



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

*На правах рукописи*

**Масютин Алексей Александрович**

**РАНДОМИЗИРОВАННЫЕ АЛГОРИТМЫ НА ОСНОВЕ  
ИНТЕРВАЛЬНЫХ УЗОРНЫХ СТРУКТУР ДЛЯ ЗАДАЧ  
КЛАССИФИКАЦИИ И РЕГРЕССИИ В ЗАДАЧАХ  
КРЕДИТНОГО РИСК-МЕНЕДЖМЕНТА**

**РЕЗЮМЕ**

диссертации на соискание ученой степени  
кандидата компьютерных наук НИУ ВШЭ

Москва — 2018

Диссертационная работа выполнена в Национальном исследовательском университете «Высшая школа экономики».

Научный руководитель: Сергей Олегович Кузнецов, д.ф.-м.н.,  
профессор, руководитель департамента анализа  
данных и искусственного интеллекта ФКН ВШЭ

## ТЕМА ДИССЕРТАЦИИ

В диссертации предложены алгоритмы прогноза вероятности дефолта и уровня потерь в случае дефолта, основанные на методах анализа формальных понятий. Предложенные алгоритмы с одной стороны превосходят по метрике качества работы используемые в банковской сфере стандартные модели, а с другой стороны сохраняют свойство интерпретируемости.

**Актуальность темы работы.** Развитие информационных технологий создает гораздо более жесткую конкурентную среду для банков и кредитных организаций. Например, с января 2017 года две крупнейшие российские телекоммуникационные компании начали предоставлять кредиты своим клиентам, хотя они никогда прежде не разрабатывали данное бизнес-направление<sup>1</sup>. Предоставление таких услуг со стороны нефинансовых компаний стало возможным благодаря внедрению современной ИТ-инфраструктуры для хранения большого объема данных о клиентах и использованию методов математического моделирования для оценки кредитоспособности клиентов. Международный лидер в сфере стратегических консалтинговых услуг McKinsey публикует исследования, согласно которым к 2025 году мировая банковская индустрия претерпит значительные изменения. Одной из основных причин трансформации является цифровизация банковских услуг, быстро растущий объем данных о клиентах и их операциях, появление новых типов рисков, связанных с использованием машинного обучения при принятии решений.

Математическое моделирование в банковском секторе находит одно из основных приложений в области управления рисками. Необходимым условием эффективного управления рисками является способность адекватно оценивать вероятность и величину риска. На данный момент, задачи оценки риска решаются широким спектром статистических инструментов, таких как скоринговые карты, рейтинговые модели, модели оценки уровня потерь в случае дефолта.

В то же время одной из основных причин многочисленных отзывов лицензий, среди прочего, является низкое качество кредитного

---

<sup>1</sup> <http://www.rbc.ru/finances/11/01/2017/587500529a794767fa723fa8>, имеются в виду две телеком-компании: Билайн и МТС.

портфеля; неадекватные оценки вероятности дефолта и/или величины потерь, в случае дефолта<sup>2</sup> (например, Пробизнесбанк, Татфондбанк в России). Так, в пресс-релизах Центрального Банка России, можно найти следующие комментарии: «кредитная организация неадекватно оценивала риски при неудовлетворительном качестве активов», «в результате расчета ожидаемых потерь, учитывая принятые риски, кредитная организация полностью потеряла свой капитал» и т. д.

В связи с увеличением объема данных о клиентах стандартные методы классификации и регрессии показывают меньшую точность по сравнению с более сложными алгоритмами, такими как градиентный бустинг и машины опорных векторов. Тем не менее управление рисками требует свойства интерпретируемости прогнозов, полученных на основе модели, что в случае сложных алгоритмов, как правило, невозможно. Кроме того, существуют определенные предписания Банка России, которые регулируют процесс оценки кредитного риска на основе математических моделей, и банки проходят детальные проверки использования моделей для оценки рисков, причем значительная часть проверок проводится с целью подтверждения стабильности работы модели и валидации ее бизнес-логики.

Данная работа предлагает алгоритмы решения задачи оценки риска, которые обладают свойством интерпретируемости, аналогичной ассоциативным правилам, при этом превосходят по точности обычные для банковской сферы методы классификации и регрессии, такие как скоринговые карты и деревья решений

Это достигается с помощью использования методов анализа формальных понятий и интервальных узорных структур. Было введено несколько новых определений и модификаций для существующих алгоритмов, с тем чтобы можно было осуществлять прогноз непрерывно распределенной целевой переменной на основе узорных структур и обрабатывать наборы данных со значительным числом наблюдений.

**Объект исследования** - интерпретируемые математические модели для оценки вероятности дефолта и оценки уровня потерь в случае дефолта.

---

<sup>2</sup> [http://www.cbr.ru/press/PR/?file=12082015\\_085127ik2015-08-12t08\\_46\\_23.htm](http://www.cbr.ru/press/PR/?file=12082015_085127ik2015-08-12t08_46_23.htm)

**Цель работы** - разработка методов оценки кредитоспособности и оценки уровня потерь в случае дефолта, которые обеспечивают более высокую точность по сравнению со скоринговыми картами и деревьями принятия решений при сохранении хорошей интерпретируемости. Для достижения данной цели были решены следующие **задачи**:

1. Разработана процедура рандомизированного поиска описаний на основе анализа формальных понятий, позволяющих решать задачу кредитного скоринга на основе признакового описания и сохраняющая свойства интерпретируемости при принятии решений.
2. Предложена модификация интервальных узорных структур с целью их применения к задачам прогноза уровня потерь для непрерывно распределенной целевой переменной.
3. Проведены вычислительные эксперименты как на внутрибанковских, так и на открытых данных, которые позволили найти оптимальные гиперпараметры предложенных алгоритмов и произвести сравнение со стандартными используемыми в банковской сфере алгоритмами.

