что приводит к увеличению заемщиков, неспособных полностью или частично выплатить полученную ссуду и, как следствие, кредитные организации теряют часть прибыли или имеют искаженные данные для анализа.

В этой связи, возникает необходимость прогнозирования кредитного риска с как можно более высокой точностью, причем затрагивая не только технические стороны вопроса, но и уделяя внимание эвристическим методам, подводящим к правильному выбору модели в условиях неполных данных. Иными словами, необходимо изучить все стороны работы кредитных организаций при анализе рисков, механизмы порождающие пропущенные значения и сформировать набор правил и методов минимизации ошибки прогноза кредитного риска, возникающей в результате наличия пропусков в данных. В ряде случаев классический подход приводит к большой ошибке или выбору неправильного инструментария для дальнейшего прогнозирования.

Методологической основой послужили работы как отечественных, так и зарубежных специалистов по анализу кредитных рисков, исследованиям неполных данных и математической статистике. Данная проблема решается по-разному – многие исследователи склонны к удалению пропусков или замещению их определёнными фиксированными значениями, однако более рациональное решение – восстановление исходной зависимости.

Важно ввести разграничение между работами посвященными только проблеме пропусков и работами, ориентированными на данную проблему в разрезе оценки кредитного риска. Так, имеется обширная литература по анализу пропусков в контексте науки о данных: традиционные методы в данном направлении развивали Д. Рубин, Р. Литтл, Р. Елашоф, П. Эллисон, М. Бокс, В. Хантер, Дж. Дрейпер, А. Демпстер, Б. Эфрон, Д. Маркер, М. Азур, Е. Стюарт. Указанные авторы особое внимание уделяли многомерным методам обработки пропущенных значений и разработке эффективных алгоритмов.

Предметная область не учитывалась авторами классической теории замещения пропусков, следовательно, не был и учтен механизм порождающий пропуски. Среди исследователей в прикладной области можно выделить следующих ученых: С. Чен, Н. Сиддики, Р. Флорес-Лопес, Б. Галлер, Ю. Кэрол. В своих работах авторы приводят апробации существующих методов, их корректировки и особенности применения. Однако никто не рассматривает распределение пропусков в рамках исходных признаков, возможно, с учетом их собственного распределения. Это позволило бы не только усовершенствовать математические методы прогнозирования пропущенных значений, но и сконструировать научно обоснованную модель их классификации.

Нельзя не отметить работы по математической статистике и алгоритмизации ее методов, которые позволили расширить существующий инструментарий. Прежде всего необходимо отметить: Д. Хосмер, С. Лемешов, В. Хардле, Б. Прантнер, С. Бурен, Ю. Прохоров, Г. Харман.

В целом, разработанные методы и модели не учитывали особенности распределения пропусков до настоящего времени. Также не были разработаны алгоритмы для их анализа, которые бы в ситуации неопределённости позволили выбрать корректный подход и, как следствие, сэкономить как временные, так и вычислительные ресурсы. В результате это открывает широкие возможности для научных исследований в данном направлении.

Цель исследования состоит в разработке методов снижения ошибки при оценке кредитного риска в условиях систематических пропусков для улучшения точности предсказания неблагоприятных исходов, вызванных невозможностью погашения заемщиком ссуды. Цель была достигнута путем решения частных научно-практических задач:

* Оценка точности прогнозных значений при различных режимах пропусков, а также в зависимости от распределения исходных величин
* Разработка методов для улучшения точности прогноза за счет преобразования статистических данных
* Разработка прототипа программного обеспечения с апробацией полученных теоретических результатов на реальных выборках
* Разработка управленческих рекомендаций по совершенствованию процесса оценки кредитного риска
* Разработка набора эвристических методов выбора способа прогнозирования пропущенных значений

Объектом исследования являются структурные подразделения коммерческих банков, занимающиеся оценкой кредитных рисков банка при выдаче займов.

Предмет исследования – процесс формирования оценки кредитного риска в условиях неполных данных (систематические пропуски).

# **ГЛАВА 1. ЗАКОНОМЕРНОСТИ И ОСОБЕННОСТИ РАБОТЫ С ПРОПУЩЕННЫМИ ЗНАЧЕНИЯМИ ПРИ ОЦЕНКЕ КРЕДИТНОГО РИСКА**

Кредитный риск понятие достаточно широкое, которое заключается в неисполнении дебитором своих обязательств перед контрагентами. Однако в рамках данного исследования понятие кредитного риска, как уже отмечалось, рассматривается в разрезе банковской сферы. Следовательно, наиболее точное определение регламентируется нормативно-правовыми документами Базельского комитета по банковскому надзору, который определяет кредитный риск, как потенциальную возможность непогашения заемщиком обязательств перед кредитором в оговоренные сроки. Отметим, что под определением понимаются две возможности: непогашение суммы кредита, которая складывается из основной суммы и процентов по кредиту, и погашение обязательств по прошествии срока, указанного в договоре. Уточненное определение дается Центральным Банком России в письме от 23.06.2004 N 70-T «О типичных банковских рисках», в котором кредитный риск определяется, как риск возникновения у кредитной организации убытков по причине неисполнения, либо частичного или несвоевременного исполнения должником финансовых обязательств перед кредитной организацией в соответствии с условиями договора[[1]](#footnote-1).

Таким образом, для моделирования кредитный риск будет определяться как вероятность просрочки или дефолта заемщика. Такая интерпретация достаточно легко позволяет формализовать возникающую ситуацию с помощью вероятностных моделей. Также к оценке риска можно подойди со стороны видов кредитного риска, которые даются организацией «Франклин&Грант. Риск консалтинг». Выделяют базовый уровень – транзакционный риск, который характеризует кредитоспособность отдельно взятого заемщика, второй уровень – портфель рисков, объединяющий совокупность кредитов, на которые оказывают влияние одни и те же факторы, и на последнем уровне рассматривается аллокационный риск, когда банковские риски распределяются по отраслям, регионам или продуктам банковской деятельности. Очевидно, что в рамках предмета исследования, в работе рассматривается транзакционный риск.

Понимание кредитного риска крайне важно при его оценке, однако не менее важную роль играет понимание важности его прогнозирования, что в свою очередь подкрепляет актуальность данного исследования. Банковская деятельность достаточно сложна в управлении, поэтому существует множество регулирующих органов. Каждый банк в отдельности может привести к подрыву отрасли в целом, например, известный «эффект домино», при котором массовое изъятие депозитов или банкротство некоторого банка в случае слухов распространяется на всю систему. Такой эффект может привести к банкротству других банков или финансовых учреждений, связанных с ним, особенно, если речь идет о банке-лидере. В результате массового банкротства по этой причине с 1961 по 1981 в США, Базельский комитет подготовил ряд документов-соглашений, которые дополняются на регулярной основе с учетом меняющихся условий рынка и различных прецедентов. Рассматривая эволюцию, представленную в таблице 1.1, разработанных соглашений, можно заметить, что внутренние методы на текущий момент наиболее актуальны и их совершенствование - это неотъемлемая часть для успешного функционирования финансово-кредитной организации[[2]](#footnote-2).

В рамках данных документов предлагается два подхода к оценке кредитного риска. Первый – стандартизированный подход, который базируется на внешних оценках рейтинговых агентств, которые, оперируя качественными и количественными данными, производят комплексную оценку риска, а затем выстраивают классы (группы), и оценка вероятности дефолта заемщика сводится к отнесению его к определенному классу по определённым характеристикам. Хотя такой подход и минимизирует издержки банков при оценке рисков, он обладает рядом недостатков: не учитываются особенности клиентской базы конкретного банка, региональная специфичность.

*Таблица 1.1.*

**Сравнительная характеристика соглашений Базель I, II, III\***

|  |  |  |  |
| --- | --- | --- | --- |
| Положение | Методология | Требования к капиталу | Риски |
| Базель I (1988 г.) | Базельским комитетом жестко установлены методологические подходы к определению и расчету капитала.  🞄Ориентация на количественные характеристики | Дифференциация коэффициентов достаточности капитала осуществляется только в зависимости от кредитного рейтинга страны | Покрытие только кредитного риска  Использование для оценки рисков только стандартизированного подхода, установленного Базельским комитетом |
| Базель II (2004 г.) | Позволяет использование внутренних методик определения рисков | Дифференциация коэффициентов достаточности | Покрытие кредитного, рыночного и операционного рисков. |

*Продолжение таблицы 1.1.*

**Сравнительная характеристика соглашений Базель I, II, III**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 🞄Ориентация на качественные показатели.  К количественным требованием добавляется вторая и третья компоненты – надзорный процесс и рыночная дисциплина | капитала происходит в зависимости от кредитного риска каждого заемщика.  🞄Повышение чувствительности к рискам | Возможность выбора между внутренними и внешними рейтинговыми системами оценки кредитного риска. |
| Базель III (2010 г.) | Возможность применения внутренних банковских методик при оценке кредитных рисков сохраняется.  🞄Введение нормативов, регулирующих выплаты собственникам и управляющим, в зависимости от их выполнения.  🞄Внедрение требований, связанных с организацией банковского надзора за соблюдением нормативов достаточности капитала | Измерение структуры собственного капитала банков.  Повышение требований к достаточности капитала.  Создание резервного запаса.  Вводится механизм контрциклического регулирования | Покрытие кредитного, рыночного и операционного рисков  Возможность выбора между внутренними и внешними рейтинговыми системами оценки кредитного риска.  Возможность привлечения зависимых рейтинговых агентств для оценки рисков |

\*Составлено по: Диверсификация модельного риска в скоринговых моделях //

ОАО Банк ЗЕНИТ. Презентация, с. 1-5

Также не разрабатываются внутренние инструменты аналитики, которые в условиях конкуренции играют важнейшую тактическую роль. Примером может послужить случай быстроменяющейся конъюнктуры на рынке, в результате чего модели необходимо корректировать в реальном времени с учетом поступающих данных, и ожидание внешних прогнозов приводит лишь к дестабилизации организации. Как следствие, активно растет второй подход – IRB.

Его обычно делят на две части – базовый и продвинутый. Первый позволяет банкам разрабатывать и применять модели только для оценки вероятности дефолта заемщика, а основные составляющие риска предлагаются надзорными органами. Второй подход же подразумевает применение разработанных моделей и к оценке драйверов риска таких как EAD или LGD. Таким образом, основной задачей банка в рамках IRB подхода является разработка и применение качественной модели для оценки вероятности дефолта заемщика на основе фундаментальных данных (обычно полученных из корпоративных информационных систем) в рамках подхода, основанного на кредитном скоринге[[3]](#footnote-3).

Учитывая, что большинство моделей опираются на статистические данные, возникает острая проблема – появление пропущенных значений, которые снижают как точность прогнозных значений, так и количество пригодных моделей для оценки кредитного риска. Следовательно, в рамках данного исследования важно расширить список инструментов, подходов и методов, позволяющих обоснованно выбрать модель для прогнозирования и улучшить точность прогнозов в условиях неполных данных. Необходимо отметить, что при прогнозировании кредитного риска физических лиц, на основе которого базируется данное исследование, никаких институциональных ограничений нет и банки имеют право самостоятельно выбирать и разрабатывать методики, что в свою очередь существенно упрощает исследования, однако разрабатываемый подход может быть обобщен на более широкий круг задач, предусмотренных положениями Базельского комитета.

# **Основные подходы к оценке кредитного риска при неполных данных**

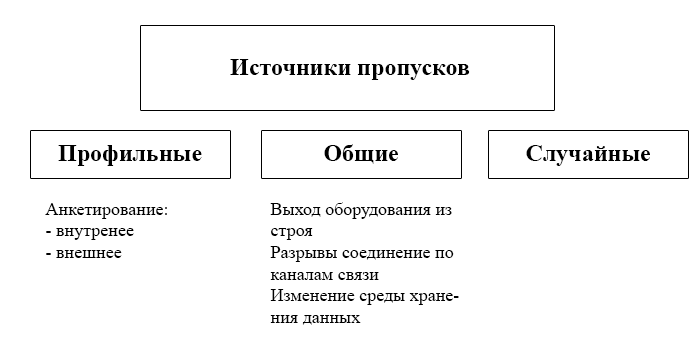
Проблема пропущенных значений всегда была актуальной, однако широкое развитие данное направление получило относительно недавно, и все глубже проникает в различные отрасли. В данном исследовании выделяются две основные причины развития: первая – это рост количества данных и, как следствие, появление возможности применения методов прогнозирования пропущенных значений, вторая – это компании, принимающие решения на основе данных. Последнюю причину можно считать весомей. И для того, чтобы понять почему, рассмотрим само понятие «на основе данных». Управление на основе данных – процесс формирования инструментов, способностей и корпоративной культуры, которая опирается в первую очередь на данные[[4]](#footnote-4). Такой тип компаний – это одно из крупнейших достижений в корпоративной среде за последние двадцать лет. На текущий момент такой переход могут себе позволить лишь немногие крупные корпорации, поскольку это требует как серьезных финансовых ресурсов, так и кадровых.

Компании такого типа выстраивают весь процесс принятия решений с помощью данных, поэтому для них крайне важны качество данных и их полнота. Достаточно вспомнить известное выражение Томсона Нгуена: “Качественные данные всегда выигрывают у качественных моделей”. К тому же качество и полнота существенно расширяют список моделей для применения. Конечно, легко понять, что одни из первых компаний, которые перешли на такой тип управления – это банки и ритейл, поскольку именно они обладают всеми необходимыми ресурсами в целом[[5]](#footnote-5).

В результате выстраивается следующая цепочка работы с данными (часто называют слои обработки данных): сбор сырых данных, их обработка или трансформация, расчет метрик и агрегация, а также заключительный этап - построение моделей и хранение вычисленных значений. Второй этап и четвертый являются ключевыми для данного исследования, так как на втором мы обогащаем или трансформируем данные путем сложной работы с пропущенными значениями, а на последнем этапе – изучаем влияние на модель и ее качество с учетом работы на втором слое. Объединяя вышесказанное, получаем предмет данного исследования, который заключается в формирования оценки кредитного риска заемщика в условиях неполных данных (систематические пропуски).

Прежде чем перейти к обсуждению существующих взглядов на проблему, рассмотрим базовые понятия и особенности работы с пропусками. Один из главных вопросов, определяющих подходы к работе с пропущенными значениями – это природа их происхождения и важность признаков, с которыми мы перейдем непосредственно к анализу. В случае банковского сектора – широко распространена практика анкетирования, причем она складывается как из внешнего анкетирования, так и внутреннего. Так, рассмотрим оба варианта: внешнее анкетирование проводится в виде опросов. Это могут быть опросы посредством электронных рассылок или личного анкетирования клиентов. Такие данные крайне важны для организации, так как по данной выборке обычно приводят суждения о всей совокупности своих клиентов и, более того, проводят мэтчинг со всей базой данных, поэтому целостность таких данных крайне важна. Однако участники анкетирования могут забыть ответить на какие-то вопросы или проигнорировать их в силу щекотливости. В результате появляются многочисленные пропуски. Увеличить выборку обычно невозможно, так как бюджет на исследования ограничен, здесь и возникает момент, когда важно с высокой точностью спрогнозировать пропуски. Второй вариант – это внутреннее анкетирование, которое проходят все клиенты банка, однако по-прежнему необязательные с правовой точки зрения поля люди склонны не заполнять, хотя для банка эти данные могут быть важны. Помимо этих двух основных источников пропусков существуют и другие, например, оборудование для записи вышло из строя или исчезло подключение к Интернету, и часть данных была утеряна.

Важно рассмотреть классификацию источников в рамках данного исследования, которая приведена на рисунке 1.1. Источники можно разделить на три типа: профильные, которые характерны для данной отрасли, общие и случайные, которые заранее вовсе можно не учесть. В результате всегда нужно иметь серьезные инструменты аналитики для работы с пропущенными значениями. Второй, не менее важный моменты работы с пропусками, – это механизмы их формирования.

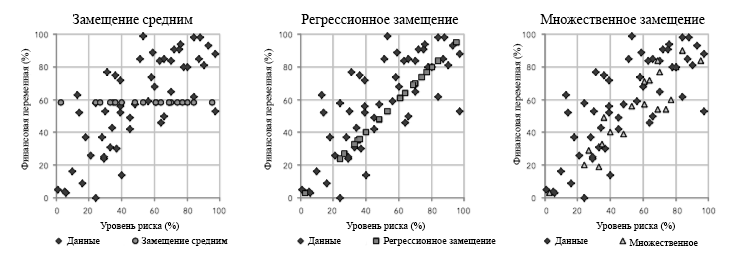


*Рис.1.1.* **Источники возникновения пропущенных значений**

В классической литературе выделяют 3 механизма: MCAR, MAR и MNAR. Первый тип MCAR (Missing Completely At Random), или полностью случайные пропуски, - механизм формирования пропущенных значений, при котором одинаковая вероятность получения пропуска для каждого объекта. В качестве примера можно привести социологический опрос, где каждому десятому участнику не задавали один случайный вопрос из списка, но на все остальные вопросы ответы были получены. В такой ситуации речь идет о механизме MCAR и, например, игнорирование или исключение пропущенных значений не приведет к искажению результатов. Второй тип – это MAR (Missing At Random), или случайные пропуски, обычно с практической точки зрения такие пропуски хоть и случайны, но формируются в результате некоторых закономерностей. Говорят, что пропуски формируются механизмом MAR, если вероятность пропуска может быть определена путем анализа других признаков в наборе данных, например, пол, возраст, образование, доход. Причем в этих признаках пропуски не содержатся. В таком механизме удаление также подходит или замена тривиальным значением. Последний механизм – это MNAR (Missing Not At Random), или пропуски не случайны, при котором данные отсутствуют в зависимости от неизвестных факторов. Такой механизм предполагает, что данные могли быть описаны на основе других признаков, однако информация по ним в наборе данных отсутствует. При таком механизме исключение невозможно, поскольку это может приводить к искажению статистических свойств выборки[[6]](#footnote-6).

