ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

**Федеральное государственное автономное образовательное учреждение  
высшего профессионального образования**

**Национальный исследовательский университет  
«Высшая школа экономики»**

**Московский институт электроники и математики**

**Факультет Прикладной математики и кибернетики**

**Кафедра Высшей математики**

Дипломная работа

по специальности 230401.65 «Прикладная математика»

**Скоринг и задача Монжа-Канторовича.**

Студент группы М95 Алифанов Д.В.

Руководитель профессор, д. ф.-м. н.

Хаметов В.М.

Зав. кафедрой доцент, к. ф.-м. н.

Кузьмина Л.И.

Москва 2014

# Аннотация

Данная работа посвящена скорингу. В ней приводится обзор методов построения скоринговых моделей. Предложены два метода построения таких моделей:

1. эмпирический байесовский,
2. основанный на решении задачи Монжа-Канторовича.

На примере данных по потребительскому кредитованию, предоставленных банком “Сбербанк России” построены рейтинги его клиентов.

**Abstract**

Presented work describes scoring. Various methods, applied in creating scoring models, characterized. Two methods offered as a basis for scoring models making:

1. empirical Bayes method
2. new method, based on Monge–Kantorovich transportation problem

With actual consumer credit “Sberbank of Russia OJSC” data calculated credit ratings.

# Содержание

[Аннотация 2](#_Toc389856330)

[Содержание 3](#_Toc389856331)

[Введение 4](#_Toc389856332)

[Глава 1. Методы построения скоринговых моделей. 6](#_Toc389856333)

[1.1. История появления и развития скоринга. 6](#_Toc389856334)

[1.2. Линейный дискриминантный анализ и линейная регрессия. 7](#_Toc389856335)

[1.3. Логистическая регрессия и пробит-регрессия. 9](#_Toc389856336)

[1.4. Нейронные сети. 10](#_Toc389856337)

[1.5. CHAID анализ. 11](#_Toc389856338)

[1.6. Другие методы. 12](#_Toc389856339)

[1.7. Сравнение различных методов. 14](#_Toc389856340)

[Глава 2. Эмпирический Байесовский подход. 16](#_Toc389856341)

[2.1. Постановка задачи. 16](#_Toc389856342)

[2.2. Построение скоринговой модели. 17](#_Toc389856343)

[2.3. Применение модели к данным. 19](#_Toc389856344)

[Глава 3. Подход, основанный на задаче Монжа-Канторовича. 32](#_Toc389856345)

[3.1. Задача Монжа-Канторовича. 32](#_Toc389856346)

[3.2. Применение задачи Монжа-Канторовича к скорингу. 35](#_Toc389856347)

[Заключение 40](#_Toc389856348)

[Список литературы 41](#_Toc389856349)

# Введение

Скоринг – это эвристический способ построения рейтингов и классификации различных объектов на группы. Он основывается на предположении о том, что люди со схожими социальными показателями ведут себя одинаково. Он применяется в банковской сфере, маркетинге, страховом деле.

Основной целью традиционного скоринга является классификация клиентов банка на “хороших” и “плохих”, исходя из которой кредитор может выбирать соответствующие действия по отношению к данному клиенту. “Плохого” клиента, к примеру, можно определить как клиента с низкой эмпирической вероятностью возвращения кредита. Но, как правило, такое определение “плохого” клиента расширяется до любого нежелательного банку поведения клиента. Классификация осуществляется на основе скоринговой карты с помощью которой рассчитывается скоринговый балл клиента.

Из литературы, посвященной скорингу, отметим несколько работ.

“Руководство по кредитному скорингу” под редакцией Элизабет Мейз, 2008 г. – единственная книга о скоринге на русском языке. Описываются общие понятия, разбираются методы построения скоринговой карты, обсуждается применение скоринга на практике. Книга состоит из статей написанных зарубежными специалистами в области финансов.

Диссертационная работа Сэмюэла Глассона “Метод цензурированной выборки для кредитного скоринга”, 2007 г. В ней исследуются инструменты анализа выживаемости, применительно к кредитному скорингу, в условиях цензурированных данных. Разбирается применение метода линейной регрессии и в частности метода Бакли-Джеймса. Практическая часть работы содержит в себе применение этих методов к оценке времени кредитного дефолта и времени выплаты очередного платежа.

Диссертационная работа Кристины Болтон “Логистические регрессии и их применение в кредитном скоринге”, 2009 г. Разбирается концепция кредитного скоринга применительно к банковскому делу в Южной Африке. Рассматриваются методы построения скоринговой модели с особым акцентом на метод логистической регрессии. Применяется этот метод для создания скоринговой модели.