## **ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ**

**Научная новизна** работы представлена двумя аспектами:

1. Разработан метод классификации «по запросу», который представляет собой рандомизированную процедуру предсказания дефолта заемщика для наборов данных с показателями финансового состояния клиента. Метод позволяет получать хорошо интерпретируемые результаты в задачах кредитного скоринга;
2. Расширены средства анализа формальных понятий для решения задачи восстановления регрессии (в случае, когда целевая имеет непрерывное распределение). Разработан алгоритм регрессии «по запросу».

Диссертация предлагает алгоритмы анализа данных, которые имеют точность, превосходящую стандартные алгоритмы, широко используемые в банковской сфере (такие как логистическая регрессия, деревья решений и скоринговые карты), сохраняя при этом свойство

интерпретируемости в том смысле, что лицо, принимающее решение, получает набор правил, релевантных для оценки кредитоспособности заемщика. Обоснование того, что методы АФП подходят для решения задач кредитного скоринга при сохранении свойства интерпретируемости, приводится в основном тексте диссертации. При этом новизна предлагаемых средств АФП заключается в следующем.

Во-первых, АФП применяется к проблеме классификации на числовых данных, причем этап построения решетки понятий пропускается (в чем и состоит концепция классификации по запросу или «ленивой» классификации). Это позволяет работать с наборами данных с произвольным числом кредитных историй, что критически важно для банков, так как массивы исторических данных достаточно велики. Кроме того, каждый заемщик с неизвестной меткой класса получает прогноз на индивидуальной основе с помощью набора правил, релевантных именно ему.

Во-вторых, вводится модификация методов АФП на основе интервальных узорных структур, что позволяет решить задачу регрессии, когда целевая переменная распределена непрерывно.

### **Основные положения, выносимые на защиту:**

1. Метод классификация «по запросу» (Query-Based Classification), который представляет собой рандомизированную процедуру предсказания неизвестной метки класса для наборов данных с большим числом наблюдений на основе интервальных узорных структур.
2. Метод регрессия «по запросу» (Query-Based Regression) который адаптирует инструментарий интервальные узорных структур для задачи восстановления регрессии, т.е. когда целевая переменная распределена непрерывно.
3. Вычислительные эксперименты, которые представляют валидацию предложенных методов, включающие сравнение с алгоритмами-аналогами как на внутрибанковских, так и на открытых данных.

Результаты получены диссертантом лично. В работах по теме диссертации диссертантом предложены ключевые научные идеи, реализованы и проведены эксперименты, написаны статьи. Вклад

остальных соавторов заключается в рецензировании программного кода экспериментов, технической помощи в постановке экспериментов, обсуждениях полученных результатов, правках текста статей, а со стороны научного руководителя, дополнительно в постановке решаемой задачи и общем руководстве исследованиями.

**Практическая значимость** подтверждена экспериментами по оценке качества работы различных алгоритмов для задач кредитного скоринга и прогнозирования уровня потерь на реальных внутрибанковских данных, а также на открытых данных. Предлагаемые методы реализованы в виде прототипа программного кода.

Предложенные методы и алгоритмы были применены в рамках пилотного проекта на наборах данных одного из топ-10 российских банков, а результаты расчетов, сравнительный анализ точности и бенчмаркинг приведены в диссертации.

Надежность полученных результатов подтверждается строгостью применения математических моделей и методов, а также путем экспериментов, сравнивающих результаты применения предлагаемых и стандартных для предметной области методов.

## **ПУБЛИКАЦИИ И АПРОБАЦИЯ РАБОТЫ**

### **Публикации повышенного уровня**

1. Masyutin A., Kashnitsky Y. Query-Based Versus Tree-Based Classification: Application to Banking Data // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, 10352 LNAI, pp. 664-673. (инд. Web of Science, Scopus)

### **Публикации стандартного уровня**

2. Masyutin A. Credit scoring based on social network data // Business Informatics, No. 3 (33), 2015, pp. 15–23. (инд. Web of Science)
3. Masyutin A. Alternative Ways for Loss-Given-Default Estimation in Retail Banking // Analysis of Images, Social Networks and Texts, 2014, Volume 436 of the series Communications in Computer and Information Science, pp. 152-162. (инд. Web of Science, Scopus)

4. Masyutin A., Kashnitsky Y., Sergei O. Kuznetsov. Lazy Classification with Interval Pattern Structures: Application to Credit Scoring, in: Proceedings of the International Workshop "What can FCA do for Artificial Intelligence?" (FCA4AI at IJCAI 2015), CEUR Workshop proceedings, Vol.1430, p. 43-54. (инд. Scopus)
5. Masyutin A., Sergei O. Kuznetsov, Continuous Target Variable Prediction with Augmented Interval Pattern Structures: A Lazy Algorithm // Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, pp.273-284. (инд. Scopus)

### **Прочие публикации**

1. Масютин А., Борисюк В., «Скоринговая карта для оценки кредитного мошенничества» // Изд-во «Риск-менеджмент в кредитной организации», №2/2015.