Исследование пропусков получает все большую важность и тесно коррелирует с развитием озера данных (программно-аппаратный комплекс различных массивов данных, объединенных в единую логическую среду), в рамках которого особое внимание уделяется качеству данных, где и происходит работа с пропущенными значениями. Именно банковский сектор, как уже отмечалось, первым пришел к таким решениям. Однако важность разграничения работ играет особую роль, поскольку общая теория работы с пропусками никак не учитывает особенностей прикладной области. Большой вклад в развитие фундаментальной теории сделали Д. Рубин, Р. Литтл, Р. Елашоф, П. Эллисон, М. Бокс, В. Хантер, Дж. Дрейпер, А. Демпстер, Б. Эфрон, Д. Маркер, М. Азур, Е. Стюарт. В своих работах они приводили различные классификации методов, измерили эффективность алгоритмов, выделяли типы пропусков. Их теории изложены и адаптированы в следующем параграфе. В рамках данного параграфа необходимо сфокусироваться на предметной области.

Важная работа известного риск-менеджера С. Чена служит отправной точкой для исследований и является одной из первых по пропускам при оценке кредитного риска[[7]](#footnote-7). С. Чен рассматривает пропущенные значения как фундаментальную проблему банка, которую банк должен держать на особом контроле, поскольку каждая переменная или наблюдение может оказать существенное влияние на прогнозирование риска. Важно определить характер пропусков, о чем уже шла речь в начале данного параграфа. Он выделяет первый этап работы с пропусками – полный кейс-анализ. Тот случай, когда банк не готов применять сложные математические модели, или сложилась ситуация, когда аналитические инструменты не пригодны. Здесь нужно ранжировать данные или признаки по степени важности или пригодности. Например, корпоративная банковская система может собирать данные со всех регионов, однако аналитик при таком анализе способен убрать часть заемщиков, не относящихся к географическому региону, для которого строится модель прогнозирования. Второй подход, выделяемый С. Ченом, называется доступный кейс-анализ. В данному случае он связан с первым, однако вместо моментального удаления пропущенных значений их отмечают специальным значением, и ищут способы обогащения данных. На практике часто используется гибридная модель из полного и доступного кейс-анализа. Третий подход, на который С. Чен обращает особое внимание, – это замещение пропусков. В банковской сфере широкое распространение получили методы замещения, основанные на центральных характеристиках распределения – среднее, мода и медиана. Также особое место занимает регрессионное замещение. Однако автор отмечает, что данные методы используются в традиционном виде (не адаптировано). В дальнейшем банки планируют внедрять множественное замещение, которое в некоторых случаях является наиболее выигрышным вариантом на данный момент. Результаты всех трех ситуаций продемонстрированы на рисунке 1.2. На первом применено элементарное замещение средним значением, а на втором с учетом тренда или регрессионной зависимости, а на третьем – множественное замещение. Однако важно заметить, что здесь очевидная зависимость от структуры данных и распределения пропусков, что и предлагается автором данного исследования учитывать при работе с пропущенными значениями. Легко смоделировать ситуацию, когда базовые центральные характеристики или регрессионные модели будут превосходить по качеству множественное замещение.



*Рис. 1.2.* **Анализ различных подходов к замещению пропусков**

В заключении автор приводит методы работы с неадекватными данными, под которыми понимаются малые выборки небольших банковских организаций. Здесь рекомендуется применение экспертных методов, в которых, вообще говоря, также можно учесть особенности распределения пропущенных значений.

В работах Р. Флорес-Лопес обсуждается подход, который в большей степени основан на объеме пропущенных значений, то есть ставится вопрос о допустимом проценте пропущенных значений[[8]](#footnote-8). В зависимости от этого рассматривается подход к оценке кредитного риска. На первом этапе, в случае совсем некорректных данных автор предлагает вовсе не рассматривать прогнозирование математическими методами, поскольку результаты могут слишком быть далеки от истинных. Если в выборке присутствует не больше 20% - можно пренебречь и удалить данные, содержащие пропуски. Особенно важно, что автор обращает внимание на необходимость учета природы пропусков при использовании математических и инструментальных методов прогнозирования кредитного риска. К тому же автор отмечает причины возникновения, специфичные для банка: неполнота данных в анкетах, умышленные пропуски, пропуски в результате агрегации данных и потеря данных в корпоративных системах. Работы Р. Флорес-Лопес достаточно тесно пересекаются с работами Н. Сиддики[[9]](#footnote-9).

В работах Б. Галлер, Ю. Кэрол фокус смещен на анализ механизмов возникновения пропусков, основываясь на работах Д. Рубина, которые обсуждались в данном параграфе. После чего предлагается перейти к выбору метода замещения[[10]](#footnote-10). Для более детального анализа степени разработанности проблемы рассмотрим таблицу 1.2, где изложены основные подходы к решению проблемы в рамках оценки кредитного риска.

*Таблица 1.2.*

**Основные подходы к решению проблемы замещения пропусков**

|  |  |  |
| --- | --- | --- |
| Авторы | Прикладной уровень исследований | Методы и подходы |
| Д. Рубин, Р. Литтл, Р. Елашоф, П. Эллисон, М. Бокс, В. Хантер, Дж. Дрейпер, А. | Нет | Многомерный анализ |
| С. Чен | Да | Продвинутые замещения |
| Работы Р. Флорес-Лопес | Да | Базовое замещение |
| Б. Галлер, Ю. Кэрол | Да | Механизмы возникновения |

# **1.2. Математические и инструментальные методы оценки кредитного риска**

Одна из основных задач данного параграфа заключается обсуждении и приведении аргументов в пользу выбранных методов, которые являются основой для построения нового подхода. Кроме того, отмечаются особенности их применения в банковском секторе с целью сфокусироваться не только на технической части вопроса, но и на практической значимости.

Математические и инструментальные методы в данной работе необходимо разделить на два класса: первый – традиционные методы прогнозирования кредитного риска, используемые соответствующими структурными подразделениями банков, второй – продвинутые методы замещения пропусков. Первый класс методов выбран таким образом, чтобы максимально точно оценить эффективность применения нового подхода при оценке кредитного риска, поэтому и отдано предпочтение классическим моделям, поскольку двойная модификация методов в конечном итоге может сгладить полученный эффект. Как правило, среди самых распространённых выделяют такой класс статистических моделей как обобщенные линейные модели. В частности, для установления вероятности банкротства используются модели с дихотомической зависимой переменной следующего вида:

(1.1)

где , - это векторы-столбцы значений объясняющих переменных и коэффициентов регрессии. Также для модели необходимо осуществить выбор функции , которая обычно задается сигмоидом, см. формулу 1.2, или функцией, основанной на нормальном распределении, обычно под такой функцией подразумевают обратную к интегральной функции стандартного нормального распределения, то есть .

(1.2)

Помимо приведенных функций, существует множество других, которые были предложены для моделирования дихотомической зависимой переменной, однако Кокс и Шел указали две основные причины в пользу логистического распределения[[11]](#footnote-11). Первая, с точки зрения математики, модель является крайне гибкой и простой в использовании, вторая причина – достаточно легко интерпретировать полученные результаты. К тому же можно интерпретировать не только выходные значения, но и оцененные параметры (например, рассмотреть их как предельные эффекты по аналогии с классической линейной регрессией). Важно заметить, что модель хорошо себя зарекомендовала для задачи классификации – объект x можно отнести к классу , если предсказанная моделью вероятность по формуле 1.1 превышает пороговое значение 0.5, и к классу в противном случае. Это особенно важно в банковском секторе, поскольку всегда известны метки классов для обучающей выборки.

В заключение обсуждения о логит-модели необходимо рассмотреть вопрос регуляризации. В прикладных исследованиях часто возникает ситуация, когда полученная модель отлично работает на примерах из обучающей выборки, но на новых данных такой же природы работает относительно плохо, такой эффект называется переобучением, и для его уменьшения рассматривают регрессии с регуляризацией[[12]](#footnote-12). Суть регуляризации состоит в том, что вектор параметров рассматривается как случайный вектор с априорной плотностью распределения . В результате меняется метод для обучения модели – вместо наибольшего правдоподобия используется его модификация, когда вводится некоторое априорное распределение исходной величины, - метод максимизации апостериорной оценки. Следовательно, необходимо найти такие параметры , которые бы максимизировали следующий функционал:

(1.3)

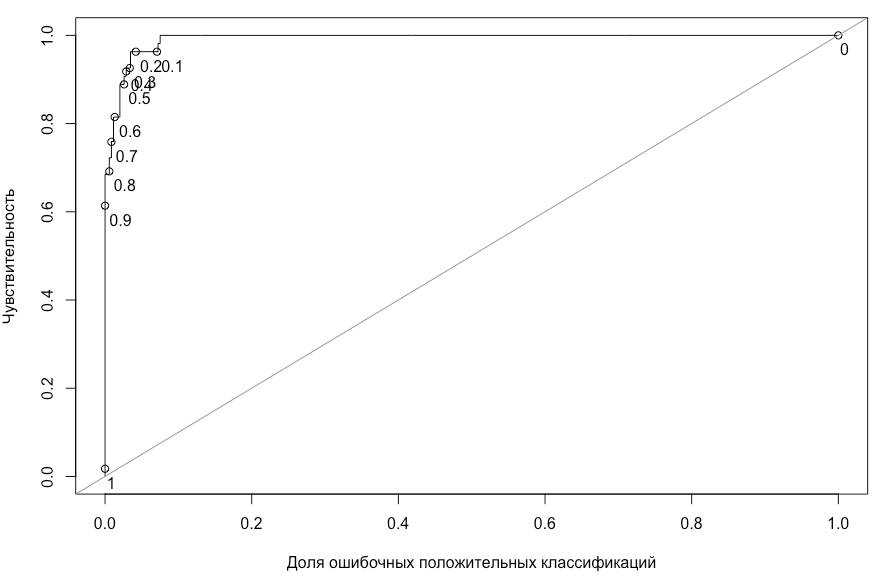
Обычно априорным распределением выступает многомерное нормальное распределение с нулевым средним и ковариационной матрицей . Также вводится предпосылка о том, что коэффициенты регрессии, не имеющие статистической значимости равны нулю. Подставив плотность априорного распределения в формулу 1.3 и прологарифмировав, получим следующую оптимизационную задачу:

(1.4)

где – параметры регуляризации. В целом, заметим, что к исходной целевой функции 1.3 L2-норму вектора параметров для регуляризации. Аналогично легко записать L1-норму, в которой слагаемое, отвечающее за штраф целевой функции, вместо возведения в квадрат берется по модулю:

(1.5)

В исследовании для оценки качества модели бинарной классификации широко используется ROC-кривые, пришедшие из теории обработки сигналов[[13]](#footnote-13). Согласно результатам Д. Хосмера и С. Лемешова кривая дает достаточно точное представление об обучении модели и позволяет проводить их сравнительный анализ[[14]](#footnote-14). Рассматриваемая кривая показывает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак при варьировании порога решающего правила. Первую ситуацию, как правило, называют чувствительностью алгоритма классификации, вторую – специфичностью. На рисунке 1.3 приведен пример ROC-кривой.



*Рис.1.3.* **ROC-кривая на примере искусственных данных в среде R**

Очевидно, что недостаточно изобразить на графике кривые, необходимо также дать их количественную интерпретацию. Для этого используется показатель AUC (площадь под кривой), рассчитанное значение которого сравнивается с нормативными значениями. В таблице 1.3 приведены правила, разработанные и рекомендуемые Д. Хосмером и С. Лемешовым, которые используются в данном исследовании[[15]](#footnote-15).

*Таблица 1.3.*

**Общие правила количественной оценки ROC-кривой**

|  |  |
| --- | --- |
| Правило | Описание |
| Если значение ROC = 0.5: | непригодность классификации |
| Если | приемлемая классификация |
| Если | отличная классификация |
| Если | превосходная классификация |

# **1.3. Методы работы с пропущенными значениями и их анализ**

Как уже обсуждалось в начале Главы 1 тема пропусков в кредитном скоринге особенно важную роль играет, однако, по мнению Р. Кабакова, ей не уделяется должного внимания. Причем не только в области кредитования, но и в учебных пособиях по статистике обсуждение данной проблемы может сводиться к нескольким абзацам[[16]](#footnote-16). Даже если учесть, что при анализе данных, работа с пропусками является неотъемлемой его частью, очень редко происходит обсуждение работы в условиях неполных данных. Возникает некоторое противоречие: с одной стороны, влияние пропущенных значений – это то, чего каждый бы исследователь хотел избежать, с другой стороны, обсуждение проблематики за пределами специализированных книг мизерно. Поэтому одной и задач данного исследования является восполнения пробела в анализе пропусков для оценки кредитного риска.

В данном параграфе рассматриваются методы прогнозирования пропусков, которые согласуются с предложенным автором исследования вероятностным подходом их моделирования. Помимо методов уделено внимание обнаружению, этапам, общим схемам работы с пропущенными значениями, которые в дальнейшем будут модифицированы с учетом особенностей нового подхода.

## **1.3.1. Базовые методы обработки пропущенных значений**

Базовые методы, несмотря на свою простоту, в определённых ситуациях обладают высокой точностью и при моделировании оценки по-прежнему будут обладать хорошими статистическими свойствами. Первый и самый тривиальный метод – это игнорирование объектов с пропущенными значениями или, как альтернатива, - удаление. Обычно такой подход хорошо работает, если количество пропусков незначительно и их прогнозирование вовсе может усугубить ситуацию, искажая данные. В целом необходимо обратить внимание, что при моделировании данные представлены матрицей объект-признак , где n – количество объектов, а d – количество признаков. Удаление может осуществляться по любому из измерений, поэтому важно ранжировать признаки или объекты. Возможно, некоторые признаки не имеют практической или статистической значимости, и их удаление вовсе не окажет влияние на итоговую модель. Тем более, как будет показано в Главе 2, скоринговые модели устойчивы в таких ситуациях.

Следующая группа методов заключается в замене пропусков специальным значением. Это группа особенно важна для разрабатываемого вероятностного подхода в данном исследовании, поскольку в качестве специального значения могут выступать параметры распределения. Установив распределение пропусков можно однозначно определить такое значение. Среди таких значений выделяют моду, медиану или среднее арифметическое. Однако признаки могут быть не только количественными, но и категориальными, поэтому, как правило, мода характерна для второго типа, остальные специальные значения для первого.

## **1.3.2. Продвинутые методы обработки пропущенных значений**

Продвинутые методы обработки нацелены прежде всего на прогнозирование пропущенных значений или сложную обработку для выявления структуры, особенностей распределения и причин появления пропусков. Затем возможно применение какого-либо метода, включая тривиальные. Рассмотрим сначала замену с помощью сингулярного разложения, которое используется для приближения матрицы матрицей меньшего ранга[[17]](#footnote-17). Сначала найдем сингулярное разложение матрицы:

, (1.6)

где , - унитарные матрицы, - диагональная матрица с сингулярными числами. Затем необходимо получить восстановленную матрицу:

, (1.7)

где получена из занулением всех диагональных элементов кроме k наибольших. То есть сначала сокращается размерность матрицы с помощью сингулярного разложения, а затем пропуски заменяются из восстановленной матрицы. Заметим, что для начальной инициализации матрицы по признакам используются средние значения.

Выполнение замены с помощью метода ближайших соседей также хорошо сочетается с новым подходом, поскольку метод исходит из предпосылки, что близкие объекты по значениям среди рассматриваемых признаков близки в признаках, значение которых может быть пропущено (гипотеза компактности). Недостаток метода заключается в наличии достаточного количества данных для восстановления.

Замена пропусков с помощью EM-алгоритма основывается на знании о том, какое распределение имеет исходная выборка[[18]](#footnote-18). В начальной постановке используется EM-алгоритм для отыскания параметров смеси нормальных распределений:

(1.8)

Зная средние значения и набор ковариационных матриц , вычисляются параметры регрессии:

, (1.9)

где – признак, который необходимо восстановить, – матрица с признаками за исключением восстанавливаемого. В алгоритме также предполагается, что среднее значение по каждому признаку X равно нулю. По оцененным параметрам из формулы 1.9 вычисляются пропущенные значения по другим признакам этого объекта:

(1.10)

Однако итоговое прогнозируемое значение необходимо усреднить по распределениям смеси следующим образом:

(1.11)

Начальная инициализация, как и в предыдущем алгоритме выполняется по средним значениям каждого признака. Затем итеративно происходит уточнение параметров распределения и пропущенных значений до сходимости или же достижения заданного числа итераций.