Диссертационная работа Маттиаса Кремпля “Адаптивные модели и их применение в кредитном скоринге”, 2011. Акцент ставится на изучении методов построения предсказывающих моделей в условиях дрейфа и задержки данных. Представлен новый метод для построения скоринговых моделей, базирующийся на методе дерева принятия решений. Представленный метод применяется для оценки дрейфа в двух наборах реальных финансовых данных.

В приведённых выше работах имеется общая проблема: применение методов к построению скоринговых моделей не обосновано. Отсюда вытекает сомнение в правильности полученных данных. Задача состоит в построении метода, применение которого было бы обосновано. В данной работе представляется метод, решающий эту задачу.

Существует множество подходов к построению скоринговой модели. В главе 1 данной работы описаны методы применяемые в вышеприведённых диссертациях. В главе 2 вводится математическая модель скоринга и разбирается эмпирический Байесовский подход к построению скоринговой модели: подход описывается теоретически, а после применяется к реальным данным банка “Сбербанк России” для построения скоринговой модели. В главе 3 вводится метод, основывающийся на задаче Монжа-Канторовича. Приведено теоретическое обоснование использования данного метода. Затем он применяется для построения скоринговой модели на данных используемых в главе 2.

# Глава 1. Методы построения скоринговых моделей.

## 1.1. История появления и развития скоринга.

Изначально скоринг разрабатывался с целью автоматизации процесса решения о выдаче кредита. До внедрения скоринга, решение о том, кому выдать кредит в каком размере, принималось кредитным экспертом. Он решал это, опираясь на опыт и собственное мнение, руководствуясь параметрами клиента, влияющими на его кредитоспособность.

В 1940-х годах началось внедрение скоринговых систем. В 1941 Давид Дюран опубликовал первую исследовательскую работу по кредитному скорингу, в которой оценивал роль различных факторов в прогнозирующей системе. После окончания Второй мировой войны, резко вырос спрос на кредитные продукты, и стало ясно, что традиционные методы принятия решения плохо работают в условиях большого числа клиентов. Взрыв спроса на кредиты, отчасти обусловленный введением кредитных карт, мотивировал кредиторов внедрять автоматизированные системы принятия решения о выдаче кредитов. Параллельное развитие вычислительной техники, способствовало этому и давало возможность обрабатывать большие массивы финансовых данных.

В 1956 году была создана компания FICO, занимавшаяся разработками в области потребительских кредитов. В 60-х годах началось внедрение компьютерных технологий в область скоринга. В 1963 году было предложено использование дискриминантного анализа данных для кредитного скоринга. И, наконец, в 1975 с принятием "US Equal Credit Opportunity Act I", скоринг был окончательно признан.

Важным шагом в развитии кредитного скоринга было появление скоринга поведения (behavior scoring) в начале 90-х. Его целью является предсказание выплат уже имеющихся клиентов.

В недавнее время развитие скоринговых систем было обусловлено регулированием со стороны внешней среды. Как часть требований к достаточности капитала, предъявляемых банкам в связи с вступлением в силу второго Базельского соглашения (Basel Committee for Banking Supervision 2001), учреждения должны внимательно следить за рисками, связанными с их кредитными портфелями. Методы кредитного скоринга позволяют это делать.

С момента введения первой скоринговой системы, использовалось множество математических и статистических методов. Среди статистических можно назвать: дискриминантный анализ, линейную регрессию, логистическую регрессию и дерево принятия решений. Другие методы пришли из математики: математическое программирование, нейронные сети, генетические алгоритмы и экспертные системы. Далее мы разберём наиболее распространённые методы и поговорим об их достоинствах и недостатках.

## 1.2. Линейный дискриминантный анализ и линейная регрессия.

Линейный дискриминантный анализ – метод для классификации объектов на заранее определённые категории. Идея в том, чтобы найти такую линейную комбинацию объясняющих переменных, которая наилучшим образом разделила бы объекты на категории. Под разделением наилучшим образом имеется ввиду такое, при котором обеспечивается максимальная дистанция между средними данных категорий. Скоринговый балл рассчитывается как линейная функция от значений атрибутов клиента:

.

Здесь,  - значения атрибутов клиента,  - параметры модели, которые максимизируют отношение

,

где  - вектор средних для *хороших* и *плохих* клиентов, - общая ковариационная матрица.

Линейный дискриминантный метод предполагает выполнение двух условий. Во-первых, ковариационные матрицы независимых переменных для обеих групп должны совпадать. Во-вторых, независимые переменные должны быть распределены нормально. Часто, в скоринге, независимые переменные дискретные или распределены не нормально. Отсюда, возникают проблемы в применении этого метода. Однако было показано, что даже в случае нарушения нормальности, данный метод широко применим. Его преимуществом можно назвать простоту применения.