### **Доклады на конференциях и семинарах**

1. 3-я Международная конференция - Анализ изображений, социальных сетей и текстов, AIST'2014, Екатеринбург, Россия.

Тема: «Alternative Ways for Loss-Given-Default Estimation in Retail Banking».

2. 24-я Международная конференция по искусственному интеллекту, семинар «What can FCA do for Artificial Intelligence?», FCA4AI, IJCAI 2015, Буэнос-Айрес, Аргентина.

Тема: «Lazy Classification with Interval Pattern Structures: Application to Credit Scoring».

3. Научно-исследовательский семинар аспирантской школы ФКН ВШЭ, 12 ноября 2015.

Тема: «Интегрированное управление риском и задачи машинного обучения», (<https://aspirantura.hse.ru/cs/announcements/164947571.html>).

4. 13-я Международная конференция по решеткам понятий и их применению, CLA'2016. Москва, Россия.

Тема: «Continuous Target Variable Prediction with Augmented Interval Pattern Structures: A Lazy Algorithm».

5. Технологии баз данных, 2016, Москва, Россия ([https://www.osp.ru/iz/tbd\\_dbms](https://www.osp.ru/iz/tbd_dbms)).



Тема: «Оценка кредита на основе анализа формальных понятий: персонализированные модели».

6. 23-й Международный симпозиум по методологии построения интеллектуальных систем, ISMIS'2017, Варшава, Польша.

Тема: «Query-based versus tree-based classification: application to banking data».

7. Семинар ИССА, ФКН ВШЭ, 28 сентября 2017 года.

Тема: «Классификация по запросу на основе описаний интервалов в задачах управления кредитным риском».

## СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из 4 разделов, заключения, приложений и библиографии из 89 источников.

В **первом разделе** (введении) обосновывается актуальность темы работы, описываются существующие проблемы, ставятся задачи исследования. Также определяются цели работы, излагаются основные результаты, обсуждается теоретическое и практическое значение работы.

Во **втором разделе** описывается история математического моделирования в банковской отрасли, которое играет ключевую роль в риск-менеджменте, также описываются широко распространенные статистические алгоритмы, используемые для классификация и задач регрессии. В контексте оценки кредитного риска выделяются два параметра: вероятность дефолта (PD), уровень потерь в случае дефолта (LGD). С точки зрения машинного обучения задача оценки PD является задачей бинарной классификации, а оценка LGD – задачей восстановления регрессии. Подчеркивается компромисс между точностью прогноза и интерпретируемостью модели, поскольку некоторые регуляторы требуют от банков причины отказа заемщикам, а также центральные банки рассматривают банковские модели, требуя обоснования экономической интуиции за ними, которая в свою очередь является аргументом того, что модели будут показывать ожидаемое и стабильно качество работы.

Рассматривается метод скоринговых карт для оценки кредитоспособности заемщика, поскольку этот метод широко применяется в банковской отрасли и используется в качестве эталона для моделей «белого ящика». Рассматривается WOE-трансформация

исходных факторов модели (Weight-of-evidence), которая предназначена для адекватного учета выбросов и немонотонных зависимостей целевой переменной от значений факторов перед подачей данных в логистическую регрессию.

Модели «черного ящика» обсуждаются на примере нейронных сетей, которые выступают антагонистами интерпретируемых моделей, предоставляющих пользователю возможность понять, почему алгоритм выдает ту или иную вероятность дефолта для клиента.

В **третьем разделе** раскрывается первое нововведение: применение АФП к проблеме классификации на выборках с большим числом наблюдений. Описываются базовые термины АФП (*узурная структура, оператор пересечения, оператор Галуа, объем и содержание понятия*).

Обозначим множество объектов с положительной меткой класса  $G_+$  (множество *положительных примеров*), а множество объектов с отрицательной меткой класса:  $G_-$  (множество *отрицательных примеров*), при этом  $G_+ \cap G_- = \emptyset$ , и  $G_+ \cup G_- = G$ .

Пусть множество описаний объектов  $D$  представляет собой кортежи интервалов, т.е.  $D = \{([a_1; b_1], \dots, [a_K; b_K]) \mid \forall i: a_i, b_i \in R\}$ , где  $K$  – размерность признакового пространства. Например, для  $K=3$  элементом  $D$  может быть следующий кортеж интервалов:  $([1;2], [-0.5;0.3], [150;340])$ .

Пусть определено отображение  $\delta: G \rightarrow D$  для каждого объекта  $g \in G$ :  $\delta(g) = ([a_1; a_1], \dots, [a_K; a_K])$ , т.е. каждый объект имеет описание в виде точки в  $K$ -мерном вещественном пространстве.

Для двух описаний  $d_1, d_2 \in D$ ,  $d_1 = ([a_1; b_1], \dots, [a_K; b_K])$  и  $d_2 = ([m_1; n_1], \dots, [m_K; n_K])$  операция пересечения  $\sqcap$  определяется следующим образом:

$$d_1 \sqcap d_2 = ([\min(a_1, m_1); \max(b_1, n_1)], \dots, [\min(a_K, m_K); \max(b_K, n_K)])$$

Если  $d_1 \sqcap d_2 = d_1$ , то пишут, что  $d_1 \sqsubseteq d_2$

*Узурной интервальной структурой* называется тройка  $(G, \underline{D}, \delta)$ , где  $\underline{D} = (D, \sqcap)$ , т.е. множество объектов со множеством возможных описаний, операцией пересечения  $\sqcap$  и отображением, ставящим в соответствие объектам  $g$  из  $G$  определенные описания  $\delta(g)$  из  $D$ .