Из российских ученых в области анализа данных наиболее значимые труды Загоруйко Н. Г. В частности, рассмотрим алгоритм ZET, в основе которого лежат следующие предпосылки: избыточность данных, линейная зависимость и локальная компактность[[19]](#footnote-19). Прогнозирование пропусков происходит следующим образом: в предварительно нормализованных данных вычисляют компетентности всех строк, не имеющих пропуска в том же признаке, со строкой, имеющей пропуск:

(1.12)

где - количество непропущенных пар значений в строках i и y, – декартово расстояние между строками. Затем выполняется аналогичная операция для столбцов и объектов:

(1.13)

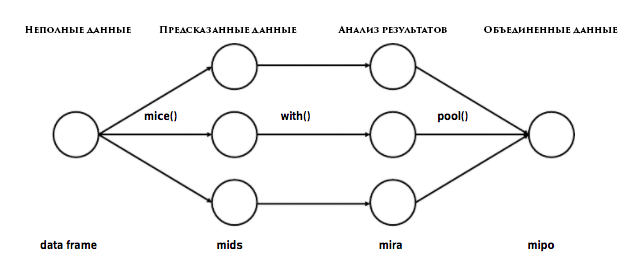
где – корреляция между столбцами, – декартово расстояние между строками. После чего среди всех строк и столбцов выбирается заданное количество строк и столбцов с наибольшей компетентностью. Далее по полученными результатам подбирается специальное значение с целью учета компетентности во взвешенной сумме отдельно для строк и столбцов, минимизируя следующий функционал:

Предсказания рассчитываются по следующим формулам:

(1.14)

где – прогноз для значений строки (столбца) k с помощью j строки (столбца) линейной регрессии вида y = ax + b. Предсказанные значение по строкам и столбцам усредняются. Данная процедура выполняется для каждого пропущенного значения.

Заключительный класс методов является одним из самых продвинутых и часто наиболее эффективным. Обратимся к многомерной оценке цепными уравнениями. Данный метод практически не освещен в отечественных источниках и его разбор будет предложен впервые в рамках данного исследования. Обычно, когда речь идет о данном подходе, приводится набор утилит MICE, реализованный учеными из университета Твенте, в котором предложен набор моделей (уравнений) и функций для генерации данных. То есть алгоритм состоит из двух основных шагов: построения модели и генерации данных, и нескольких промежуточных. Важно заметить, что алгоритм предназначен для матриц данных с большим количеством объектов и признаков. Рассмотрим схему на рисунке 1.4, где показан весь процесс замещения пропущенных значений.



*Рис.1.4.* **Основные этапы при многомерном замещении**

На рисунке 1.4 можно видеть три основных шага – предсказание пропущенных значений, их анализ и объединение. Для более точных прогнозов выделяют три класса хранений: mids, mira и mipo[[20]](#footnote-20). Анализ начинается с левой части рисунка, где на вход подаётся набор неполных данных . На начальном этапе происходит замещение некоторыми базовыми значениями, которые берутся из определённого распределения. Данная процедура выполняется функцией mice и значения хранятся в классе mids. Кроме того, данные разделяются на некоторое количество наборов, объекты которых сходны между собой. Далее применяется функция (из некоторого набора, например, класс обобщенных линейных моделей) для уточнения значений до тех пор, пока их характеристики не будут схожи с наблюдаемыми объектами. На последнем шаге вычисленные оценки объединяют в один усредненный вектор и вычисляют дисперсию. Далее на основе полученных оценок применяют метод выбросов Рубина и его правила для объединения полученных оценок в конечную матрицу объект-признак[[21]](#footnote-21).

На текущий момент в рамках подхода MICE реализованы и исследованы свойства следующих методов (в зависимости от типа переменных): для любого типа - предсказывающее среднее обычное и взвешенное, случайные выборки, решающие деревья, случайный лес, для численных – безусловное среднее, линейная регрессия Байеса, линейная регрессия с бутстреп подходом, метод случайных индикаторов, для бинарных переменных – логистическая регрессия и бутстрап логистическая регрессия, для порядковых переменных - модель пропорционального отношения шансов, для номинальных переменных – полиноминальная логистическая регрессия и линейный дискриминантный анализ. Можно заметить, что в рамках данного подхода используется широкий набор вероятностных моделей, которые также могут быть выбраны обоснованно с учетом аппарата, предложенного в данном исследовании[[22]](#footnote-22).

## **1.3.3.** **Программное обеспечение и особенности базисной выборки**

Одной из основополагающих частей данного исследования является проведение многочисленных экспериментов и построение ряда моделей как на примере искусственных данных, так и на реальных выборках. Для этих целей необходимо обоснование выбранного программного комплекса, который позволит реализовать поставленные задачи. Кроме того, в данном подпараграфе обсуждается полученная выборка, объем которой также требует особенных механизмов обработки.

Важным элементом при работе с кредитным риском и апробации полученных результатов является наличие достаточного объема данных для построения и тестирования моделей, поскольку такие данные, как правило, являются коммерческой тайной, следовательно, - труднодоступны. Поэтому в качестве базы для построения скоринговой модели были использованы данные, опубликованные американской платформой для прогнозного моделирования и аналитических соревнований Kaggle, образованной в 2010 году[[23]](#footnote-23).

Компании размещают на данном ресурсе реальные данные, а исследователи со всего мира соревнуются между собой в разработке лучших моделей. Одно из таких соревнований, проводимых Kaggle, было направлено на повышение навыков построения скоринговых моделей, позволяющих предсказывать вероятность того, что заемщик может испытать финансовые трудности в ближайшие два года, что приведет к его дефолту. В качестве исторических данных, была предоставлена информация по 250 000 заемщикам. Причем данные представлены двумя выборками – для тестирования модели и для ее построения. Именно данная выборка служит основой для апробации результатов исследования. Важно заметить, что в данной выборке пропуски получены естественным путем, а не искусственно сгенерированы, что повышает ценность полученных результатов.

Разведывательный анализ представлен в таблице 1.4, в которой описаны выборка для обучения, тестовая выборка, количество пропущенных значений и эталонная выборка, то есть выборка, для которой известны метки классов. В качестве зависимой переменной выступает бинарная переменная Y, характеризующая факт просрочки по кредиту от 90 дней и более, что подразумевает наступление дефолта заемщика. Факторные переменные включают возраст, доход, соотношение долга по револьверному кредиту к кредитному лимиту, количество раз, когда заемщик допускал просрочку по кредиту от 30 до 59 дней за последние 2 года, коэффициент задолженности, равный отношению суммы ежемесячных выплат по задолженности, алиментов и расходов на жизнь к ежемесячным доходам и количество открытых кредитов (рассрочка, автокредит, ипотека) и кредитных линий (кредитных карт).

*Таблица 1.4.*

**Данные для построения и тестирования статистических моделей**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Назначение выборки | | |
| Описание | Выборка для обучения модели | Выборка для тестирования | Эталонная выборка с вероятностями |
| Количество наблюдений | 150 000 | 101 503 | 101 503 |
| Количество  пропусков | 31 655 | 22 729 | 0 |

Важную роль при проведении исследований играет программное обеспечение, на котором планируется выполнение расчетов, поскольку часто какой-либо подход невозможно реализовать в силу ограниченности выбранной среды. В качестве основного ПО в работе используется язык программирования R, который предназначен для статистической обработки данных и является одним из самых продвинутых инструментов для моделирования[[24]](#footnote-24). Выделим основные его преимущества: работа с векторами и матрицами, высокая скорость обработки больших массивов данных, уникальные графические возможности, которые играют важную роль в моделировании пропущенных значений, широкий набор методов для имитационного моделирования, а также простота в создании новых статистических процедур. Последняя особенность обеспечивает высокий потенциал расширения возможностей и включения самых новых методик анализа данных. В совокупности это открывает широкие возможности для данного исследования[[25]](#footnote-25).

# **Выводы**

В первой главе выполнен системный анализ как отечественных, так и зарубежных работ по рассматриваемой проблематике для формирования методологической базы. Результаты структурированы и четко введено разграничение между применяемыми на практике подходами к замещению пропусков и подходом автора данной работы. Приведены существующие классификации источников пропусков в банковском секторе и разработаны новые, которые позволяют ускорить процесс выбора модели в условиях неопределённости. Раскрыта сущность процессов при оценке кредитного риска в банковской сфере, что создает базис для формирования теория замены пропущенных значений в рамках выделенной предметной области. Приведены передовые теоретические модели для работы с пропусками, особенно важно отметить применение модели цепных уравнений, которая ранее не была освещена в отечественных источниках. Особый акцент сделан также на подобранных реальных данных с ресурса Kaggle для апробации нового подхода с увязкой с современным программным обеспечением – статистической средой R.

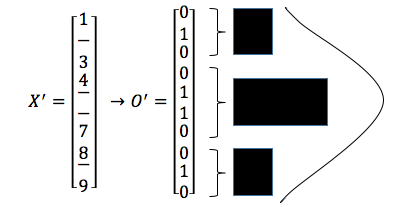
# **2. ПРОГНОЗИРОВАНИЕ ИСХОДОВ ПРИ РАЗЛИЧНЫХ РЕЖИМАХ ПРОПУСКОВ**

Изучение распределения случайной величины занимает центральное место в моделировании, основанном на вероятностном подходе. Такое знание позволяет строить новые модели, обладающие более высокой точностью или же изучать поведение старых при различных распределениях. Таким образом, цель данной главы состоит в изучении условной схемы совместного распределения пропусков и исходной величины на примере одномерного анализа. Под одномерным анализом понимается использование алгоритмов, применимых только для одной исходной величины. Важно отметить, что одномерный анализ носит исключительно теоретический характер для пространственных данных, которыми и оперируют скоринговые модели, однако данные выкладки позволяют в дальнейшем сделать обобщение на многомерный случай, либо одномерную ситуацию, когда известны дополнительные факторы, дающие возможность упорядочивания исходной величины.

В рамках данной и следующей глав активно используются методы имитационного моделирования в программной среде R. Набор искусственных данных получен из источника GitHub[[26]](#footnote-26), поскольку представляет удачно сгенерированные ряды для численных экспериментов в контексте моделирования рисковых ситуаций. Набор данных, состоящий из 2000 наблюдений, включает такие факторы как доход индивида, возраст и размер кредита. Важно обратить внимание, что все величины имеют равномерное распределение, что позволяет достаточно легко создать генератор, который на основе равномерного распределения будет генерировать случайные величины другого типа.

Необходимо определить ключевые понятия для данной главы – что есть распределение пропусков и что понимается под совместным распределением. Предположим, что есть некоторый вектор X размерности , например, характеризующий возраст заемщиков. Для упрощения ситуации обозначим некоторые предпосылки:

1. равномерно распределены на отрезке [a, b].
2. Пропуски полностью случайные, то есть не зависят от значений других факторов – наблюдаемых или ненаблюдаемых:
3. Распределение пропусков известно и задается матрицей-индикатором пропущенных значений, в данном случае – вектор-строкой:



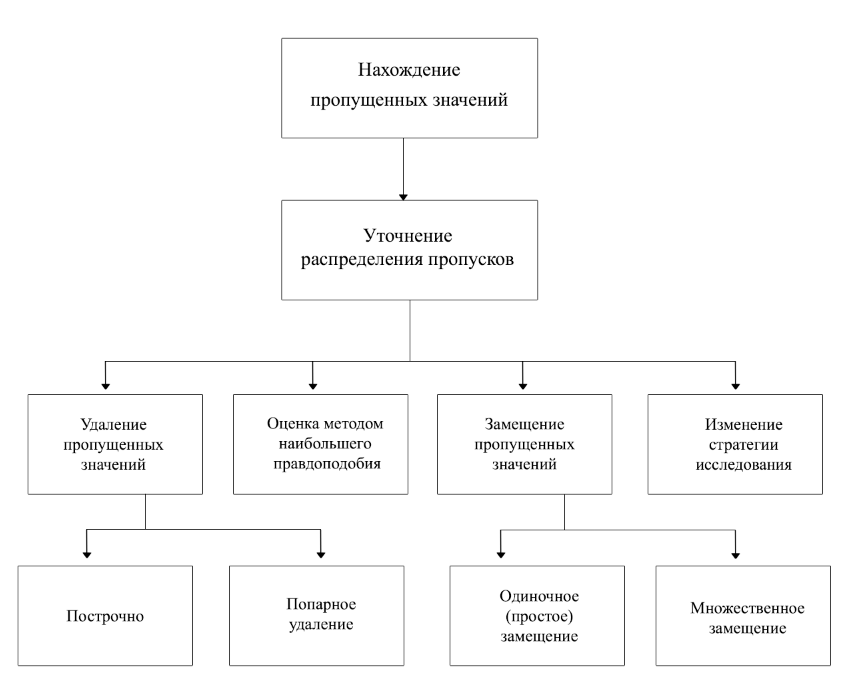
*Рис.2.1.* **Моделирование режима пропусков на примере одномерного вектора**

Более общее определение матрицы-индикатора будет дано в Главе 3, однако можно заметить, что данная матрица получена из исходной путем замены пропущенных значений единицами, а существующих – нулями. Следовательно, подобное определение позволит идентифицировать эмпирическое распределение. В случае пространственных данных на практике практически невозможно определить распределение пропусков без дополнительной информации – других факторов или изучения процесса их появления. Однако такой подход может быть крайне полезен в случае временных рядов, где данные упорядочены естественным образом или панельных данных – частичное упорядочивание.

Рассмотрим рисунок 2.2, на котором изображена гистограмма распределения самой случайной величины – белые столбцы, и серым показано распределение пропусков. Внутри каждого интервала показано отношение между пропущенными значениями и существующими. Например, рассмотрим интервалы , в которых около 30% составляют пропущенные значения; учитывая, что пропуски распределены нормально, то, вообще говоря, значения, предсказанные на хвостах будут точнее, чем центральные в силу большего количества данных для интерполяции. Следовательно, даже получение приближенных распределений пропусков и случайной величины позволяет построить или изменить дальнейшую стратегию развития исследования.

*Рис.2.2.* **Совместное распределение пропусков и случайной величины**

Таким образом, можно расширить общую схему работы с пропущенными значениями, которую предложил Роберт И. Кабаков путём введения дополнительного этапа – исследование совместного распределения[[27]](#footnote-27). Рассмотрим изменённую схему работы с пропущенными значениями на рисунке 2.3, в которой помимо новой вершины также добавлена новая ветвь – альтернатива удалению пропусков, если известно распределение.



*Рис.2.3.* **Модифицированная схема работы с пропущенными значениями**

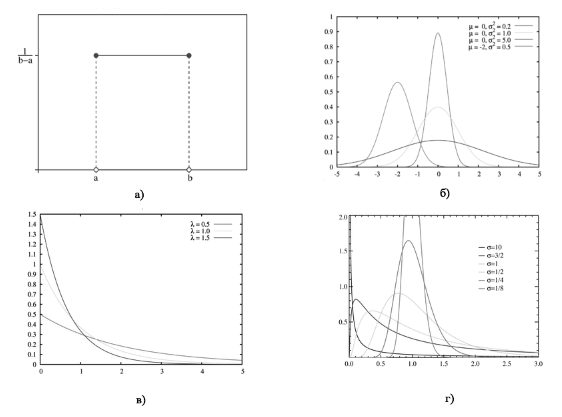
Полученная схема является одним из элементов новизны данной работы и нацелена на оптимизацию работ по анализу банковских данных. Раскроем новые этапы, предложенные автором данного исследования.

В целом дополнительный этап исследования пропусков имеет ряд преимуществ: первое – более точное определение метода работы с пропусками, как будет далее показано при определённых распределениях удаление пропусков вовсе не влияет на точность прогнозов, второе – возможность пересмотреть концепцию самого исследования. Иногда возникает ситуация, при которой наличие чрезмерно большого количества пропусков не позволяет подобрать метод, который бы обеспечил нужную точность прогноза. Тогда можно изучить распределение пропусков и исходных величин для внесения каких-либо корректировок в само исследование. Все дальнейшие исследования в рамках данной главы будут посвящены второму этапу – уточнение распределения пропусков за исключением небольшого отступления для пункта по изменению стратегии исследования.

Отметим, что новый элемент в схеме анализа пропусков раскрывает широкие возможности для аналитических отделов банков с точки зрения работы с большими данными. Это позволит существенно ускорить выбор теоретической модели, что в условиях быстро растущих данных крайне важно для экономии временных ресурсов. В статье Т. Миколова по планировании стратегии анализа больших банковских данных указывается, что тренировка моделей на больших сетах может занимать несколько недель[[28]](#footnote-28). Представим, что ведется перебор хотя бы пяти моделей, тогда результат крайне не приемлем для банка – расчеты могут требовать до месяца ожиданий, а поскольку скоринг нацелен на немедленные изменения – это не приемлемо, так как простой может повредить точность прогнозов[[29]](#footnote-29). Таким образом, реализуя схему на рисунке 2.3, можно сократить временные ресурсы в несколько раз.