Схожий метод линейной регрессии, также используется для формирования скоринговой модели. В случае двух категорий, он эквивалентен методу линейного дискриминантного анализа и выражает зависимость одной переменный (зависимой) от других (независимых). В общем виде представляется так:



 - зависимая переменная;

 - объясняющие независимые переменные;

 - неизвестные коэффициента регрессии, которые находятся методом наименьших квадратов;

 - ошибка.

Для применения модели линейного скоринга требуется выполнение следующего предположения: связь между зависимой и независимыми переменными должна быть линейной. В противном случае, точность оценки значительно ухудшается. Ошибки же должны быть независимы и распределены нормально.

Как и в случае дискриминантного анализа, в условиях кредитного скоринга, предположения, требуемые для применения линейной регрессии, нередко нарушаются. Линейная регрессия может дать оценку вероятности вне диапазона [0,1], что является неприемлемым. К примеру, логистическая регрессия лишена этого недостатка.

## 1.3. Логистическая регрессия и пробит-регрессия.

Данные виды регрессии больше подходят для построения скоринговой модели, так как допускают категорийное представление данных. Модель логистической регрессии задаётся следующим образом:



где  - оценка вероятности того, что клиент “плохой”*,  -* вектор неизвестных параметров регрессии, который вычисляется через условие максимизации отношения правдоподобия.

Модель логистической регрессии базируется на функции логарифм. В свою очередь, пробит-регрессия базируется на нормальном распределении и задаётся следующим образом:

,

где . Вектор  находится также как и в модели логистической регрессии.

Так как логистическая регрессия и пробит-регрессия используют схожие по форме распределения, результаты применения данных моделей также схожи. Логистическая регрессия пользуется большим предпочтением, так как вычисления проще, чем в пробит-регрессии и имеется больше инструментов для работы с ней. За счёт своей бинарной природы, логистическая регрессия предпочтительней линейной регрессии в использовании для построения скоринговых моделей. На практике же было выяснено, что разница в точности предсказываемых результатов незначительна. Тем не менее, наблюдается преобладание логистической регрессии в скоринговых системах.

## 1.4. Нейронные сети.

Искусственные нейронные сети являются симуляцией нейронных сетей имеющихся в природе. Возникло это понятие при попытке смоделировать процессы, происходящие в мозге человека.

Нейронные сети, также называемые многослойным перцептроном, особенно подходят для решения задачи классификации. Они широко используются в различных сферах: финансах, компьютерных науках, физике и медицине. Популярность нейронных сетей отчасти обуславливается возможностью моделировать сложные ситуации без особых затрат со стороны использующего этот метод. По своей природе нейронные сети автоматически обнаруживают любую нелинейную ситуацию в данных и подстраиваются под неё. Также многослойные нейронные сети являются универсальными аппроксиматорами, то есть могут аппроксимировать любую функцию сколь угодно точно.

Нейронные сети состоят из слоев которые, в свою очередь, состоят из узлов. Есть 3 типа слоёв в сетях: входной, скрытые, выходной. Входной слой образуют атрибуты клиента, такие как пол, возраст и т.п.

Выход  для k-го узла с m входами представляется так:

,

где  - активационная функция,  - вектор входных данных,  - весовой вектор который обозначает силу связи между узлами.

Основным недостатком является то, что не смотря на возможность добиться высокой точности прогноза, понять причины, по которым было принято то или иное решение, невозможно.

В контексте кредитного скоринга было показано, что нейронные сети работают не хуже традиционных методов.

## 1.5. CHAID анализ.

Данный метод отлично подходит для нахождения связей между данными, особенно если связи нелинейные. Он применяется для построения деревьев принятия решений, и имеет много общего с классическими методами, такими как дискриминантный анализ и линейная регрессия.

Аббревиатура CHAID расшифровывается как Chi-squared Automated Interaction Detector.

Гибкость данного метода делает его привлекательным для использования, но это не означает, что его стоит использовать вместо традиционных методов. В случае, когда встречаются строгие теоретические предположения о распределении, традиционные методы предпочтительней. Как техника исследования или в случае, когда традиционные методы не срабатывают, CHAID анализ является непревзойдённым инструментом.

CHAID строит не бинарные деревья (т.е. деревья у которых может быть более двух ветвей) на основе относительно простого алгоритма, который особенно хорошо подходит для анализа больших массивов данных. Алгоритм основывается на применении теста хи-квадрат.

## 1.6. Другие методы.

*Дерево принятия решений.*

Метод разделяет данные на подмножества, каждое из которых более однородно в своем поведении, нежели исходное множество данных. Каждое из этих подмножеств делится далее, по такому же алгоритму. Результат деления именуется «листом» это дерева. Имеются и другие методы, работающие по схожему принципу.