Определим также отображение из пространства объектов в пространство описаний и обратно, обозначив его символом  $\circ$  :

$$A^\circ = \bigcap_{g \in A} \delta(g) \text{ для } A \subseteq G, \\ d^\circ = \{g \in G \mid d \subseteq \delta(g)\} \text{ для } d \in D.$$

Вводятся новые определения для  $\alpha$ -слабых описаний. Описание  $d_+ \in D$  называется  $\alpha$ -слабым положительным описанием, если:

$$\frac{|d_+^\circ \cap G_-|}{|G_-|} \leq \alpha, \text{ и } \exists A \subseteq G_+ : d_+ \subseteq A^\circ$$

Описание  $d_- \in D$  называется  $\alpha$ -слабым отрицательным описанием, если:

$$\frac{|d_-^\circ \cap G_+|}{|G_+|} \leq \alpha, \text{ и } \exists A \subseteq G_- : d_- \subseteq A^\circ$$

Предлагается алгоритм классификации по запросу («ленивая классификация»). На вход алгоритма подается множество положительных примеров  $G_+$ , множество отрицательных примеров  $G_-$ , а также множество тестовых объектов  $G_{test}$  с соответствующими описаниями, определенными отображением  $\delta$ . На выходе алгоритм возвращает число  $\Delta \in R$  для каждого тестового объекта  $g_{test} \in G_{test}$ . Данное число является аналогом скорингового балла при оценке платежеспособности заемщика, т.е. на его основе возможно построение решающих правил «если  $\Delta(g_{test}) > x$ , то метка класса  $g_{test}$  положительная, иначе отрицательная». Идея алгоритма состоит в проверке для каждого тестового объекта  $g_{test}$ , является ли он похожим на объекты из множества положительных примеров  $G_+$  или множества отрицательных примеров  $G_-$ . Сходство определяется как суммарная поддержка  $\alpha$ -слабых положительных (отрицательных) описаний, содержащих описание тестового объекта. Поддержкой  $\alpha$ -слабого положительного описания  $d_+$  называется  $|d_+^\circ \cap G_+|$ , т.е. число объектов из множества положительных примеров  $G_+$ , удовлетворяющих описанию  $d_+$ . Поддержкой  $\alpha$ -слабого отрицательного описания  $d_-$  называется  $|d_-^\circ \cap G_-|$ , т.е. число объектов из множества отрицательных примеров  $G_-$ , удовлетворяющих описанию  $d_-$ . Пусть существует  $p$

штук  $\alpha$ -слабых положительных описаний и  $n$  штук  $\alpha$ -слабых отрицательных описаний, причем и первые, и последние содержат описание тестового объекта  $\delta(g_{test})$ , т.е.  $\forall i = 1, \dots, p: d_{+i} \sqsubseteq \delta(g_{test})$  и  $\forall j = 1, \dots, n: d_{-j} \sqsubseteq \delta(g_{test})$ .

Суммарной поддержкой  $\alpha$ -слабых положительных описаний называется  $P = \sum_{i=1}^p |d_{+i}^\circ \cap G_+|$ , а суммарной поддержкой  $\alpha$ -слабых отрицательных описаний называется  $N = \sum_{j=1}^n |d_{-j}^\circ \cap G_-|$ . На основе величины  $\Delta = P - N$  производится оценка того, насколько тестовый объект более похож на объекты из множества положительных или отрицательных примеров, она является аналогом скорингового балла для оценки платежеспособности заемщика. В работе также рассматриваются другие меры сходства и схемы голосования на основе  $\alpha$ -слабых описаний (см. раздел 3.4 диссертации).

Алгоритм представляет из себя итеративную процедуру и использует три гиперпараметра: *размер подвыборки* (*subsample\_size*), *количество итераций* (*number\_of\_iterations*) и *альфа-порог* ( $\alpha$ ).

Первый гиперпараметр представляет собой долю объектов, случайно извлекаемых, из множества примеров (положительных или отрицательных). На каждой итерации извлекается подвыборка объектов из  $G_+$  и  $G_-$ , и рассчитывается пересечение описаний объектов в подвыборке с описанием тестового объекта  $g_{test}$ :

$$d = \delta(g_1) \sqcap \dots \sqcap \delta(g_k) \sqcap \delta(g_{test})$$

$$\text{где } k/|G| = \text{subsample\_size}.$$

Второй гиперпараметр алгоритма представляет собой количество раз (т.е. количество итераций), которое подвыборка случайно извлекается из  $G_+$  и  $G_-$ . Данный гиперпараметр также настраивается через поиск по сетке. Третий гиперпараметр  $\alpha$  определяет, какие описания  $d$  будут считаться  $\alpha$ -слабыми и будут учтены для последующего предсказания метки класса тестового объекта. Описания, полученные в результате прохождения итераций алгоритма, но не являющиеся  $\alpha$ -слабыми не сохраняются.