## **2.1.** **Описание распределений пропусков**

Рассмотрим основные распределения, которые будут использованы для анализа пропусков – они же представляют и режимы. Существует практически бесконечное множество распределений случайных величин, однако выделим наиболее часто встречающиеся, они же в некотором смысле базовые. Плотности рассматриваемых распределений представлены на рисунке 2.4.



*Рис.2.4.* **Плотность вероятности случайных величин: а) равномерное б) нормальное в) экспоненциальное г) логнормальное**

Кроме того, далее будет описан алгоритм ядерной оценки плотности распределения, для которого в качестве ядра может быть использована одна из функций плотности.

Итак, на рисунке 2.4 под буквой а) изображена плотность равномерного распределения на интервале [a, b], для которого плотность вероятности на заданном интервале постоянна[[30]](#footnote-30):

(2.1)

Вообще равномерное распределение является отправной точкой для исследования, так как при данном распределении пропусков, как будет показано далее, будут получаться наиболее точные прогнозы. Важно отметить, что пропуски, например, при заполнении данных носят, как правило, систематический характер, что позволяет сделать предположение об их равномерном распределении. Также в рамках имитационного моделирования появляется возможность генерирования различных распределений, прибегая к использованию метода обратного преобразования. Рассмотрим преобразование Бокса - Мюллера для генерирования нормального распределения[[31]](#footnote-31). Данное преобразование активно используется при генерировании пропусков, подчиняющихся нормальному закону.

Пусть даны - независимые равномерно распределённые случайные величины на интервале (0, 1]. Рассмотрим величины , которые определяются по следующим формулам:

(2.2)

Тогда величины будут независимыми и нормально распределёнными с нулевым математическим ожиданием и единичной дисперсией. Поскольку получается стандартное нормальное распределение в рамках данных преобразований, то можно записать общий случай, когда необходимо получить распределение с произвольным математическим ожиданием и дисперсией :

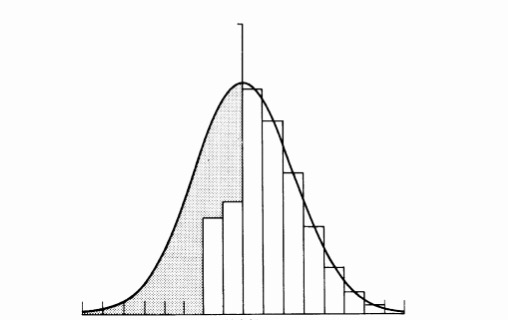
(2.3)

Следующий вариант – это подчинение пропусков нормальному закону, плотность изображена на рисунке 2.4 б) и формулой 2.4:

(2.4)

Нормальное распределение требует более тщательного изучения в силу своей вариативности – избыточный эксцесс или пологость, также тяжело оценить поведение распределения на хвостах[[32]](#footnote-32). Кроме того, могут возникнуть затруднения с интерпретацией такого распределения для пропусков.

Распределения, представленные на рисунке 2.4 под буквами в) и г) играют важную роль при изучении пропусков, которые сгущаются в одном месте. Достаточно нередкое явление, которое отмечали Wachter и Trussel в своей известной работе по изучению роста населения[[33]](#footnote-33). На рисунке 2.5 представлен один из экспериментов, когда большая часть пропусков была сосредоточена в одном месте.



*Рис.2.5.* **Сгущение пропусков в работе Wachter и Trussel**

Подобная ситуация со сгущением пропусков отлично моделируется экспоненциальным или логнормальным распределениями. В дальнейшем для моделирования пропусков и случайной величины с подобными распределениями можно также воспользоваться формулами 2.2, а затем вычислить экспоненту для полученных значений.

Заключительный этап данного раздела посвящен обсуждению моделирования распределений в статистической среде R, поскольку это единственный вариант установления приблизительного режима пропусков[[34]](#footnote-34). Существуют две основные концепции представления эмпирического распределения – гистограмма и ядерная оценка плотности. Последний метод используется для более точной оценки плотности распределения и в каждой точке моделируется следующим образом:

(2.5)

где x является последовательностью, состоящей из n элементов, K – некоторое симметричное ядро, достаточно часто используют Гауссово ядро, которое легко получить из формулы 2.4, h характеризует степень сглаживания. В целом манипуляции с параметром h позволяют получать более точные оценки плотности[[35]](#footnote-35).

Достаточно сложная процедура внесения пропусков в детерминированный вектор с распределением отличным от равномерного. Для этого необходима генерация булевого вектора такой же размерности с соответствующим распределением. Данный алгоритм предложен ниже в виде псевдокода:

*Листинг 2.1.*

**Алгоритм заполнения пропусков с учетом заданного распределения**

Заполнить (X*, пропуски, тип распределения*)

* Вход: вектор X размерности , количество пропусков, которые необходимо внести в соответствии с указанным типом распределения
* Выход: сгруппированный список с соответствующим распределением пропусков

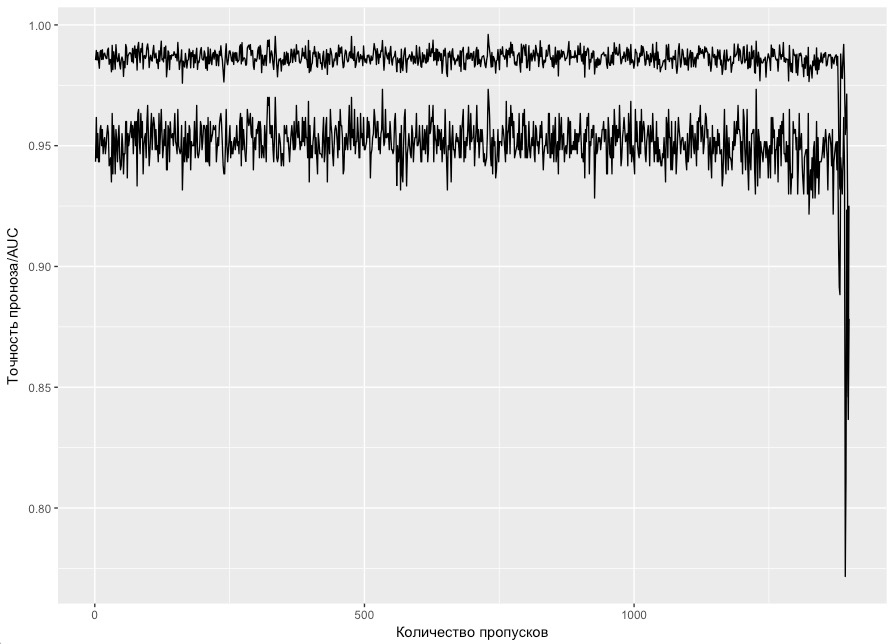
*for для всех u ∈ n:*

Каждая модель в данной работе будет представлена гистограммой изучаемой величины, внутри которой моделируется режим пропусков, а также будет произведено наложение ядерной оценки. Совокупная модель в некоторой степени дает представление об их совместном поведении и позволяет либо реализовать второй этап по схеме, предложенной на рис. 2.3, либо производить дополнительные уточнения. Данная схема реализована на языке R и выложена в репозиторий, который выступает приложением к данной работе.

## **2.2. Исследование режимов пропусков**

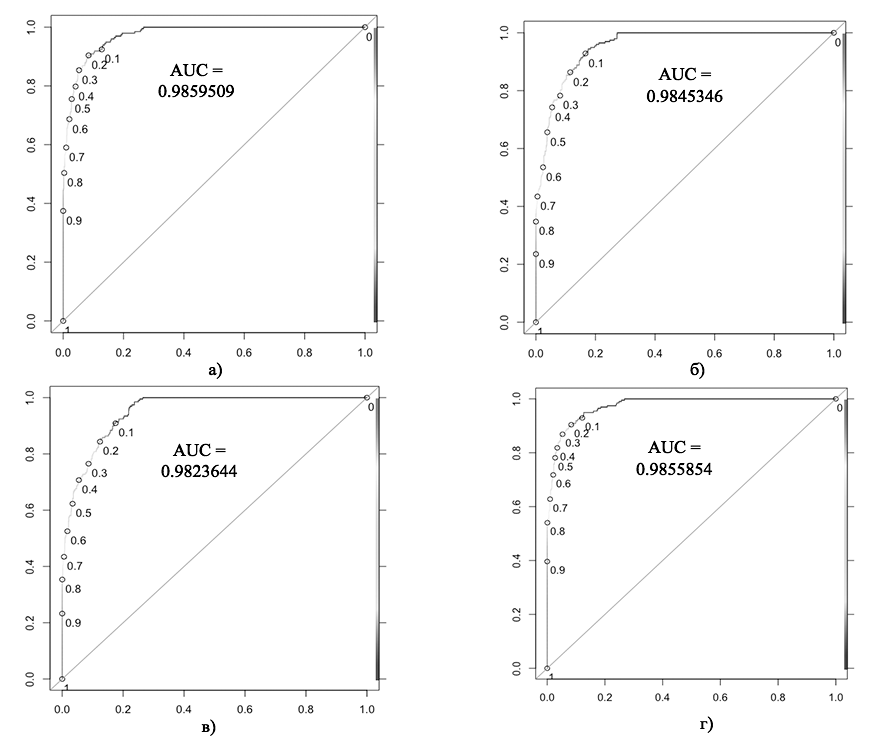
Рассмотрим ситуацию, когда пропуски и исходная величина имеют совместное равномерное распределение. В качестве скоринговой модели выступает логистическая регрессия, описанная в Главе 1. Для рассматриваемых данных ее статистически значимая спецификация имеет следующий вид:

На рисунке 2.6 рассматривается случай, когда пропущенное значение заменяется средним арифметическим по выборке, однако в случае одномерного анализа возможны другие средние характеристики, которые не оказывают существенного влияния на точность прогнозных значений: мода, медиана или средние внутригрупповые. Также допустим вариант с удалением пропущенных значений. Результаты по всем средним характеристикам приведены в таблице 2.1



*Рис.2.6.* **Бутстреп-анализ точности прогнозных значений**

Детально анализ точности прогнозов в рамках равномерного распределения можно изучить с помощью бутстреп-процедуры. Результаты приведены на рисунке 2.6. Важно отметить, что с ростом количества пропусков точность практически никак не меняется, лишь при стремлении к пороговому значению (сохраняя лишь необходимый минимум наблюдений для построения модели) точность падает на 15 – 20%.



*Рис.2.7.* **Испытания для совместного равномерного распределения:**

а) 10% исключений б) 30% в) 60% г) 90%

Поскольку в данной главе изучается преимущественно одномерный анализ, то в качестве переменной для экспериментов будет выступать возраст заемщика. На рисунке 2.7 представлены 4 испытания, в каждом из них в выборку вносился определённый процент пропусков – а) 10% б) 30% в) 60% г) 90%, а затем заменялись средним значением по выборке. Можно заметить, что показатель AUC практически не изменялся, однако заметны отклонения в точности прогнозов.

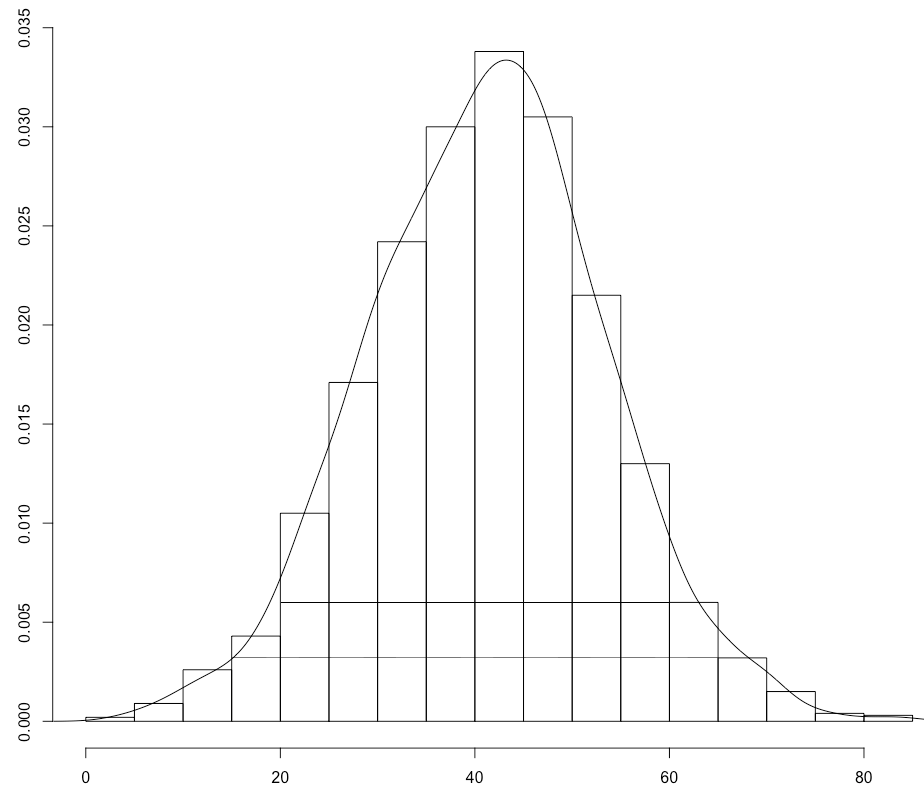
*Таблица 2.1.*

**Замена пропусков при равномерном распределении**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Тип | E[X] | | D[X] | |
| Точность | AUC | Точность | AUC |
| 1 | Удаление | 0.949 | 0.985 | 0.0001 | 0.0000 |
| 2 | Медиана | 0.949 | 0.986 | 0.0001 | 0.0001 |
| 3 | Мода | 0.948 | 0.985 | 0.0001 | 0.0000 |
| 4 | Среднее арифметическое | 0.951 | 0.986 | 0.0000 | 0.0000 |
| 5 | Среднее внутригрупповое | 0.952 | 0.987 | 0.0000 | 0.0000 |

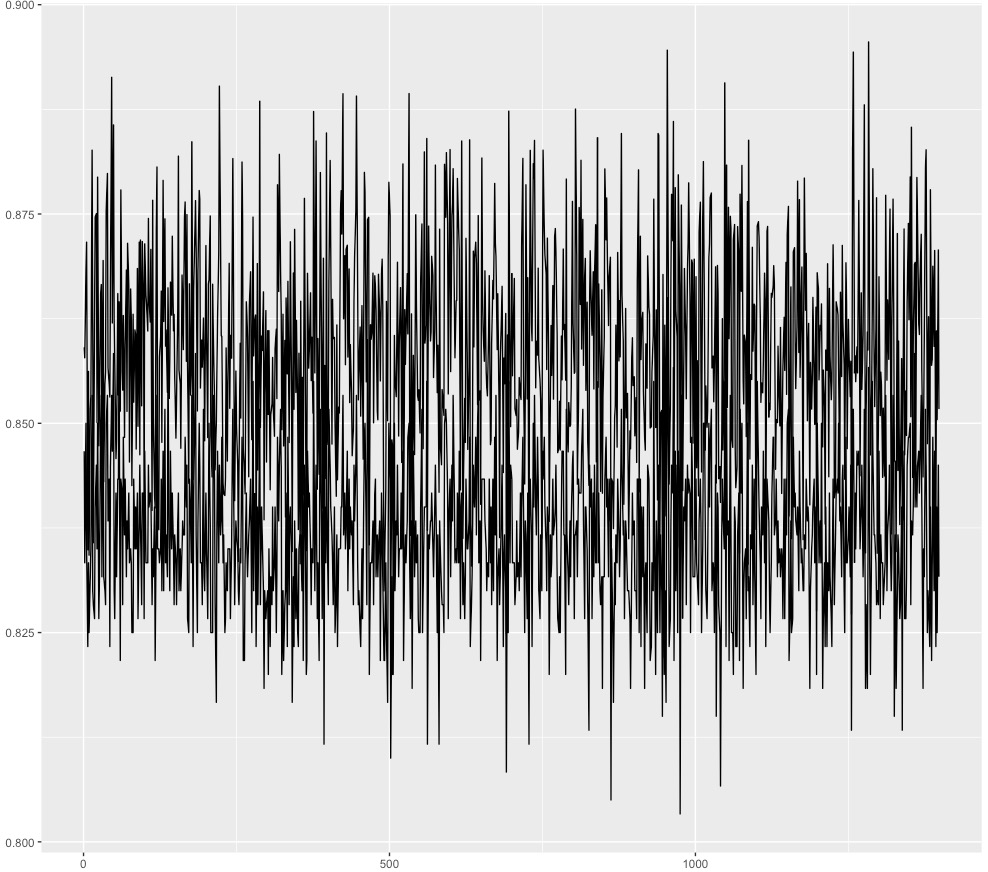
Результаты в таблице 2.1 заметно не различаются между собой, однако явно выделяются оценки под номерами 4 и 5, обладая минимальной дисперсией и наиболее высокой точностью прогнозных значений. Замена средним арифметическим является наиболее привлекательным вариантом, во-первых, достаточно простая формула и легко реализуемая на ЭВМ. Медиана и особенно мода требуют больше ресурсов, к тому же упорядочение выборки иногда может приводить к искажению результатов, например, когда речь идет о работе с временными рядами.