Достоинства этого метода – простота и интуитивность. Метод способен работать с отсутствующими наблюдениями. Особенно он применим в случае, когда о данных до их исследования практически ничего неизвестно и нельзя построить какие-либо догадки или гипотезы.

Главный недостаток этого метода – сложность компьютерных расчетов. Вследствие громоздкости получаемых деревьев, процесс изучения модели трудоёмкий. Изменения в ситуации может привести к пересмотру всего дерева решений.

В основном метод используется как вспомогательный. К примеру для определения переменных, которые наиболее сильно объясняют поведение зависимой переменной.

*Метод k ближайших соседей.* Непараметрический метод классификации объектов. Основывается на метрике, определяющей схожесть между данными.

Первоначально вводятся тренировочные данные, разделенные на классы. Затем вводятся оцениваемые данные и определяется схожесть между введёнными и тренировочными данными. На основе метрики выбирается k ближайших соседей. Новый элемент относят к тому классу, к которому принадлежит большинство его соседей.

Количество соседей k определяется компромиссом между компенсацией и дисперсией. Чем меньше класс, тем меньше выбирается k. При этом необязательно, что при больших k результат будет лучше.

Одно из преимуществ данного метода – легко добавить новые данные, не изменяя при этом модель. Непараметрическая сущность этого метода позволяет работать с иррациональностями в функциях риска на пространстве признаков.

Отсутствие формального метода для выбора k и невозможность вероятностной интерпретации результата, так как результатом являются ожидаемые частоты, являются главными недостатками метода. Данные сложности могут быть решены использованием методом Байесовской аппроксимации.

Данный метод мало используется в скоринге. Одной из причин этого является то, что для классификации одного объекта необходимо иметь базу по всем объектам.

Более новый *метод опорных векторов*, построенный на машинном обучении, показал себя не хуже традиционных скоринговых методов. Он состоит из двух процессов: первый преобразует входные данные к данным высокой размерности в пространстве признаков; второй классифицирует данные с помощью линейного классификатора. Классификатором может выступать, например, линейный дискриминантный анализ.

## 1.7. Сравнение различных методов.

Был проведён ряд сравнительных исследований для скоринговых методов. Критериями для ранжирования служили процент ошибок при классификации и ROC-кривая. Исследовались 8 наборов данных.

|  |  |
| --- | --- |
| Метод | Средняя оценка |
| Нейронные сети | 3.2 |
| Опорных векторов | 3.3 |
| Логистическая регрессия | 4.3 |
| Линейный дискриминантный анализ | 5.3 |
| Линейные LS-SVM | 5.5 |
| Расширенное дерево Байеса | 5.6 |
| Наивный байесовский классификатор | 7.8 |
| Радиально базисные функции | 9.1 |
| k-ближайших соседей (k=100) | 9.5 |
| Линейный SVM | 10.1 |
| Квадратичный дискриминантный анализ | 10.8 |
| Дерево принятия решений | 10.8 |
| Линейное программирование | 11.9 |
| Дерево принятия решений | 13.0 |
| Дерево принятия решений | 13.7 |
| k-ближайших соседей (k=10) | 14.1 |
| Дерево принятия решений | 14.7 |

Из таблицы видно, что нейронные сети и метод опорных векторов явились наилучшими на исследуемых 8 наборах данных. Кроме того традиционные методы, такие как линейный и дискриминантный анализ показали себя конкурентоспособными. Отсюда следует, что, вероятно, большинство данных для кредитного скоринга лишь немного нелинейны. Вследствие чего линейные методы показали себя на уровне с нелинейными.

Не существует оптимальной скоринговой модели для любой ситуации. Выбор модели зависит от данных и цели, на которую направленно создание модели. Кроме того, метод, оценивающий наилучшим образом, не обязательно будет лучшим в данной ситуации.

# Глава 2. Эмпирический Байесовский подход.

В данной главе разберём эмпирический Байесовский подход и с помощью него построим скоринговую модель. Построение будем вести исходя из статистики по потребительским кредитам банка “Сбербанк России”.

## 2.1. Постановка задачи.

Предположим, имеется некий банк, занимающийся кредитованием частных лиц. В банк за получением кредита обращаются клиенты. Решение о выдаче кредита банк выносит на основе информации о клиенте.

Информацию о клиенте банк получает из разных источников: от самого клиента, от кредитного бюро и из других источников. Мы будем рассматривать информацию, предоставляемую самим клиентом. Банк получает её через заполненную заемщиком анкету.