Процедура повторяется для каждого тестового объекта для отдельно для множества положительных и множества отрицательных примеров, создавая множество  $\alpha$ -слабых положительных и отрицательных описаний. Конечным результатом работы алгоритма для тестового объекта, является разница между суммарной поддержкой  $\alpha$ -

слабых положительных описаний и суммарной поддержкой  $\alpha$ -слабых отрицательных описаний. На основе данной разницы по всему набору тестовых объектов рассчитывается метрика качества работы алгоритма: коэффициент Джини, т.е. мера того, насколько точно возможно предсказать метку класса для тестового объекта, зная разницу суммарных поддержек.

Алгоритм проверяется как на внутриванковских данных, так и на открытых данных Kaggle. Множество положительных примеров – это набор заемщиков, которые допустили дефолт по своему кредиту. Целевой атрибут (дефолт по кредиту) определяется как более 90 дней просрочки в течение первых 12 месяцев после выдачи кредита. Множество остальных заемщиков представляет собой множество отрицательных примеров. Каждое множество состоит из 1000 объектов. Набор тестовых данных состоит из 300 объектов и извлекается из той же генеральной совокупности, что и множества положительных и отрицательных примеров. Объясняющими переменными являются различные показатели, такие как сумма кредита, срок, процентная ставка, отношение платежа к доходу, возраст заемщика, подтвержденный/неподтвержденный документарно доход, показатели кредитной истории и т.д. Всего набор переменных, используемых для классификации, содержал 28 числовых атрибутов. Чтобы оценить точность алгоритма классификации были рассчитаны коэффициенты Джини для каждой комбинации гиперпараметров на основе 300 прогнозов на тестовом наборе. Коэффициент Джини рассчитывается на основе разницы между суммарным количеством объектов в положительных описаниях и аналогично – в отрицательных. Данная разница рассматривается как мера ранжирования заемщиков, аналогичная показателю скорингового балла. Реализован подбор гиперпараметров алгоритма по сетке.

## Коэффициенты Джини для различных значений гиперпараметров алгоритма (QBCA)

Alpha-threshold	Number of iterations	Subsample size									
		0.1%	0.2%	0.3%	0.4%	0.5%	0.6%	0.7%	0.8%	0.9%	
0.0%	100	40%	44%	39%	18%	1%	0%	0%	0%	0%	
	150	35%	46%	35%	5%	0%	0%	0%	0%	0%	
	200	42%	37%	36%	12%	5%	1%	0%	0%	0%	
	500	39%	44%	44%	25%	6%	1%	0%	0%	0%	
	1000	44%	47%	44%	41%	11%	3%	0%	0%	0%	
	2000	44%	48%	46%	36%	17%	4%	0%	0%	0%	
0.1%	100	33%	37%	40%	40%	44%	43%	34%	32%	34%	
	150	41%	34%	33%	43%	41%	47%	41%	37%	37%	
	200	40%	40%	34%	42%	51%	43%	44%	41%	36%	
	500	37%	42%	47%	49%	51%	49%	43%	41%	34%	
	1000	37%	42%	46%	48%	49%	48%	43%	43%	37%	
	2000	39%	43%	45%	49%	51%	49%	46%	41%	38%	
0.2%	100	29%	38%	42%	32%	43%	37%	46%	43%	37%	
	150	27%	42%	41%	41%	36%	47%	48%	45%	41%	
	200	32%	40%	43%	42%	42%	49%	46%	47%	48%	
	500	39%	46%	46%	48%	47%	48%	51%	48%	51%	
	1000	41%	50%	48%	47%	49%	53%	52%	52%	47%	
	2000	38%	48%	50%	48%	47%	53%	52%	53%	50%	
0.3%	100	35%	38%	39%	42%	39%	45%	34%	45%	39%	
	150	27%	43%	44%	42%	42%	39%	37%	40%	46%	
	200	34%	46%	47%	45%	49%	47%	45%	45%	52%	
	500	31%	45%	49%	50%	49%	46%	50%	51%	47%	
	1000	37%	48%	49%	49%	49%	47%	52%	51%	51%	
	2000	38%	46%	48%	51%	51%	50%	50%	52%	52%	
	5000	40%	47%	46%	51%	52%	51%	49%	51%	53%	
	10000	40%	44%	43%	46%	46%	48%	50%	52%	54%	
	20000	40%	43%	42%	46%	47%	49%	50%	52%	53%	

## Коэффициенты Джини на уточненной области поиска по сетке

Alpha-thresh-old	Number of iterations	Subsample size						
		1.0%	1.1%	1.2%	1.3%	1.4%	1.5%	
0.3%	500	51%	49%	48%	43%	41%	38%	
	1000	52%	51%	48%	45%	43%	39%	
	2000	54%	53%	49%	47%	46%	38%	
	5000	55%	52%	50%	47%	46%	40%	
	10000	56%	53%	50%	47%	47%	40%	
	20000	55%	53%	51%	46%	48%	41%	

Алгоритм классификации по запросу в сравнении с классическими моделями, принятыми в банках и другими бенчмарками на внутрибанковских данных

	Gini on test sample
Logistic regression	47.38%
Scorecard (Logistic based on WOE-transformation)	51.89%
CART (minsize= 50)	54.75%
MLCA (s = 1%, a=0.3%, n=10000)	56.30%
AdaBoostClassifier	54.72%
KNeighborsClassifier	44.00%
NaiveBayes	48.91%
RandomForestClassifier	53.42%

Алгоритм был также протестирован на открытых данных платформы Kaggle конкурса 2012 года «Give me some credit»<sup>3</sup>. Данные имеют бинарную целевую переменную (метка класса) в зависимости от того, был ли заемщик дефолтным или нет.