В практической деятельности равномерное распределение для исходной величины крайне редко встречается, мы скорее работаем с распределениями близкими к нормальному с различными аномалиями – нестандартное поведение на хвостах или асимметрия. Поэтому необходимо перейти ко второму режиму пропусков – когда исходная величина имеет нормальное распределение, а пропуски равномерное. Воспользуемся формулой 2.2 для преобразования типа распределения и листингом 2.1 для установления пропусков. Полученный результат представлен на рисунке 2.8. Подобная схема указывает на то, что при достаточно большом количестве пропусков возникают проблемы со значениями, которые расположены на хвостах, с другой стороны, сохранение наиболее часто встречающихся значений позволяет строить точные средние прогнозы. Проведем испытания с помощью бутстреп-процедуры: 10 заходов по 1000 имитаций.



*Рис.2.8.* **Равномерное распределение пропусков при**

Проанализируем ситуацию, представленную на графике 2.9, где верхняя линия представляет показатель AUC, в нижняя точность прогнозных значений. Примечательно, что изменение распределения на нормальное никак не влияет на точность прогноза с ростом количества пропусков при условии, что пропуски вносились равномерно и заменялись выборочным средним. Если обобщить данную ситуацию, то можно наблюдать: логистическая регрессия не обладает высокой чувствительностью к наличию достаточно большого числа пропусков, что может быть еще одним объяснением ее популярности в банковской среде. Например, те же нейронные сети крайне чувствительны к таким ситуациям, некоторые вовсе нельзя применять. Далее рассмотрим таблицу 2.2, аналогичную таблице 2.1, где покажем, что в случае одномерного анализа выбор метода замены пропусков важной роли не играет. Аналогично предыдущей ситуации: все оценки обладают практически одинаковой точностью (важно учитывать погрешности) и минимальной дисперсией.



*Рис.2.9.* **Бутстреп-анализ точности прогнозных значений**

*Таблица 2.2.*

**Замена пропусков при нормальном распределении**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Тип | E[X] | | D[X] | |
| Точность | AUC | Точность | AUC |
| 1 | Удаление | 0.857 | 0.855 | 0.0000 | 0.0001 |
| 2 | Медиана | 0.858 | 0.856 | 0.0001 | 0.0001 |

*Продолжение таблицы 2.2.*

**Замена пропусков при нормальном распределении**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 3 | Мода | 0.860 | 0.851 | 0.0001 | 0.0000 |
| 4 | Среднее арифметическое | 0.836 | 0.859 | 0.0000 | 0.0001 |
| 5 | Среднее внутригрупповое | 0.863 | 0.862 | 0.0001 | 0.0001 |

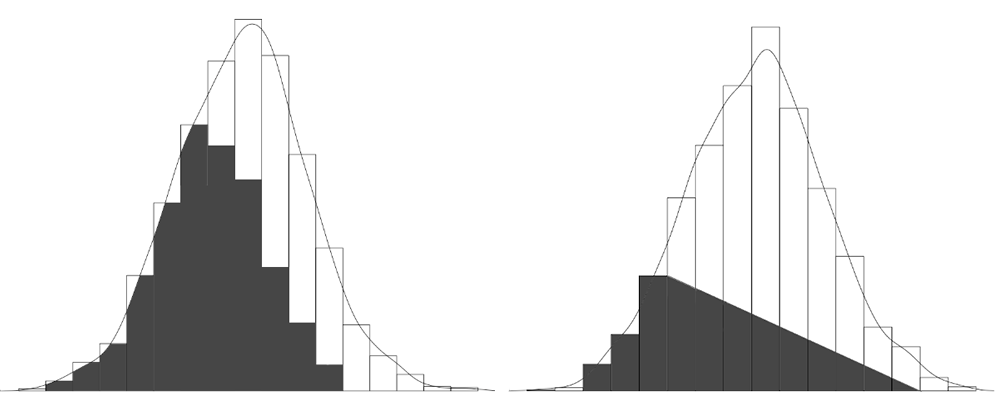
Далее в работе рассматриваются практические рекомендации в разрезе оценки рисков. Поскольку данная глава служит базой для многомерного анализа, который является основой скоринговых моделей, то в практические рекомендации включена часть, направленная на усовершенствование расчетов при составлении скоринговых карт, а также изложенная методология может быть использована для прогнозирования пропущенных значений.

Например, в последнем режиме пропусков, когда исходная величина распределена по нормальному закону и при этом наблюдается существенный срез данных на хвостах или в целом много пропусков равномерно распределённых, а пространственный прогноз необходим, то можно воспользоваться медианной регрессией:

(2.6)

и получить более точные оценки, самое главное – менее чувствительные к пропускам или их заменам[[36]](#footnote-36). Таким образом, после изучения второго этапа на рисунке 2.3, перейдем к этапу – изменение стратегии исследования. Однако прежде рассмотрим еще некоторые режимы с целью обобщения медианной регрессии.

На практике часто возникает ситуация, когда пропуски концентрируются в какой-либо области исходного распределения, как это было показано на рис. 2.5, например, в случае с клиентами банка – более обеспеченная часть склонна к сокрытию части информации о себе. Подобная ситуация может быть описана распределениями 2.4 в) или г). На рисунке 2.10 представлены наиболее часто встречающиеся ситуации в комбинации с нормальным распределением. Конечно, концентрация пропусков может быть не только со скосами, но в центре или на любом из хвостов.



*Рис.2.10.* **Концентрация пропусков в распределении**

В подобных ситуациях интерполяция значений даже продвинутыми методами исключена, поскольку либо полностью, либо большая часть наблюдаемых значений отсутствует. Обычно исследователи удаляют подобную переменную, однако в данной работе предлагается к рассмотрению альтернативный вариант – медианная регрессия или ее обобщение квантильная регрессия. Это позволит сохранить переменную и извлечь важную информацию.

## **2.3. Квантильная регрессия**

Квантильная регрессия может выступать отличным инструментом при определённых режимах пропусков или в случае, если наблюдается большая ошибка при замене пропусков. Кроме того, поскольку квантильная регрессия относится к семейству непараметрических методов, то появляется возможность ослабить предпосылки, которые характерны, например, для обычного метода наименьших квадратов: выбросы и тип распределения оказывают меньшее влияние.

В практических целях для моделирования рисков такой подход позволит извлечь из данных больше информации. В частности, при построении скоринговой карты банк, как правило, располагает минимальным объемом информации; при большом количестве пропусков происходит исключение рассматриваемой переменной или замена с большой ошибкой[[37]](#footnote-37). Поэтому перед банковскими отделами, занимающимися скорингом, ставится фундаментальная задача – обогащение экономических данных, которое в основном происходит за счет расчетов различных агрегатов и производных характеристик. В данном разделе предлагается дополнительный способ обогащения данных за счет изучения типа распределения пропусков, а затем использование непараметрических методов интерполяции, или же просто изучение поведения характеристик. Тем самым минимизируются банковские риски[[38]](#footnote-38).

Однако начнем рассмотрение с базовой идеи, частного случая квантильной регрессии – это медианная регрессия (условная медиана), которая представленна формулой 2.6 в предыдущем параграфе. При данной регрессии моделируется ситуация, когда выше некоторого числа и ниже расположены по 50% выборки. Такой подход крайне важен, когда мы идентифицировали достаточно большое количество пропусков на хвостах распределения, см. рис. 2.8. Важно заметить, что оценки по медианной регрессии будут состоятельными при таком режиме пропусков в отличие от условного среднего. Заметим, что если исходная величина имеет симметричное распределение, то асимптотически разницы между условным средним и медианной регрессией нет, поэтому при таком режиме пропусков можно разработать веса и для МНК-оценок.

Изучая ситуации на рисунке 2.10, когда пропуски имеют экспоненциальное распределение или логнормальное, в общем случае – сконцентрированы в некоторой части исходного распределения, то медианной регрессии недостаточно. Рассмотрим ее обобщение на случай квантиля произвольного порядка. Пусть даны некоторые , где - вектор факторов регрессионного уравнения, тогда:

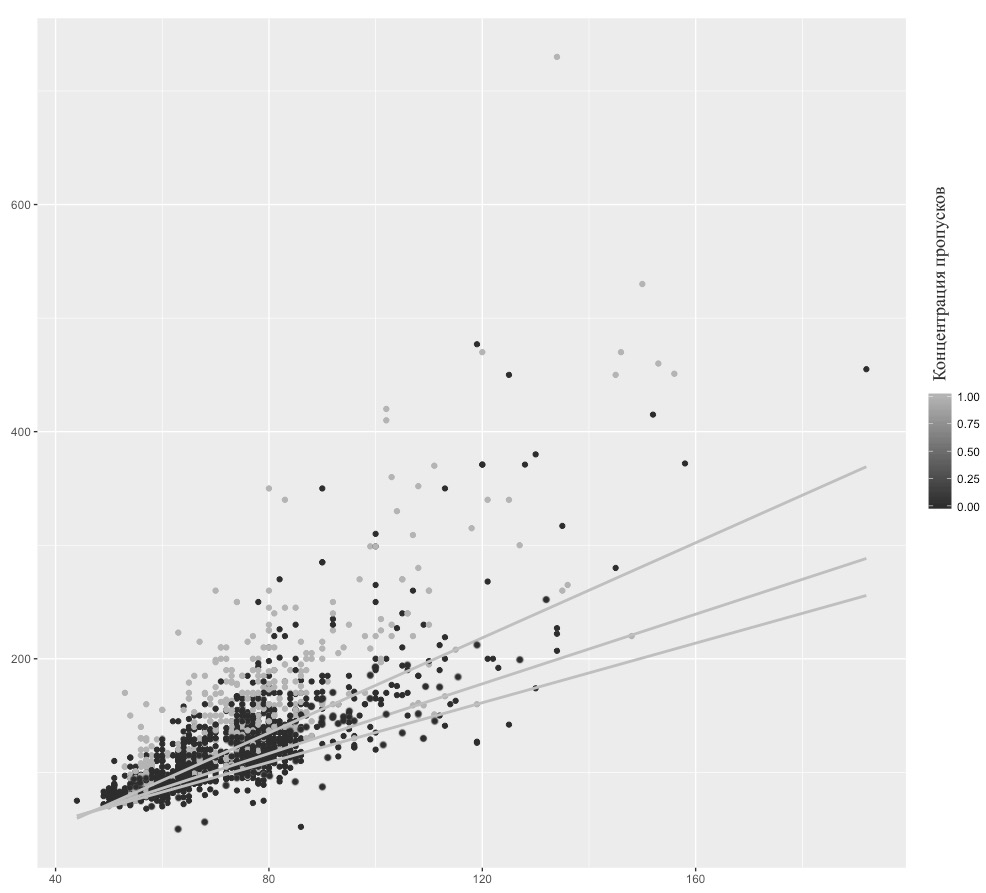
(2.7)

Для удобства перепишем соотношение 2.7 в следующей форме:

(2.8)

где обозначает условную квантиль порядка . Рассмотрим на примере изучаемых данных в данной главе условную ситуацию, которая часто возникает в практической деятельности. Предположим, что заемщики упорядочены по уровню их доходов. При увеличении их дохода растет и склонность к сокрытию информации. Поэтому разумно разбить множество заемщиков.

Нижняя квантиль распределения представляет заемщиков с низким уровнем дохода, верхняя – с высоким. Поскольку пропуски сконцентрированы именно в левосторонней области распределения, то, следуя схеме на рисунке 2.3, можно рассмотреть следующую стратегию работы: для заёмщиков с высоким и средним уровнем дохода рассмотреть экспертные методы прогнозирования, а для остальных применим аппарат квантильной регрессии, тем самым сохранив важную статистическую информацию для дальнейших исследований. Случай одного регрессора показан на рисунке 2.11, где также представлена гистограмма концентрации пропусков, что соответствует ситуации Вотчера и Труззеля, см. рис. 2.5.



*Рис.2.11.* **Квантильная регрессия с концентрацией пропусков**

# **Выводы**

В данной главе изучен механизм пропусков при различных режимах, что открывает новые возможности работы с пропущенными значениями и позволяет заложить основы для дальнейшего исследования методов работы с неполной информацией при многомерном анализе. Кроме того, данный механизм исследован в разрезе моделирования оценки банковского риска при выдаче займов на примере логистической регрессии, что подтвердило ее устойчивость при различных режимах. Также предложена модифицированная схема работы с пропущенными значениями, в которую включен этап по уточнению условного распределения пропусков. Результат модификации продемонстрирован на примере с квантильной регрессией, когда высокая концентрация пропусков не позволяет их заменить на какие-либо значения, но потеря данных недопустима.

# **ГЛАВА 3. ОПРЕДЕЛЕНИЕ РЕЖИМА ПРОПУСКОВ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО КОМПЛЕКСА R**

Определение режима пропусков, то есть установление условного распределения пропусков в рамках изучаемой случайной величины, предполагает не только установление его типа, но и полную процедуру восстановления пропущенных значений, которая включает выбор оптимального метода прогнозирования пропусков и сравнительный анализ с базовыми методами. Относительно базовых методов можно провести аналогию с классической регрессионной моделью: если полученные результаты хуже, чем прогноз по среднему – такая модель считается полностью непригодной. Данная глава освещает установление режима пропусков в зависимости от типа данных – пространственный анализ или временные ряды.

Однако прежде чем перейти к эмпирическому исследованию, необходимо рассмотреть общее определение матрицы-индикатора, то есть ее многомерный вариант, поскольку в большинстве случаев только правильно построенная с полными анализом матрица объект-признак может дать возможность установить паттерн распределения пропусков. Например, если один из признаков хотя бы линейно связан с другим, то, упорядочив один из них, можно проверять гипотезы относительно другого, в котором есть пропущенные значения.

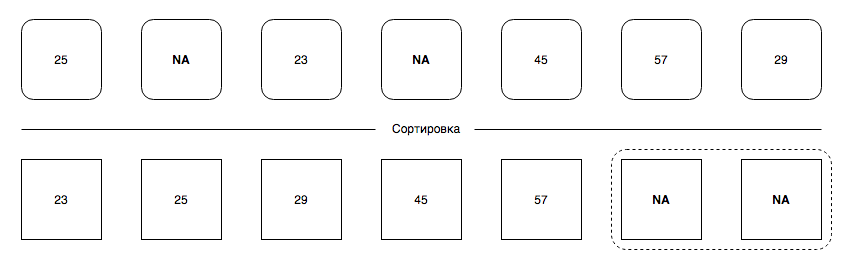
## **3.1. Общее определение матрицы-индикатора**

Одномерный случай матрицы-индикаторы был представлен во второй главе и подходит для анализа, например, временного ряда. В прогнозировании оценки кредитного риска тоже возникают ситуации, требующие сужения исследования до временного ряда, например, когда рассматривается транзакционная история клиента за некоторый период с целью установления его поведения. Однако при обучении общей модели (первичное прогнозирование вероятности возврата денежных средств заемщиков), конечно, рассматриваются либо пространственные, либо панельные данные. Для такого случая проводится анализ многомерной по признакам матрицы, которая представлена формулой 3.1, где 0 – это отсутствие пропущенного значения, а 1 – наличие.

(3.1)

Заметим, что в данной матрице пропуски находятся только в векторе , остальные же данные полностью заполнены. Такая структура широко распространена в анализе данных, поскольку исходные данные можно легко привести к подобному виду. Выделим и рассмотрим ключевую особенность такой структуры.

При восстановлении пропущенных значений в рамках изучаемого подхода необходимо сначала построить условное совместное распределение изучаемой случайной величины и пропусков в ней.



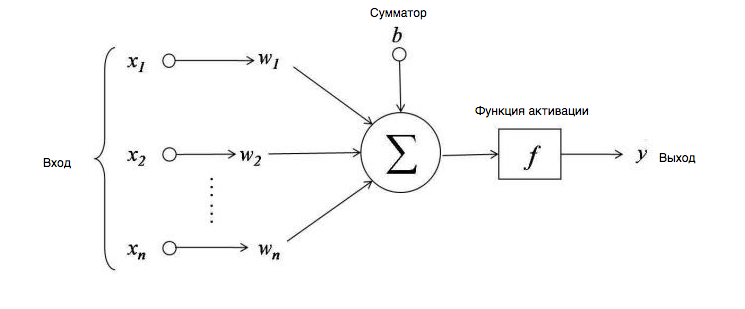
*Рис.3.1.* **Сортировка вектора с пропущенными значениями**

В случае пространственных данных значения переменной необходимо сортировать по возрастанию. При выполнении данной процедуры возникает ситуация неопределённости расположения пропусков. Рассмотрим рисунок 3.1, который иллюстрирует данный процесс. В программном комплексе R пропуски обычно маркируют ключевым словом NA и на рисунке они распложены на 2 и 4 позициях первой строки[[39]](#footnote-39). Такое расположение абсолютно случайное. Во второй строке представлен отсортированный вектор, где пунктиром обозначена область, элементы которой выбывают из рассмотрения. Разумеется, чтобы построить полное распределение, необходимо определить, или восстановить, индексы, характеризующие положение каждого пропуска.