В анкете заемщик указывает следующие данные: пол, возраст, семейное положение, наличие детей, ежемесячный доход, наличие недвижимости и прочее.

На основании этих данных разобьём клиентов на группы, в которых они схожи по определённым признакам. Для каждого клиента Байесовским методом найдём рейтинг – эмпирическая вероятность того, что клиент вернёт кредит при условии, что он принадлежит данной группе.

Найдя распределение рейтингов, мы тем самым построим скоринговую модель.

Для применения метода необходимо, чтобы данные удовлетворяли следующим условиям:

* независимость – клиенты не имеют сговора по выплате кредита;
* однородность – данные взяты из одной генеральной совокупности;
* равновероятность – клиенты равновероятно распределяются по группам .

Их выполнение проверяется ниже.

## 2.2. Построение скоринговой модели.

Введем вероятностное пространство . Обозначим в этом пространстве  – клиент банка.

Каждый клиент банка имеет набор характеристик согласно заполненной анкете. Например: в браке или нет, уровень дохода, разбитый по категориям, наличие машины и прочие характеристики. Согласно этим характеристикам введём разбиение пространства  на множества 

1. ;
2. ;
3. ;
4. .

Таким образом, множество клиентов разбито на группы .

Введём случайные величины.

 – количество клиентов в j-й группе.



Исходя из данных, мы можем построить совместное эмпирическое распределение вероятностей .

Здесь и далее, записывая , будем иметь ввиду эмпирическую вероятность события A определяемую как отношение количества исходов удовлетворяющих событию A к общему количеству исходов.

Далее мы применим эмпирический Байесовский подход, который основывается на формуле Байеса

 ,

где  – априорная эмпирическая вероятность события A,

 – эмпирическая вероятность события B при условии А,

 – эмпирическая вероятность события A при условии B, которую называют апостериорной вероятностью,

 – эмпирическая вероятность события B.

Данная формула позволяет переоценить вероятность события A, учитывая тот факт, что произошло событие B.

Из определения условной вероятности можем записать:

 (1)

Далее заметим, что .

Выразив из (1)  и подставив в эту формулу выражение для , получим:



Выраженная величина является рейтингом клиента из j-го множества. Исходя из её значений, осуществляется классификация на “хороших” и “плохих”.

## 2.3. Применение модели к данным.

В нашем распоряжении имеются данные о 1977 клиентах банка “Сбербанк России” включаю информацию о том, вернули они кредит или нет.

Для применения Байесовского подхода необходимо убедиться в выполнении 3-х гипотез:

* о независимости – клиенты не имеют сговора по выплате или не выплате кредита;
* об однородности – данные взяты из одной генеральной совокупности;
* о виде распределения – данные распределены равновероятно.

*Гипотеза о независимости*

Для проверки этой гипотезы воспользуемся ранговым критерием Спирмена. Статистикой данного критерия является коэффициент ранговой корреляции, определяемый следующим образом.

Даны два ряда наблюдений:  и . На основании этих наблюдений построим пары рангов . Под рангом  понимаем номер места, занимаемого наблюдением  в вариационном ряду . Аналогично понимаем ранг . Затем, переставляем пары рангов в порядке возрастания первой компоненты. Получившийся ряд обозначим .

Коэффициент корреляции находится по формуле: .

Критическая область критерия . Для нахождения  воспользуемся тем, что закон распределения  стремится к , при больших n. Отсюда . Здесь  - функция распределения стандартного Гауссова закона.

При уровне значимости , =1.959964. Граница критической зоны . Найденный коэффициент корреляции .

Таким образом, статистика критерия не попадает в его критическую область, и мы можем принять гипотезу о независимости при уровне значимости 0.05.

*Гипотеза об однородности*

Формулируется гипотеза следующим образом. Даны две выборки  и  из распределений и соответственно, с функциями распределений  и . Тогда гипотеза об однородности .

Для проверки данной гипотезы воспользуемся критерием Смирнова.

Статистикой данного критерия является , где  - эмпирические функции распределения, построенные по выборкам и . Критическая область задаётся в виде . При больших n и m границу критической области  можно принять равной , где .  - функция распределения Колмогорова. .

Таким образом, гипотеза об однородности отвергается, в случае если . При уровне значимости 0.05 .  - граница критической области. Статистика .

Статистика критерия не попадает в критическую область, и мы можем принять гипотезу об однородности при уровне значимости 0.05.

*Гипотеза о виде распределения*

Сформулируем гипотезу. Нам дана выборка  из распределения  с функцией распределения , которая неизвестна. Необходимо проверить, что  - функция распределения равномерного распределения на отрезке [0;0.05].



Для этого воспользуемся критерием согласия Пирсона.