Алгоритм классификации по запросу и бенчмарки для открытого набора данных Kaggle

metric \ algo	Scorecard	QBCA	Xgboost
Valid. Gini	0.5806	0.6624	0.708

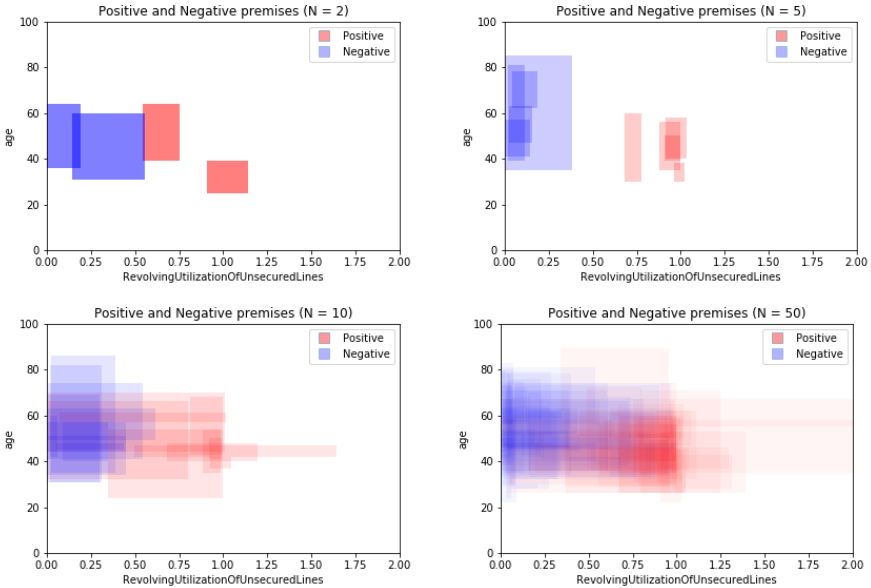
Проводится анализ чувствительности точности работы алгоритма в зависимости от настроек гиперпараметров.

Также представлена визуализация коллекций  $\alpha$ -слабых описаний, которая позволяет интерпретировать результат модели для клиента. При рассмотрении метки целевого класса тестового объекта (хорошего

<sup>3</sup> <https://www.kaggle.com/c/GiveMeSomeCredit>

или плохого) алгоритмы строят портреты хороших и плохих клиентов по историческим данным в многомерном пространстве признаков.

Ниже приведены несколько примеров двухсторонних областей для разных значений числа итераций:



Положительные описания изображены красным цветом, а отрицательные - синим. Приводится обоснование, что такой набор областей в пространстве признаков дает понимание того, почему конкретный заемщик считается высоко рискованным или низкорискованным с точки зрения кредитования.

**Четвертый раздел** содержит второе нововведение: адаптация АФП к регрессионной задаче (т.е. для случая непрерывно распределенной целевой переменной). Чтобы сделать методы АФП применимыми к этому случаю, вводится определение *расширенной интервальной узорной структуры*.

Определим расширенную интервальную узорную структуру как четверку  $(G, D, \delta, h)$ , где  $G$  – множество объектов,  $D$  – множество



возможных описаний объектов,  $d \in D$ . Описание  $d$  для нашей предметной области представляет собой кортеж интервалов, в котором теперь выделяются два элемента  $d_x$  и  $d_y$  ( $d_y$  – интервал значений для *целевого атрибута*  $y \in R$ , а  $d_x$  – кортеж интервалов значений для *объясняющих атрибутов*  $x$ , на основе которых прогнозируется величина  $y$ ). Также как в разделе 3 используется отображение  $\delta: G \rightarrow D$  и дополнительно к нему эмпирическая функция плотности  $h \in H$ , где  $H$  – семейство функций плотности для целевого атрибута. Мы также будем использовать обозначения  $\delta_x$  и  $\delta_y$ , чтобы разделять описания, содержащие объясняющие атрибуты и целевой атрибут соответственно. Определение оператора пересечения  $\Pi$  остается неизменным.

Пусть дано произвольное множество объектов  $A_0 \in G$ , т.е.:

$$A_0 = \{g_1, g_2, \dots, g_J\},$$

$$\delta(g_j) = (\delta_x, \delta_y) = ([x_{1j}; x_{1j}], \dots, [x_{Kj}; x_{Kj}], [y_j; y_j])$$

для  $j = 1, \dots, J$

где  $J$  – число объясняющих атрибутов. Отображение  $^\circ$  определим следующим образом:

$$A_0^\circ = (d_0, h_0)$$

где  $d_0 = \{d_{x0}, d_{y0}\}$  и  $d_{x0} = \delta_x(g_1) \Pi \dots \Pi \delta_x(g_J)$ , а описание целевого атрибута есть  $d_{y0} = \delta_y(g_1) \Pi \dots \Pi \delta_y(g_J)$ , что на самом деле является интервалом  $[y_{min}, y_{max}]$ , а  $h_0$  – отображение  $d_{y0} \rightarrow [0; 1]$ , т.е. эмпирическая функция плотности, построенная по наблюдениям значений целевого атрибута в  $A_0$ :

$$h([\tau_{i-1}, \tau_i]) = \frac{\sum_{g \in A} 1_{[\tau_{i-1}, \tau_i] \subseteq \delta_y(g)}}{|A|}, \forall i = 1, \dots, K$$

где  $\tau_0 = y_{min}, \tau_K = y_{max}$ , и  $\Delta\tau_i = \tau_i - \tau_{i-1} = \frac{y_{max} - y_{min}}{K}$ ,  
 $1$  – функция-индикатор.