Для этого можно воспользоваться различными методами математической статистики, однако в общем случае необходимо решить либо регрессионную задачу, либо задачу кластеризации. Если большинство признаков имеют количественный тип и данные достаточно полные, следует воспользоваться регрессией. В данной ситуации представляет интерес лишь прогнозируемое значение зависимой переменной, поэтому неважно, какой тип регрессии используется или каким способом отыскивается искомое значение. Можно воспользоваться классической множественной линейной регрессией, представленной формулой 3.2[[40]](#footnote-40):

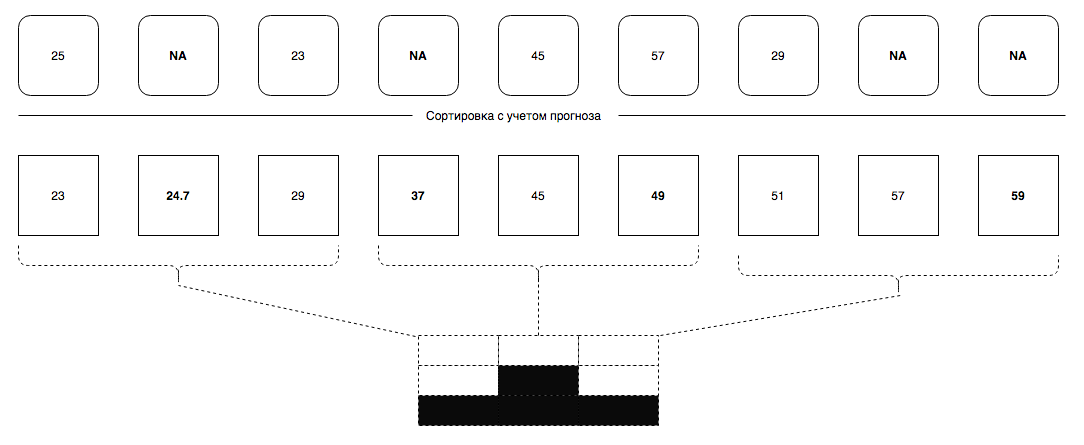
(3.2)

Либо же применить более сложные методы, например, использовав нейронные сети для решения задачи регрессии. Общий случай представлен на рисунке 3.2. Однако применение нейронной сети сопряжено с рядом рисков, например, ее переобучение или слишком долгий процесс обучения[[41]](#footnote-41). Кроме того, полученные результаты могут не иметь существенно более высокой точности. Поэтому в эмпирическом исследовании будут использоваться классические как линейные, так и нелинейные регрессионные модели для прогнозирования.



*Рис.3.2.* **Схема базовой нейронной сети для решения задачи регрессии**

Так как при использовании различных методов регрессионного анализа, получаем количественный результат, а не конкретный индекс, – возникает вопрос относительно того, можно ли данное значение принять для замещения пропуска. Для раскрытия данного вопроса модифицируем рисунок 3.1, добавив в него прогнозные значения, сортировку с их учетом и дополнительные элементы.



*Рис.3.3.* **Сортировка вектора с восстановленными значениями и гистограммой распределения пропусков**

На рисунке 3.3 восстановлено положение пропущенных элементов с указанием их прогнозных значений. Однако, как можно заметить, на гистограмме их распределение отличается от равномерного, следовательно, от типа поведения как самих пропусков, так и исходной величины будет зависеть и выбор оптимальных методов для прогнозирования. Например, если распределение близкое к нормальному, то, очевидно, что на его хвостах данные более однородны и применив метод цепных уравнений можно достичь более точных результатов прогнозирования. Таким образом, зная базовые значения пропущенных значений, появляется возможность восставить порядок в исходном векторе, и поскольку в данном случае идет работа с интервалами (построение гистограммы), то точность самих прогнозных значений может оказаться неприемлемой для исходной задачи, особенно когда решается вопрос выдачи займа.

Нельзя не отметить еще один широко распространённый подход для восстановления индексов в векторе с помощью кластеризации. Располагая всеми данными из матрицы, которая представлена формулой 3.1, можно выделить однородные группы, усилив влияние главного признака (в задаче регрессии является зависимой переменной). Получив все группы – необходимо расположить элементы в правильном порядке. Однако данный подход обладает рядом недостатков.

Первый недостаток: количество полученных кластеров должно согласовываться с оптимальным количеством интервалов, на которые разбивается наблюдаемый диапазон изменения случайной величины при построении гистограммы плотности её распределения. Например, согласование с эмпирическим правилом Стёрджеса[[42]](#footnote-42).

Второй, отсутствие приближенных оценок значений пропусков, что, вообще говоря, может нарушить границы вхождения того или иного пропуска. Кроме того, построение модели кластеризации требует как вычислительных ресурсов, так и временных: необходимость нормализации данных, построение специальных матриц в случае, когда данные имеют смешанный тип – категориальные, порядковые или количественные переменные[[43]](#footnote-43). Кроме того, разные методы кластеризации могут давать принципиально различающиеся результаты.

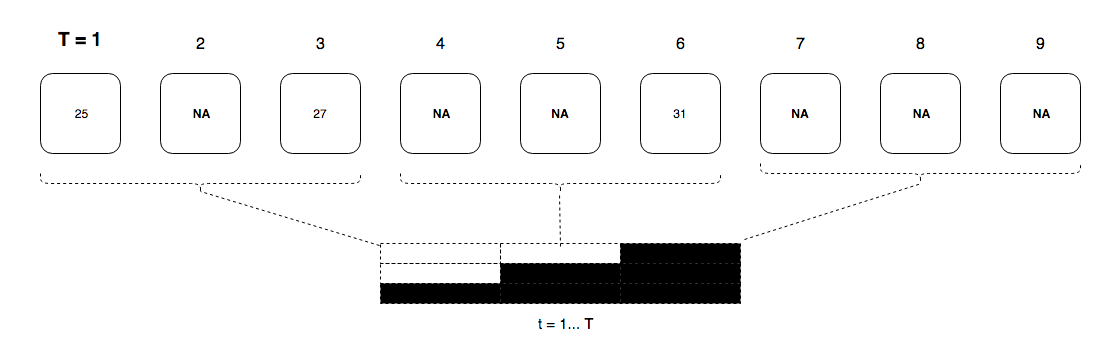
## **3.2. Определение режима пропусков с использованием временных рядов**

Временные ряды играют важную роль при оценке кредитного риска, поскольку они могут быть как частью панельных данных, так и иметь самостоятельную функцию. Первый вариант означает, что если рассмотреть панельную структуру, то можно явно выделить из нее временные данные и работать непосредственно с ними. Восстановив пропущенные значения, данные необходимо вернуть обратно в панельную структуру и продолжить исследования. Второй вариант означает, что временной ряд интересен как отдельная сущность для анализа. Рассмотрим подробнее такие ситуации.

При общем прогнозировании оценки кредитного риска исследователь сфокусирован прежде всего на общей выборке из всех клиентов банковской организации. Обучив модель, ее добавляют в банковскую систему управления данными и затем используют при выдаче займов[[44]](#footnote-44). Однако вернемся к алгоритмам оценки кредитного риска из Главы 1, где демонстрируется, что, как правило, алгоритм дает ответ на вопрос – выдавать кредит или нет, сопровождая данный результат вероятностной оценкой. Именно на основе значения этой вероятности и принималось решение алгоритмом о выдаче займа, но такая вероятность может часто оказываться на границе порога активации. То есть клиенту все-таки можно было бы выдать кредит. Таких клиентов на практике принято отсекать и проводить дополнительные исследования. Как раз для таких ситуаций и рассматривают его транзакционную историю, которая имеет временную структуру. Причем транзакционная история может быть не только в банке, который выдает кредит, но и история из других источников. Обычно такие источники предоставляют информацию с пропусками, которые подлежат восстановлению[[45]](#footnote-45).

Еще одна ситуация, когда анализируют временные ряды для оценки кредитного риска – это клиенты специальных категорий, например, привилегированный сегмент. Обычно таких клиентов небольшое количество и построить классические модели с приемлемыми результатами нет возможности, поэтому анализируют историю их поведения.

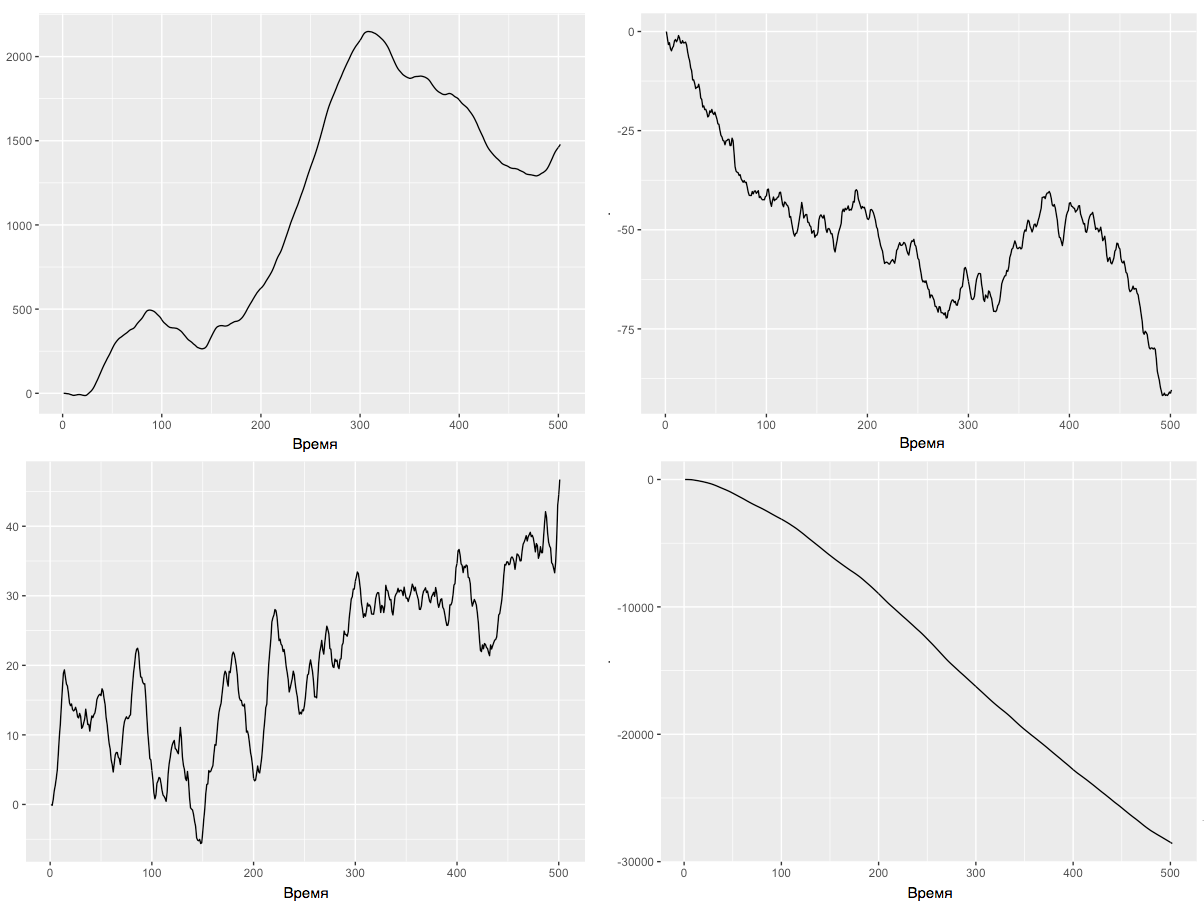
В предыдущем параграфе анализировались методы восстановления положения элементов в векторе. Для временных рядов этот шаг опускается, поскольку эти данные упорядочены естественным образом, что существенно упрощает построение распределения. Рассмотрим рисунок 3.4, на котором представлена гистограмма распределения временного ряда. Важно заметить, что данные сгруппированы по времени, поэтому относительно времени имеем равномерное распределение, а вот сами пропуски могут иметь различный режим. Например, на данном рисунке имеем скошенное распределение и для таких режимов увязывается теория с восстановлением пропущенных значений с помощью квантильной регрессии.



*Рис.3.4.* **Гистограмма распределения временного ряда**

Введем некоторые замечания относительно дальнейших экспериментов. Поскольку транзакционная история клиента является крайне конфиденциальной информацией – ее запрещено использовать даже в скрытом виде: агрегированные или сгруппированные данные. Поэтому для расчетов по временным рядам будут сгенерированы специальными образом выборки и внесены пропущенные значения с заранее неизвестным распределением.

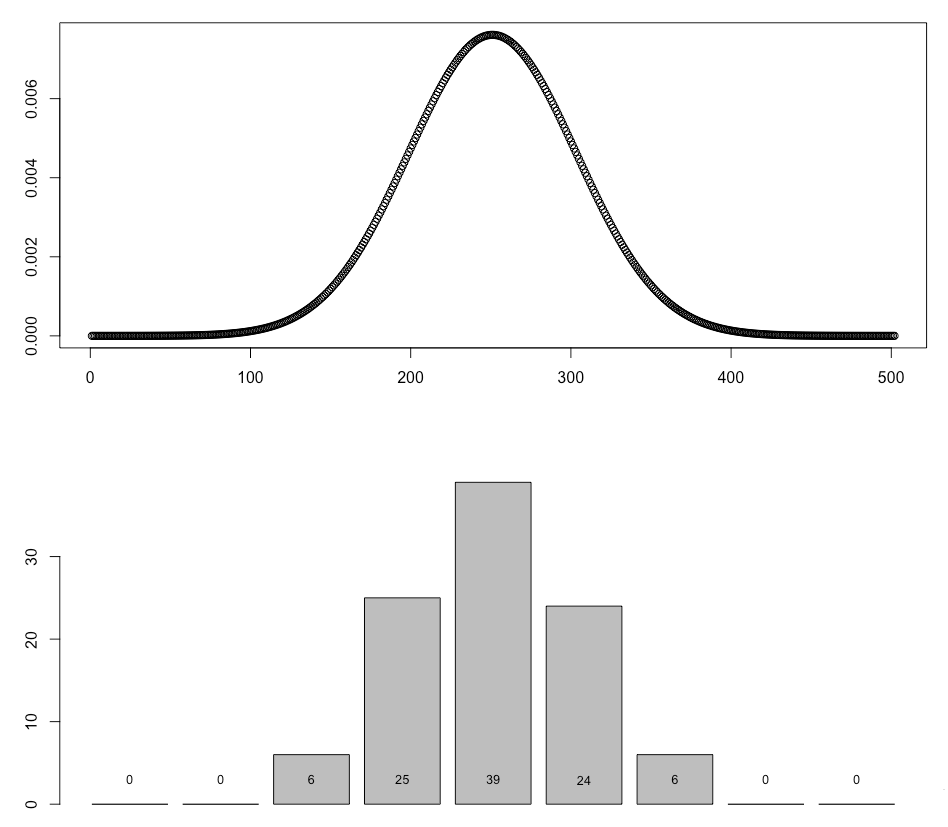
Рассмотрим алгоритм генерирования пропусков с определённым распределением. Заметим, что такой алгоритм технически является уникальным и ранее в других источниках не воспроизводился. Алгоритм был разработан автором данного исследования на статистическом языке R согласно листингу 2.1, предложенному во второго главе и успешно применен на тестовых данных.



*Рис.3.5.* **Источник временных рядов**

Первый шаг алгоритма – это получение данных временного ряда. Поскольку рассматриваются искусственные временные ряды, то необходимо получить корректные данные, отражающие реальные процессы. Для этого шага сгенерируем не просто набор случайных значений, а полноценный временной ряд, который подчиняется определённому закону. В качестве такого закона выбран класс моделей временных рядов ARIMA[[46]](#footnote-46). Именно эти модели являются источником изучаемых данных. На рисунке 3.5 приводится иллюстрация временных рядов, сгенерированных случайным образом в соответствии с законами данного класса моделей путем конфигурации AR и MA компонентов[[47]](#footnote-47).

Внесем пропуски c заданным распределением. Для этого сгенерируем вектор вероятностей, подчиняющийся закону распределения K с таким же количеством элементов, как и в исходном векторе. Затем рассчитаем основные моменты распределения. Полученный вектор вероятностей используем как вектор весов для внесения пропусков.



*Рис.3.6.* **Плотность распределения пропущенные значений**

Получив данные временного ряда с пропусками, определим количество интервалов для построения гистограммы в соответствии с правилом Стёрджеса, которое представлено формулой 3.3:

(3.3)

где N – количество наблюдений в выборке. Получив количество интервалов, проведем соответствующее расщепление данных, и для каждой группы вычислим количество пропущенных значений. На рисунке 3.6 представлен результат работы алгоритма, где верхний график на рисунке – это плотность распределения весовых коэффициентов, а нижний – плотность распределения пропусков. Как можно заметить, пропуски распределены в соответствии с плотностью нормального распределения.