Статистикой критерия является . Здесь  - частота попадания наблюдений в i-й отрезок,  - вероятность попадания в i-й отрезок. Если проверяемая гипотеза верна, при больших n статистика подчиняется распределению хи-квадрат с k-1 степенью свободы.

Гипотеза отвергается в случае, если статистика превышает критическое значение .

Значение статистики посчитаем с помощью программного пакета Statistica - =24,19468, k-1=39. Граница критической области при уровне значимости  .

Таким образом значение статистики не превышает критического уровня и гипотеза о равномерном распределении клиентов по группам принимается при уровне значимости 0.05.

Итак, данные удовлетворяют всем гипотезам, приведённым выше, и мы можем приступить к нахождению рейтингов.

Имеющиеся данные содержат множество различных характеристик клиентов. Для построения будем использовать 4 из них, наиболее значимых. Взять большее количество элементов нам не позволяет ограниченность нашей выборки (1977 элементов).

Выбранные характеристики: возраст и пол заёмщика, наличие детей, выплаты по кредиту в % от суммарного дохода заёмщика. Характеристика возраст принимает 3 значения – 18-29, 30-45, 46-…; пол заёмщика два значения – мужской и женский; наличие детей два значения – есть дети, и нет детей; выплаты принимают 5 значений – (<5%),(6-10%), (11-16%), (17-22%),(23-55%).

Каждое конкретное значение характеристики назовём свойством заёмщика. Разобьём всех наших клиентов на множества, опираясь на наличие конкретного свойства у данного клиента. К примеру,  – множество клиентов у которых нет детей.

Приведём эти обозначения:

 - Дети, = Нет детей, = Есть дети;

 - Возраст, =Возр1(18-29), =Возр2(30-45), =Возр3(46-…);

 - Пол, =Женский, =Мужской;

- выплаты по кредиту в % от суммарного дохода заемщика, =Вып1(<5), =Вып2(6-10), =Вып3(11-16), =Вып4(17-22), =Вып5(23-55).

Образуем новые множества , как комбинацию всех возможных свойств клиента -  по всевозможным  К примеру, множество  состоит из женщин в возрасте 18-29 лет без детей, выплачивающих <5% от своего суммарного дохода. Количество таких множеств равно 60.

Таблица 2.1. Кодировка множеств.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Возр1 | Возр2 | Возр3 | Женский | Мужской | Нет детей | Есть дети | Вып1 | Вып2 | Вып3 | Вып4 | Вып5 |
| 1 | + |  |  | + |  | + |  | + |  |  |  |  |
| 2 | + |  |  | + |  | + |  |  | + |  |  |  |
| 3 | + |  |  | + |  | + |  |  |  | + |  |  |
| 4 | + |  |  | + |  | + |  |  |  |  | + |  |
| 5 | + |  |  | + |  | + |  |  |  |  |  | + |
| 6 | + |  |  |  | + | + |  | + |  |  |  |  |
| 7 | + |  |  |  | + | + |  |  | + |  |  |  |
| 8 | + |  |  |  | + | + |  |  |  | + |  |  |
| 9 | + |  |  |  | + | + |  |  |  |  | + |  |
| 10 | + |  |  |  | + | + |  |  |  |  |  | + |
| 11 |  | + |  | + |  | + |  | + |  |  |  |  |
| 12 |  | + |  | + |  | + |  |  | + |  |  |  |
| 13 |  | + |  | + |  | + |  |  |  | + |  |  |
| 14 |  | + |  | + |  | + |  |  |  |  | + |  |
| 15 |  | + |  | + |  | + |  |  |  |  |  | + |
| 16 |  | + |  |  | + | + |  | + |  |  |  |  |
| 17 |  | + |  |  | + | + |  |  | + |  |  |  |
| 18 |  | + |  |  | + | + |  |  |  | + |  |  |
| 19 |  | + |  |  | + | + |  |  |  |  | + |  |
| 20 |  | + |  |  | + | + |  |  |  |  |  | + |
| 21 |  |  | + | + |  | + |  | + |  |  |  |  |
| 22 |  |  | + | + |  | + |  |  | + |  |  |  |
| 23 |  |  | + | + |  | + |  |  |  | + |  |  |
| 24 |  |  | + | + |  | + |  |  |  |  | + |  |
| 25 |  |  | + | + |  | + |  |  |  |  |  | + |
| 26 |  |  | + |  | + | + |  | + |  |  |  |  |
| 27 |  |  | + |  | + | + |  |  | + |  |  |  |
| 28 |  |  | + |  | + | + |  |  |  | + |  |  |
| 29 |  |  | + |  | + | + |  |  |  |  | + |  |
| 30 |  |  | + |  | + | + |  |  |  |  |  | + |
| 31 | + |  |  | + |  |  | + | + |  |  |  |  |
| 32 | + |  |  | + |  |  | + |  | + |  |  |  |
| 33 | + |  |  | + |  |  | + |  |  | + |  |  |
| 34 | + |  |  | + |  |  | + |  |  |  | + |  |
| 35 | + |  |  | + |  |  | + |  |  |  |  | + |
| 36 | + |  |  |  | + |  | + | + |  |  |  |  |
| 37 | + |  |  |  | + |  | + |  | + |  |  |  |
| 38 | + |  |  |  | + |  | + |  |  | + |  |  |
| 39 | + |  |  |  | + |  | + |  |  |  | + |  |
| 40 | + |  |  |  | + |  | + |  |  |  |  | + |
| 41 |  | + |  | + |  |  | + | + |  |  |  |  |
| 42 |  | + |  | + |  |  | + |  | + |  |  |  |
| 43 |  | + |  | + |  |  | + |  |  | + |  |  |
| 44 |  | + |  | + |  |  | + |  |  |  | + |  |
| 45 |  | + |  | + |  |  | + |  |  |  |  | + |
| 46 |  | + |  |  | + |  | + | + |  |  |  |  |
| 47 |  | + |  |  | + |  | + |  | + |  |  |  |
| 48 |  | + |  |  | + |  | + |  |  | + |  |  |
| 49 |  | + |  |  | + |  | + |  |  |  | + |  |
| 50 |  | + |  |  | + |  | + |  |  |  |  | + |
| 51 |  |  | + | + |  |  | + | + |  |  |  |  |
| 52 |  |  | + | + |  |  | + |  | + |  |  |  |
| 53 |  |  | + | + |  |  | + |  |  | + |  |  |
| 54 |  |  | + | + |  |  | + |  |  |  | + |  |
| 55 |  |  | + | + |  |  | + |  |  |  |  | + |
| 56 |  |  | + |  | + |  | + | + |  |  |  |  |
| 57 |  |  | + |  | + |  | + |  | + |  |  |  |
| 58 |  |  | + |  | + |  | + |  |  | + |  |  |
| 59 |  |  | + |  | + |  | + |  |  |  | + |  |
| 60 |  |  | + |  | + |  | + |  |  |  |  | + |