Мы будем использовать композицию отображения  $^\circ$  похожим образом, как это делается, в обычных интервальных узорных структурах, но определение множества объектов, соответствующих описанию, будет производиться только по той части описания, которая относится к объясняющим атрибутам:

$$A_0^{\circ\circ} = (d_0, h_0)^\circ \stackrel{\text{def}}{=} d_{x0}^\circ = A_1$$

Для того, чтобы перейти к итоговому прогнозу целевого атрибута полезно ввести определение  $\alpha$ -слабого описания с  $\omega$ -допустимым выбросом. Расширенная интервальная узорная структура  $d = (d_x, d_y) \in D$  называется  $\alpha$ -слабым описанием с  $\omega$ -допустимым выбросом, если:

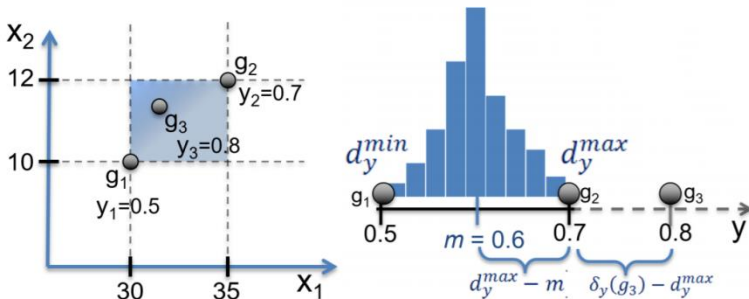
$$1 - \frac{|\{g \in A \mid d_y^{min} - \omega(m - d_y^{min}) \leq \delta_y(g) \leq d_y^{max} + \omega(d_y^{max} - m)\}|}{|A|} \leq \alpha$$

где  $\delta_y(g)$  – значение целевого атрибута для объекта  $g$ , множество  $A = d_x^*$ ,  $d_y$  – интервал  $[d_y^{min}; d_y^{max}]$  для целевого атрибута, а  $m$  – медиана эмпирической функции плотности  $h$ , построенной на интервале  $d_y$  на основе значений целевого атрибута среди объектов из  $A$ .

Рассмотрим пример использования введенных понятий.

Пусть множество объектов есть  $G = \{g_1, g_2, g_3\}$  описания которых имеют два объясняющих признака  $x_1, x_2$  и один целевой атрибут  $y$ :

Объекты\Атрибуты	$x_1$	$x_2$	$y$
$g_1$	30	10	0.5
$g_2$	35	12	0.7
$g_3$	31.5	11.5	0.8



Пусть  $A_0 = \{g_1, g_2\}$ .

Тогда  $\delta_x(g_1) = ([30; 30], [10; 10])$ ,  $\delta_y(g_1) = [0.5; 0.5]$

$\delta_x(g_2) = ([35; 35], [12; 12])$ ,  $\delta_y(g_2) = [0.7; 0.7]$

$d_0 = (d_{0x}, d_{0y})$

$$d_{0x} = \delta_x(g_1) \sqcap \delta_x(g_2) = ([30; 35], [10; 12])$$

$$d_{0y} = \delta_y(g_1) \sqcap \delta_y(g_2) = [0.5; 0.7]$$

$$h_0 = \{0.5, 0.7\}$$

$$A_0^\circ = (d_0, h_0)$$

$$A_0^{\circ\circ} = (d_0, h_0)^\circ = d_{0x}^\circ = A_1 = \{g_1, g_2, g_3\}$$

$$d_1 = ([30; 35], [10; 12], [0.5; 0.8])$$

$$h_1 = \{0.5, 0.7, 0.8\}$$

$$A_0^{\circ\circ\circ} = A_1^\circ = (d_1, h_1)$$

Описание  $d_0 = ([30; 35], [10; 12], [0.5; 0.7])$  является 1/3-слабым описанием с 1-допустимым выбросом, т.к. медиана 0.5 и 0.7 равняется 0.6.

Первым этапом регрессии «по запросу» (QBRA) является поиск  $\alpha$ -слабых описаний с  $\omega$ -допустимым выбросом, вторым - предсказание для тестового объекта на основе найденных описаний. Размер подвыборки – это гиперпараметр, который представляет собой количество объектов, которые случайным образом извлекаются из множества объектов  $G$ . Далее фиксируются  $\alpha$  и  $\omega$ . После вычисляется описание  $d_0 = \delta(g_1) \sqcap \dots \sqcap \delta(g_k) \sqcap \delta(g_t)$  и эмпирическая функция плотности для значений целевого атрибута. Если  $d_0$  является  $\alpha$ -слабым описанием с  $\omega$ -допустимым выбросом, то оно добавляется к коллекции описаний, которые будут использоваться для прогнозирования. После завершения поиска описаний проходит следующий этап, который формирует прогноз целевого атрибута на основе найденных описаний. Итоговое предсказание было определено как медиана смеси распределений целевого атрибута из всех  $\alpha$ -слабых описаний с  $\omega$ -допустимым выбросом.