Проведем эксперимент для базовых методов замены пропусков, к которым относят замену средним и медианой. Сравнение будет производится с модифицированным средним и медианой для замены пропусков на примере режима, представленного на рисунке 3.6. Определим различия между обычными методами и модифицированными. Первые подразумевают работу с полной выборкой и их характеристики рассчитываются без учета особенностей пропусков. Вторые рассчитываются в рамках целого алгоритма, который имеет ряд преимуществ:

1. Учет режима пропущенных значений: алгоритм на вход получает условное распределение пропусков, и для каждой группы рассчитывается среднее значение для замещения. Затем данные объединяются в общий вектор.
2. Использование комбинации базовых характеристик, например, изучив особенности распределения пропущенных значений, алгоритму можно явно передать указания, в какой ситуации необходимо применять среднее, а в какой - медиану.
3. Способность работать с одномерной структурой. Большинство алгоритмов для замещения пропусков не адаптированы для работы без вспомогательных переменных, что является существенным недостатком при анализе временных рядов[[48]](#footnote-48).

Далее целесообразно привести псевдокод алгоритма, разработанного автором исследования, реализующего идею нового подхода по замещению пропусков в данных. Отметим, что алгоритм полностью воспроизведен на языке R. Алгоритм представлен листингом 3.1.

Отметим некоторые положения: определение режима пропусков ложится на исследователя на основе матрицы, которая получена на первом шаге алгоритма (матрица, отражающая структуру пропусков). Параметры-указания – это специальные параметры, которые определяют методы замены пропусков на каждом из участков. В данной ситуации исследователю (банковскому аналитику) следует тщательно изучить полученные эмпирические распределения, и только потом переходить к формированию списка указаний. Важно отметить, что правило Стерджесса не является ограничителем для алгоритма. Исследователь может подобрать любое из существующих правил по разбиению данных в зависимости от поставленной задачи или иных факторов.

*Листинг 3.1.*

**Алгоритм построения распределения и замещения пропусков**

Построить распределение (X *[Входные данные]*)

* Вход: X - вектор с пропусками или матрица объект-признак с указанием столбца, в котором наблюдаются пропуски
* Выход: матрица, отражающая структуру пропусков

Разбить данные на группы по правилу Стерджесса [Пользовательское правило]

Подсчитать количество пропусков на каждом участке

Занести данные по пропускам в матрицу и отметки участков

Переход к следующему преобразованию

* Вход: Матрица предыдущего шага и указания относительно каждого участка

Определение режима пропусков

Формирования вектора весовых коэффициентов в соответствии с распределением

Замещение пропусков на каждом участке в соответствии с режимом и параметрами-указаниями

* Выход: целостные данные
* Завершение работы алгоритма

Основная метрика для сравнения качества моделей – MSE, которая представлена формулой 3.3[[49]](#footnote-49). Поскольку ответы заранее известны, данная метрика полностью отражает качество прогнозов.

(3.3)

Значения в таблице 3.1 – это результат имитационного моделирования. Для каждого подхода и способа замещения пропущенных значений генерировался временной ряд со своей структурой. В таблицу заносились усредненные значения метрики качества.

В таблице 3.1 представлены результаты работы алгоритма и указанием степени превосходства по сравнению с традиционным подходом. Значения в ячейках таблицы – это вычисленная метрика MSE. Как можно заметить, предложенный подход позволяет увеличить точность прогноза даже на базовых методах в среднем в 2 – 3 раза. Комбинированный же подход вовсе дает наилучший результат. К стандартному подходу комбинация не применима, поскольку ничего не известно о поведении пропусков, следовательно, невозможно определить, как дифференцировать методы прогнозирования.

*Таблица 3.1.*

**Сравнительный анализ методов замещения пропусков**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Стандартный подход | Модифицированный подход | Степень превосходства |
| Среднее | 421.1359 | 122.6024 | 3.43 |
| Медиана | 363.7749 | 128.5909 | 2.83 |
| Комбинация | - | 110.4345 | - |

Рассмотрим еще один подход, обладающий высокой точностью прогнозирования пропусков. Для этого возьмем более сложную комбинацию, состоящую из моделей временного ряда и характеристик распределения. Модели будут применяться на тех участках, где концентрация пропусков минимальна, например, первый и последний квартиль. Те группы (высокая концентрация пропущенных значений), где модель не применима и в зависимости от дисперсии – будет применяться медиана, мода или среднее. Результаты представлены в таблице 3.2.

*Таблица 3.2.*

**Сравнительный анализ методов замещения пропусков**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Стандартный подход | Примесь | Модифицированный подход | Степень превосходства |
| Регрессия | 157.3567 | Медина, Среднее | 123.4565 | 1.27 |
| ARIMA | 135.7985 | Мода, Среднее | 129.5439 | 1.04 |
| Комбинация | - | ARIMA,  регрессия, характеристики распределения | 105.4567 | - |

Однозначно заметны улучшения в сравнении со стандартным подходом, особенно выделяется с наилучшим результатом итоговая комбинация, поскольку в ней задействован каскад моделей, что позволяет лавировать между ними, и в зависимости от распределения пропусков выбирать приемлемую модель. Однако такой способ обладает существенным недостатком – временные затраты. Если рассмотреть автоматизацию данного процесса, что характерно для банковского сектора, то необходима будет очень тщательная настройка системы с применением машинного обучения.

Для подведения итогов по работе с временными рядами рассмотрим рисунок 3.7, на котором изображен размер ошибки MSE для каждого подхода и ее соотношения. Для всех методов новый подход на основе изучения распределения пропусков дает лучший результат.

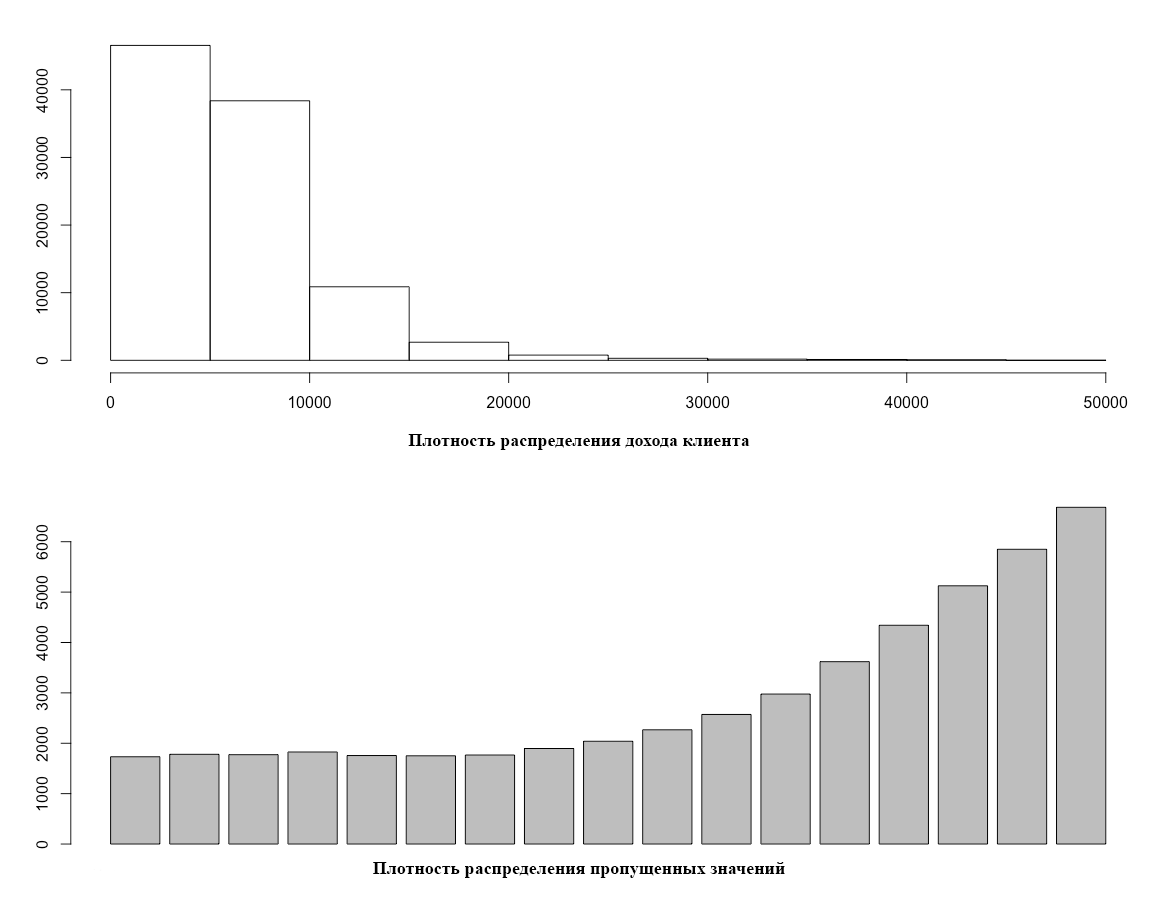
*Рис.3.7.* **Итоговое сравнение методов замещения пропусков на основе соотношения ошибки прогноза MSE**

Особенно заметны улучшения для простейших методов замены пропусков – в среднем результат выше в несколько раз. Важно обратить внимание, что работа проделана с одномерной структурой, имеющей эмпирическое нормальное распределение. Конечно, существует бесконечное множество других распределений, с которыми может столкнуться исследователь, но, зная поведение пропусков, всегда можно подобрать оптимальное соотношение методов для получения наилучшей обработки. Сложные методы работают эффективно только на больших матрицах объект-признак, поэтому на рисунке 3.7 результаты не такие впечатляющие. К тому же, при таком подходе снижается их точность в связи с урезанием выборки[[50]](#footnote-50).

Таким образом, предложенная методология замены пропущенных значений во временных рядах расширяет инструментарий исследователя, давая возможность получения более точных конечных прогнозов, что особенно актуально для банковского сектора, где минимальное улучшение оказывает существенное влияние на финансовый результат.

## **3.3 Определение режима пропусков в рамках многомерного анализа**

Особенности и методы работы с многомерной структурой данных были рассмотрены в начале данной главы: упорядочивание элементов вектора и прогнозирование пропущенных значений. Необходимо по аналогии с анализом временных рядом рассмотреть эмпирические результаты, однако заметим одно важное отличие от них – здесь будут использованы продвинутые методы замещения пропусков и реальные данные, которые были описаны в первой главе, в параграфе 1.3.3.



*Рис.3.8.* **Совместное условное распределение исходной величины и пропусков в ней**

На рисунке 3.8 установлен режим пропущенных значений для переменной - доход клиента, предварительно восстановив порядок значений с помощью кластерного анализа. Поскольку доход клиента наиболее значимая переменная при выдаче займа – прогнозирование пропусков в ней требует максимальной точности. Заметим важный факт, что с ростом дохода – клиент склонен скрывать его или вовсе не указывать. Данная гипотеза рассматривалась в Главе 2. Зная режим пропусков, можно разработать стратегию исследования: обычные клиенты имеют равномерное распределение пропусков – для такого случая подойдут продвинутые методы замещения пропусков. Случай с клиентами, у которых высокий доход более интересный – здесь применим квантильную регрессию. В качестве вспомогательной переменной можно выбрать коррелирующую с доходом переменную, например, возраст клиента. В конечном итоге объединим результаты и сравним с случаем, если бы не было установления режима пропусков.

*Таблица 3.3.*

**Сравнительный анализ методов замещения пропусков**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Стандартный подход | Модифицированный подход | Примесь | Степень превосходства |
| Многомерная оценка цепными уравнениями | 3644.232 | 3487.7768 | Квантильная регрессия | 5% |
| kNN-оценка | 3949.781 | 3801.9867 | 4% |

В сравнении с традиционными методами замены пропусков прирост в точности прогнозирования в среднем растет на 5%. Это может показаться незначительным, но стоит обратить внимание, что доля клиентов из премиального сегмента небольшая, поэтому и прирост соответствующий. Однако в рамках сегмента получен существенный результат и точно определены границы работы методов.

*Рис.3.9.* **Итоговое сравнение результатов на основе метрики MSE**

# **Выводы**

В данной главе проведено обобщение предложенного подхода при работе с пропусками на случай различных типов данных. Введены метрики анализа качества результатов. Рассмотрены модели восстановления последовательностей, включая кластерный анализ и нейронные сети. Ключевой момент – это результаты, полученные при апробации предложенного подхода как на искусственных данных, так и на реальной выборке. В итоге получен существенный прирост в точности прогнозирования пропущенных значений. Приведены сравнительные таблицы. Важно отметить, что исследование проведено в разрезе двух типов данных – временные ряды и пространственные многомерные структуры, представленные матрицами объект-признак. Обратим внимание, что обобщенная теория замены пропусков во временных рядах ранее не предлагалась. В главе также предложен уникальный алгоритм, реализующий идеи автора исследования по новому подходу к работе с пропущенными значениями.

# **ЗАКЛЮЧЕНИЕ**

В работе выполнен системный анализ как отечественных, так и зарубежных работ по рассматриваемой проблематике; предложены классификации пропусков с учетом особенностей предметной области и разработан совершенно новый подход к анализу пропущенных значений в данных, который позволяет снизить ошибку при оценке кредитного риска с целью улучшения точности предсказания неблагоприятных исходов, вызванных невозможностью погашения заемщиком ссуды. Кроме того, важным элементом работы является изучение прогнозирования пропусков с учетом основных тенденций в области анализа данных, таких как большие данные и машинное обучение. Современные банковские организации уже не обходятся без их применения.

Основные итоги выполненного исследования заключаются в следующем: разработаны классификации пропусков в зависимости от их источника с учетом банковских систем: профильные, общие и случайные. Анализ источника позволяет разработать комплекс мер для дальнейшей работы с ними. В процессе исследования были предложены эвристические методы выбора способа прогнозирования пропущенных значений в зависимости от режима пропусков. Среди предложенных рекомендаций особенно следует отметить изменение стратегии исследования с применением квантильной регрессии на случай высокой концентрации пропусков. Кроме того, при распределениях, базирующихся на нормальном, рекомендуется использоваться базовые характеристики распределения исходной случайной величины на участках, где наблюдается сгущение пропусков. Такой подход выигрывает как в скорости обработки данных, так и в точности прогнозных значений.

Ключевой этап работы состоит в разработанных алгоритмах по улучшению точности прогнозирования пропусков и апробации полученных теоретических результатов на реальных выборках. В Главах 2 и 3 предложен подход, который базируется на установлении совместного условного распределения исходной случайной величины и пропусков, причем исследование проведено для различных типов данных – временные ряды и многомерные структуры. Отметим, что методология работы с пропусками во временных рядах ранее не предлагалась к рассмотрению. В результате в зависимости от типа данных удалось повысить точность прогнозирования пропущенных значений в несколько раз, и тем самым снизить ошибку при оценке кредитного риска.

Точность оценок при прогнозировании и скорость обработки данных – это две взаимосвязанные проблемы современного анализа данных, особенно актуальные для банковского сектора. При работе со скоринговыми моделями банкам требуется быстро адаптировать модели под регулярно поступающие данные, следовательно, появляется необходимость в новых методологиях и подходах. Таким образом, структурным подразделения коммерческих банков, занимающимся оценкой кредитных рисков при выдаче займов физическим лицам в работе с пропущенными значениями автором данного рекомендуется включить этап уточнения режима пропусков в соответствии со схемой, представленной на рисунке 2.3 Главы 2. Это позволит решить рассмотренные фундаментальные проблемы: повысить точность прогнозных значений за счет изучения поведения пропусков, а также ускорить обработку пропущенных значений, поскольку предоставляется возможность моментального определения метода для замещения пропусков, следовательно, у ЭВМ нет необходимости в переборе всех методов. Также отметим рекомендацию по ведению специального словаря моделей с ранжированием: при исследовании режимов необходимо отмечать наиболее пригодные модели и ситуации по их применению. Несмотря на то, что предложенный подход является универсальными – особенности данных могут оказывать также влияние на специфику выбора модели. Последняя рекомендация состоит в ведении внутрибанковских классификаций источников пропусков, что дает возможность заранее понять, какие данные могут быть восстановлены и какие ресурсы необходимы для выполнения данной задачи. Установив три компонента при анализе пропусков – классификация источников, словарь моделей и новый подход при восстановлении пропусков, банковские организации могут существенно оптимизировать свою деятельность в направлении анализа данных при оценке кредитного риска.