Заметим, что - образуют разбиение всего множества клиентов:  и .

Построим совместное эмпирическое распределение двух дискретных случайных величин –  и , где ={0,1}, ={, i=1:60}. Строить его будем как отношение количества клиентов удовлетворяющих паре значений случайных величин (X,Y) к общему количеству клиентов.

Зафиксируем количество клиентов соответствующих каждой возможной паре (X,Y).

Таблица 2.2. Количество клиентов вернувших и не вернувших кредит в каждой группе. 0 – вернули кредит, 1 – не вернули кредит.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| D X Y | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 0 | 20 | 45 | 44 | 21 | 15 | 12 | 23 | 31 | 15 | 4 | 18 | 28 | 35 | 14 | 12 | 19 | 34 | 45 | 21 | 11 |
| 1 | 12 | 22 | 20 | 10 | 10 | 3 | 17 | 17 | 8 | 3 | 12 | 23 | 13 | 5 | 6 | 5 | 25 | 15 | 4 | 3 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| D X Y | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| 0 | 18 | 36 | 47 | 19 | 5 | 20 | 49 | 52 | 30 | 28 | 10 | 21 | 20 | 16 | 8 | 11 | 27 | 27 | 8 | 6 |
| 1 | 6 | 11 | 11 | 6 | 3 | 5 | 10 | 16 | 4 | 10 | 5 | 17 | 13 | 3 | 4 | 6 | 17 | 10 | 4 | 5 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| D X Y | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| 0 | 49 | 65 | 65 | 28 | 14 | 49 | 69 | 58 | 27 | 20 | 5 | 12 | 7 | 3 | 2 | 2 | 6 | 6 | 4 | 1 |
| 1 | 12 | 32 | 15 | 9 | 2 | 14 | 27 | 18 | 9 | 3 | 2 | 3 | 3 | 2 | 2 | 1 | 2 | 4 | 0 | 1 |

Построим совместное эмпирическое распределение вероятностей. Для этого разделим количество клиентов вернувших и не вернувших кредит в каждой из групп на общее количество клиентов.