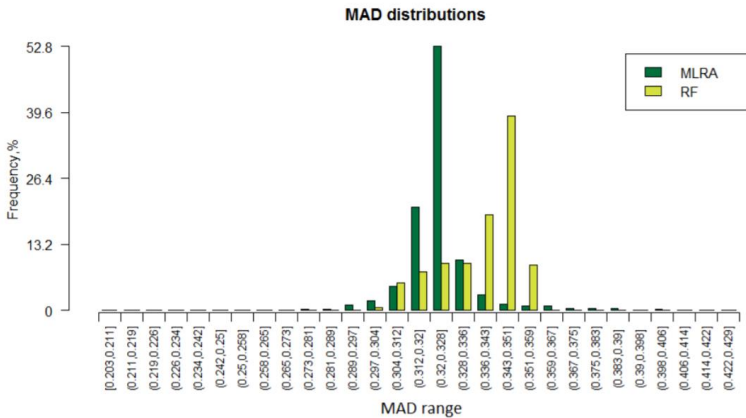
Для оценки качества работы алгоритма были использованы финансовые данные балансов и отчетов о прибылях и убытках 612 корпоративных клиентов одного из топ-10 крупнейших российских банков. Среди прочих факторов мы использовали отношение активов к обязательствам, отношение долга к собственному капиталу, прибыль до налогов и процентных платежей, доходность активов и т.д. Эти клиенты были оценены во время ранних сигналов о неплатежеспособности, и была собрана статистика по проценту возвращенной задолженности.

Точность предсказания была оценена с точки зрения среднего абсолютного отклонения (mean absolute deviation - MAD):

$$MAD = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N}$$

где  $y_i$  является целевым атрибутом (уровень возврата долга) для  $i$ -го объекта из тестового множества и  $\hat{y}_i$  является прогнозом для  $y_i$ . Алгоритм сравнивался с моделями случайных лесов и некоторыми другими стандартными методами.

Распределение  $MAD$  показывает, что алгоритм «ленивой» регрессии позволяет получить более низкую ошибку прогноза, чем у моделей случайных лесов.



Распределения представляют собой величину MAD, достигаемую для большего числа запусков алгоритма с различными комбинациями значений гиперпараметров.

Другие бенчмарки приведены ниже:

	MAD value
Naive model (median value for all test objects)	0.35
LinearRegression	0.30
DecisionTreeRegressor	0.28
Random Forest	0.24
AdaBoostRegressor	0.28
GradientBoostingRegressor	0.25

В заключении подводятся итоги исследования, и подчеркивается, что ключевым требованием практического

моделирования в риск-менеджменте выступает интерпретируемость, независимо от точности модели. АФП предлагает привлекательные инструменты для извлечения знаний из данных, поскольку извлеченные описания могут рассматриваться как посылки ассоциативных правил. Кроме того, результаты экспериментальных вычислений показывают, что предлагаемые рандомизированные алгоритмы для задач кредитного скоринга и прогноза уровня потерь превосходят стандартные методы, используемые в банках, такие как скоринговые карты и деревья решений в терминах Джини и среднего абсолютного отклонения. Предлагаемые методы классификации и регрессии могут конкурировать со стандартными статистическими процедурами, принятыми в банковской сфере, предоставляя интерпретируемые наборы правил для принятия решений относительно конкретного заемщика.

В приложении приведен программируемый код для QVCA и QVRA. Предоставляются некоторые ключевые функции для расчета оператора пересечения, расчета объема и содержания понятия, поиска альфа-слабых описаний и схем голосования для построения окончательного прогноза. Язык, на котором предоставлен код, - R (<https://www.r-project.org/>) поскольку он обладает интуитивным синтаксисом, так что идея реализации алгоритма удобно преподносится читателю. Однако для производственных реализаций рекомендуется использовать другие языки, такие как Java или Spark (для распределенных систем).

### **Результаты**

1. Разработаны методы рандомизированного поиска правил для задачи классификации на основе АФП.
2. Введено понятие альфа-слабого описания и других гиперпараметров алгоритмов классификации и регрессии «по запросу». Проведен анализ точности прогнозов с учетом значений гиперпараметров алгоритма, настроенных на данных для кредитного скоринга и прогноза уровня потерь в случае дефолта.
3. Разработан алгоритм, позволяющий использовать технику интервальных узорных структур в задаче регрессии с несколькими гиперпараметрами: количество итераций, альфа-порог, размер подвыборки, величина  $\omega$ -допустимого выброса, штраф за высокое

значение дисперсии целевой переменной в правой части расширенной узорной структуры (штраф за высокое отклонение).

4. Предложено понятие *расширенной интервальной узорной структуры*, введено понятие  $\omega$ -допустимого выброса для  $\alpha$ -слабых описаний. Новые определения позволяют решать задачу восстановления регрессии с помощью методов АФП.

5. Предложенные алгоритмы проанализированы с точки зрения интерпретируемости решения. Произведен сравнительный анализ точности работы предложенных алгоритмов с моделями кредитного скоринга и моделями других типов.

6. Разработан алгоритм классификации «по запросу» с тремя гиперпараметрами: *число итераций, альфа-порога, размер подвыборки*. Предложен анализ точности предсказаний алгоритма в зависимости от значений гиперпараметров, предложено интуитивное объяснение полученных результатов.

7. Предложенные алгоритмы реализованы в виде программной системы на языке R.