# **СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ**

1. Cox, D.R., and Hinkley, D.V. (1974). Theoretical Statistics. New York: Wiley.
2. Everitt B.S. (2002) Cambridge Dictionary of Statistics, CUP. ISBN 0-521-81099-X
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001. 533 pages.
4. Кабаков Р.И. Анализ и визуализация данных на языке R. Москва: ДМК Пресс, 2016. 587 pp.
5. R (язык программирования) [Электронный ресурс] // Википедия: [сайт]. URL: https://en.wikipedia.org/wiki/R\_(programming\_language) (дата обращения: 01.05.2018).
6. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P (2001). “A multivariate technique for multiply imputing missing values using a sequence of regression models.” Survey Methodology, 27, 85–95
7. Rubin DB (2003). “Nested multiple imputation of NMES via partially incompatible MCMC.” Statistica Neerlandica, 57(1), 3–18.
8. R Documentation and manuals [Электронный ресурс] // Rdocumentation: [сайт]. URL: https://www.rdocumentation.org/ (дата обращения: 01.05.2018).
9. О типичных банковских рисках: <Письмо> Банка России от 23.06.2004 N 70-Т
10. Козлов Д. Диверсификация модельного риска в скоринговых моделях. Агрегат-ный подход/ Д.Козлов, В. Левин//ОАО Банк ЗЕНИТ Москва .- Презентация. – 18С.
11. Андерсон К. Аналитическая культура: От сбора данных до бизнес-результатов. Отдельное издание изд. Манн, Иванов и Фербер, 2017. 336 с.
12. Chen S. Jumping the hurdle of missing risk data, Vol. |, No. 5, May 2013. pp. 1-4.
13. Florez-Lopez R. Effects of missing data in credit risk scoring. A comparative analysis of methods to gain robustness in presence of sparce data // Credit Scoring and Credit Control X, Vol. 5, No. 1, 29 2007. pp. 1-20.
14. Siddiqi N. Intelligent Credit Scoring, Building and Implementing Better Credit Risk Scorecards. Second Edition ed. New Jersey: Wiley, 2017. 464 pp.
15. Кабаков Р.И. Анализ и визуализация данных на языке R. Москва: ДМК Пресс, 2016.
16. Kaggle // Википедия: [сайт]. URL: https://en.wikipedia.org/wiki/Kaggle (дата обращения: 01.05.2018).
17. Abayomi K, Gelman A, Levy M (2008). “Diagnostics for multivariate imputations.” Journal of the Royal Statistical Society Series C: Applied Statistics, 57(3), 273–291.
18. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. — 270 с. — ISBN 5-86134-060-9
19. Hastie, Trevor. 8.5 The EM algorithm // The Elements of Statistical Learning / Trevor Hastie, Tibshirani, Friedman. — New York : Springer, 2001. — P. 236–243.
20. William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery. 2.6 Singular Value Decomposition // Numerical Recipes in C. — 2nd edition. — Cambridge: Cambridge University Press. — ISBN 0-521-43108-5.
21. Pepe, Margaret S. The statistical evaluation of medical tests for classification and prediction. — New York, NY : Oxford, 2003. — ISBN 0-19-856582-8.
22. Rubin DB (1996). “Multiple Imputation after 18+ Years.” Journal of the American Statistical Association, 91(434), 473–489.
23. Hand, David J.; Adèr, Herman J.; Mellenbergh, Gideon J. (2008). Advising on Research Methods: A Consultant's Companion. Huizen, Netherlands: Johannes van Kessel. pp. 305–332. ISBN 90-79418-01-3.
24. Samant IV UB, MacK CD, Koepsell T, Rivara FP, Vavilala MS (2008). “Time of hypotension and discharge outcome in children with severe traumatic brain injury.” Journal of Neurotrauma, 25(5), 495–502.
25. Internal ratings-based approach (credit risk) // Википедия: [сайт]. URL: https://en.wikipedia.org/wiki/Internal\_ratings-based\_approach\_(credit\_risk) (дата обращения: 01.05.2018).
26. Плэтт В. Стратегическая разведка. Основные принципы. Учебное пособие.. — М.: ФОРУМ, 1997. — 376 с. — ISBN 5-86225-600-8.
27. Creditset in CSV format // Github: [сайт]. URL: https://gist.github.com/Bart6114/8675941 (дата обращения: 01.05.2018).
28. Alexander N. Gorban, Balázs Kégl, Donald Wunsch, and Andrei Zinovyev (2008). Principal Manifolds for Data Visualization and Dimension Reduction. LNCSE 58. Springer.
29. Кнут. Э.Д. Искусство программирования. Вып. 2. Вильямс, 2017. С. 832.
30. Watcher and Trussel, The Biological Standard of Living in Comparative Perspective, 1997
31. Хардле В. Прикладная непараметрическая регрессия. Кембридж: Издательство Кембриджского университета, 1993. С. 345.
32. Labs B. Скоринг в МФО на минимальных данных // MyShared. 2010. URL: http://www.myshared.ru/slide/1283203/ (дата обращения: 21.05.2017).
33. Sturges H. (1926). The choice of a class-interval. J. Amer. Statist. Assoc., 21, 65-66.
34. The R Development Core Team. Simulating AR, MA, and ARMA Time Series, 1-5, 2003
35. Buuren S.V. Multivariate Imputation by Chained Equations // R-project. 2017. URL: https://cran.r-project.org/web/packages/mice/mice.pdf (дата обращения: 01.05.2018).
36. Mikolov T., Deoras A., Povey D., Strategies for Training Large Scale Neural Network Language Models, HLT-COE, CLSP, Johns Hopkins University, Baltimore, MD, USA, pp. 3, 2017
37. SAS Credit Scoring // Integrate scorecard development, deployment and monitoring for better decisions. 2017. URL: https://www.sas.com/content/dam/SAS/en\_us/doc/productbrief/sas-credit-scoring-100665.pdf (дата обращения: 01.05.2018).
38. Park, Sung Y.; Bera, Anil K. (2009). "Maximum entropy autoregressive conditional heteroskedasticity model". Journal of Econometrics. Elsevier. 150 (2): 219–230.
39. Shore, H (2012). "Estimating Response Modeling Methodology Models". WIREs Comp Stat. 4: 323–333.
40. Austin, M.P., and T.M. Smith, 1989. A new model for the continuum concept. Vegetatio 83
41. Koenker, Roger (2005). Quantile Regression. Cambridge University Press. ISBN 0-521-60827-9.
42. Data Enrichment // Techopedia. 2017. URL: https://www.techopedia.com/definition/28037/data-enrichment (дата обращения: 01.05.2018).
43. Cleaning Data in R [Электронный ресурс] // DataCamp: [сайт]. URL: https://www.datacamp.com/courses/cleaning-data-in-r (дата обращения: 01.05.2018).
44. Barlow, Jesse L. (1993). "Chapter 9: Numerical aspects of Solving Linear Least Squares Problems". In Rao, C.R. Computational Statistics. Handbook of Statistics. 9. North-Holland. ISBN 0-444-88096-8
45. Lawrence, Jeanette (1994). Introduction to neural networks : design, theory and applications. California Scientific Software. ISBN 1883157005. OCLC 32179420
46. A List of Common and Uncommon Types of Variables [Электронный ресурс] // Statisticshowto: [сайт]. URL: http://www.statisticshowto.com/types-variables/ (дата обращения: 01.05.2018).
47. Credit Risk Modelling for Banks - 2018 European Forum [Электронный ресурс] // Finance: [сайт]. URL: https://finance.knect365.com/credit-risk-modelling-for-banks/agenda/1
48. Scott M Zoldi, 2013. Big Data Developments in Transaction Analytics // Credit Scoring and Credit Control XIII August 28-30, 2013
49. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. — М.: Дело, 2007. — 504 с. — ISBN 978-5-7749-0473-0.
50. Wackerly, Dennis; Mendenhall, William; Scheaffer, Richard L. (2008). Mathematical Statistics with Applications (7 ed.). Belmont, CA, USA: Thomson Higher Education. ISBN 0-495-38508.

1. О типичных банковских рисках: <Письмо> Банка России от 23.06.2004 N 70-Т [↑](#footnote-ref-1)
2. Козлов Д. Диверсификация модельного риска в скоринговых моделях. Агрегат-ный подход/ Д.Козлов, В. Левин//ОАО Банк ЗЕНИТ Москва .- Презентация. – 18С. [↑](#footnote-ref-2)
3. Internal ratings-based approach (credit risk) // Википедия: [сайт]. URL: https://en.wikipedia.org/wiki/Internal\_ratings-based\_approach\_(credit\_risk) (дата обращения: 01.05.2018). [↑](#footnote-ref-3)
4. Андерсон К. Аналитическая культура: От сбора данных до бизнес-результатов. Отдельное издание изд. Манн, Иванов и Фербер, 2017. 336 с. [↑](#footnote-ref-4)
5. Samant IV UB, MacK CD, Koepsell T, Rivara FP, Vavilala MS (2008). “Time of hypotension and discharge outcome in children with severe traumatic brain injury.” Journal of Neurotrauma, 25(5), 495–502. [↑](#footnote-ref-5)
6. Hand, David J.; Adèr, Herman J.; Mellenbergh, Gideon J. (2008). Advising on Research Methods: A Consultant's Companion. Huizen, Netherlands: Johannes van Kessel. pp. 305–332. ISBN 90-79418-01-3. [↑](#footnote-ref-6)
7. Chen S. Jumping the hurdle of missing risk data, Vol. |, No. 5, May 2013. pp. 1-4. [↑](#footnote-ref-7)
8. Florez-Lopez R. Effects of missing data in credit risk scoring. A comparative analysis of methods to gain robustness in presence of sparce data // Credit Scoring and Credit Control X, Vol. 5, No. 1, 29 2007. pp. 1-20. [↑](#footnote-ref-8)
9. Siddiqi N. Intelligent Credit Scoring, Building and Implementing Better Credit Risk Scorecards. Second Edition ed. New Jersey: Wiley, 2017. 464 pp. [↑](#footnote-ref-9)
10. Rubin DB (1996). “Multiple Imputation after 18+ Years.” Journal of the American Statistical Association, 91(434), 473–489. [↑](#footnote-ref-10)
11. Cox, D.R., and Hinkley, D.V. (1974). Theoretical Statistics. New York: Wiley. pp. 1-15 [↑](#footnote-ref-11)
12. Everitt B.S. (2002) Cambridge Dictionary of Statistics, CUP. ISBN 0-521-81099-X [↑](#footnote-ref-12)
13. Pepe, Margaret S. The statistical evaluation of medical tests for classification and prediction. — New York, NY : Oxford, 2003. — ISBN 0-19-856582-8. [↑](#footnote-ref-13)
14. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001. 533 pages. [↑](#footnote-ref-14)
15. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001. 533 pages. [↑](#footnote-ref-15)
16. Кабаков Р.И. Анализ и визуализация данных на языке R. Москва: ДМК Пресс, 2016. 587 pp. [↑](#footnote-ref-16)
17. William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery. 2.6 Singular Value Decomposition // Numerical Recipes in C. — 2nd edition. — Cambridge: Cambridge University Press. — ISBN 0-521-43108-5. [↑](#footnote-ref-17)
18. Hastie, Trevor. 8.5 The EM algorithm // The Elements of Statistical Learning / Trevor Hastie, Tibshirani, Friedman. — New York : Springer, 2001. — P. 236–243. [↑](#footnote-ref-18)
19. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. — 270 с. — ISBN 5-86134-060-9 [↑](#footnote-ref-19)
20. Abayomi K, Gelman A, Levy M (2008). “Diagnostics for multivariate imputations.” Journal of the Royal Statistical Society Series C: Applied Statistics, 57(3), 273–291. [↑](#footnote-ref-20)
21. Rubin DB (2003). “Nested multiple imputation of NMES via partially incompatible MCMC.” Statistica Neerlandica, 57(1), 3–18. [↑](#footnote-ref-21)
22. R Documentation and manuals [Электронный ресурс] // Rdocumentation: [сайт]. URL: https://www.rdocumentation.org/ (дата обращения: 01.05.2018). [↑](#footnote-ref-22)
23. Kaggle // Википедия: [сайт]. URL: https://en.wikipedia.org/wiki/Kaggle (дата обращения: 01.05.2018). [↑](#footnote-ref-23)
24. R (язык программирования) [Электронный ресурс] // Википедия: [сайт]. URL: https://en.wikipedia.org/wiki/R\_(programming\_language) (дата обращения: 01.05.2018). [↑](#footnote-ref-24)
25. Кабаков Р.И. Анализ и визуализация данных на языке R. Москва: ДМК Пресс, 2016. 587 pp. [↑](#footnote-ref-25)
26. Creditset in CSV format // Github: [сайт]. URL: https://gist.github.com/Bart6114/8675941 (дата обращения: 01.05.2018). [↑](#footnote-ref-26)
27. Alexander N. Gorban, Balázs Kégl, Donald Wunsch, and Andrei Zinovyev (2008). Principal Manifolds for Data Visualization and Dimension Reduction. LNCSE 58. Springer. [↑](#footnote-ref-27)
28. Mikolov T., Deoras A., Povey D., Strategies for Training Large Scale Neural Network Language Models, HLT-COE, CLSP, Johns Hopkins University, Baltimore, MD, USA, pp. 3, 2017 [↑](#footnote-ref-28)
29. SAS Credit Scoring // Integrate scorecard development, deployment and monitoring for better decisions. 2017. URL: https://www.sas.com/content/dam/SAS/en\_us/doc/productbrief/sas-credit-scoring-100665.pdf (дата обращения: 01.05.2018). [↑](#footnote-ref-29)
30. 38. Park, Sung Y.; Bera, Anil K. (2009). "Maximum entropy autoregressive conditional heteroskedasticity model". Journal of Econometrics. Elsevier. 150 (2): 219–230. [↑](#footnote-ref-30)
31. Кнут. Э.Д. Искусство программирования. Вып. 2. Вильямс, 2017. С. 832. [↑](#footnote-ref-31)
32. Shore, H (2012). "Estimating Response Modeling Methodology Models". WIREs Comp Stat. 4: 323–333. [↑](#footnote-ref-32)
33. Watcher and Trussel, The Biological Standard of Living in Comparative Perspective, pp. 18-22, 1997 [↑](#footnote-ref-33)
34. Austin, M.P., and T.M. Smith, 1989. A new model for the continuum concept. Vegetatio 83: 35 [↑](#footnote-ref-34)
35. Хардле В. Прикладная непараметрическая регрессия. Кембридж: Издательство Кембриджского университета, 1993. С. 345. [↑](#footnote-ref-35)
36. Koenker, Roger (2005). Quantile Regression. Cambridge University Press. ISBN 0-521-60827-9. [↑](#footnote-ref-36)
37. Labs B. Скоринг в МФО на минимальных данных // MyShared. 2010. URL: http://www.myshared.ru/slide/1283203/ (дата обращения: 21.05.2017). [↑](#footnote-ref-37)
38. Data Enrichment // Techopedia. 2017. URL: https://www.techopedia.com/definition/28037/data-enrichment (дата обращения: 01.05.2018). [↑](#footnote-ref-38)
39. 43. Cleaning Data in R [Электронный ресурс] // DataCamp: [сайт]. URL: https://www.datacamp.com/courses/cleaning-data-in-r (дата обращения: 01.05.2018). [↑](#footnote-ref-39)
40. Barlow, Jesse L. (1993). "Chapter 9: Numerical aspects of Solving Linear Least Squares Problems". In Rao, C.R. Computational Statistics. Handbook of Statistics. 9. North-Holland. ISBN 0-444-88096-8 [↑](#footnote-ref-40)
41. Lawrence, Jeanette (1994). Introduction to neural networks : design, theory and applications. California Scientific Software. ISBN 1883157005. OCLC 32179420 [↑](#footnote-ref-41)
42. Sturges H. (1926). The choice of a class-interval. J. Amer. Statist. Assoc., 21, 65-66. [↑](#footnote-ref-42)
43. A List of Common and Uncommon Types of Variables [Электронный ресурс] // Statisticshowto: [сайт]. URL: http://www.statisticshowto.com/types-variables/ (дата обращения: 01.05.2018). [↑](#footnote-ref-43)
44. Credit Risk Modelling for Banks - 2018 European Forum [Электронный ресурс] // Finance: [сайт]. URL: https://finance.knect365.com/credit-risk-modelling-for-banks/agenda/1 (дата обращения: 01.05.2018). [↑](#footnote-ref-44)
45. Scott M Zoldi, 2013. Big Data Developments in Transaction Analytics // Credit Scoring and Credit Control XIII August 28-30, 2013 [↑](#footnote-ref-45)
46. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. — М.: Дело, 2007. — 504 с. — ISBN 978-5-7749-0473-0. [↑](#footnote-ref-46)
47. Buuren S.V. Multivariate Imputation by Chained Equations // R-project. 2017. URL: https://cran.r-project.org/web/packages/mice/mice.pdf (дата обращения: 01.05.2018). [↑](#footnote-ref-47)
48. Buuren S.V. Multivariate Imputation by Chained Equations // R-project. 2017. URL: https://cran.r-project.org/web/packages/mice/mice.pdf (дата обращения: 01.05.2018). [↑](#footnote-ref-48)
49. Wackerly, Dennis; Mendenhall, William; Scheaffer, Richard L. (2008). Mathematical Statistics with Applications (7 ed.). Belmont, CA, USA: Thomson Higher Education. ISBN 0-495-38508. [↑](#footnote-ref-49)
50. Barlow, Jesse L. (1993). "Chapter 9: Numerical aspects of Solving Linear Least Squares Problems". In Rao, C.R. Computational Statistics. Handbook of Statistics. 9. North-Holland. ISBN 0-444-88096-8 [↑](#footnote-ref-50)