Таблица 2.3. Совместное эмпирическое распределение вероятностей.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 |  | 0 | 1 |
| 1 | 0,010116 | 0,00607 | 31 | 0,005058 | 0,002529 |
| 2 | 0,022762 | 0,011128 | 32 | 0,010622 | 0,008599 |
| 3 | 0,022256 | 0,010116 | 33 | 0,010116 | 0,006576 |
| 4 | 0,010622 | 0,005058 | 34 | 0,008093 | 0,001517 |
| 5 | 0,007587 | 0,005058 | 35 | 0,004047 | 0,002023 |
| 6 | 0,00607 | 0,001517 | 36 | 0,005564 | 0,003035 |
| 7 | 0,011634 | 0,008599 | 37 | 0,013657 | 0,008599 |
| 8 | 0,01568 | 0,008599 | 38 | 0,013657 | 0,005058 |
| 9 | 0,007587 | 0,004047 | 39 | 0,004047 | 0,002023 |
| 10 | 0,002023 | 0,001517 | 40 | 0,003035 | 0,002529 |
| 11 | 0,009105 | 0,00607 | 41 | 0,024785 | 0,00607 |
| 12 | 0,014163 | 0,011634 | 42 | 0,032878 | 0,016186 |
| 13 | 0,017704 | 0,006576 | 43 | 0,032878 | 0,007587 |
| 14 | 0,007081 | 0,002529 | 44 | 0,014163 | 0,004552 |
| 15 | 0,00607 | 0,003035 | 45 | 0,007081 | 0,001012 |
| 16 | 0,009611 | 0,002529 | 46 | 0,024785 | 0,007081 |
| 17 | 0,017198 | 0,012645 | 47 | 0,034901 | 0,013657 |
| 18 | 0,022762 | 0,007587 | 48 | 0,029337 | 0,009105 |
| 19 | 0,010622 | 0,002023 | 49 | 0,013657 | 0,004552 |
| 20 | 0,005564 | 0,001517 | 50 | 0,010116 | 0,001517 |
| 21 | 0,009105 | 0,003035 | 51 | 0,002529 | 0,001012 |
| 22 | 0,018209 | 0,005564 | 52 | 0,00607 | 0,001517 |
| 23 | 0,023773 | 0,005564 | 53 | 0,003541 | 0,001517 |
| 24 | 0,009611 | 0,003035 | 54 | 0,001517 | 0,001012 |
| 25 | 0,002529 | 0,001517 | 55 | 0,001012 | 0,001012 |
| 26 | 0,010116 | 0,002529 | 56 | 0,001012 | 0,000506 |
| 27 | 0,024785 | 0,005058 | 57 | 0,003035 | 0,001012 |
| 28 | 0,026302 | 0,008093 | 58 | 0,003035 | 0,002023 |
| 29 | 0,015175 | 0,002023 | 59 | 0,002023 | 0 |
| 30 | 0,014163 | 0,005058 | 60 | 0,000506 | 0,000506 |

Рисунок 2.1. Эмпирическая функция распределения при X=0.

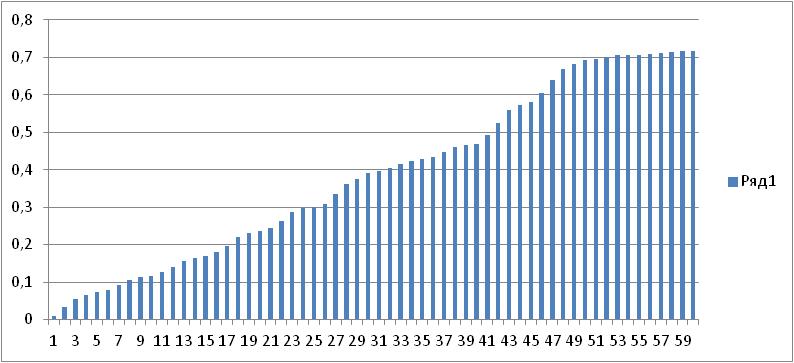


Рисунок 2.2. Эмпирическая функция распределения при X =1.

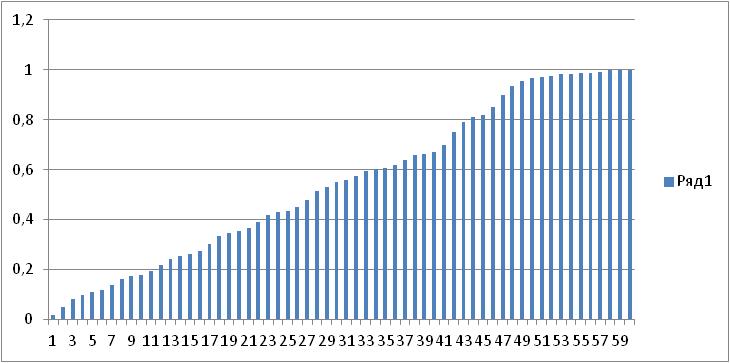
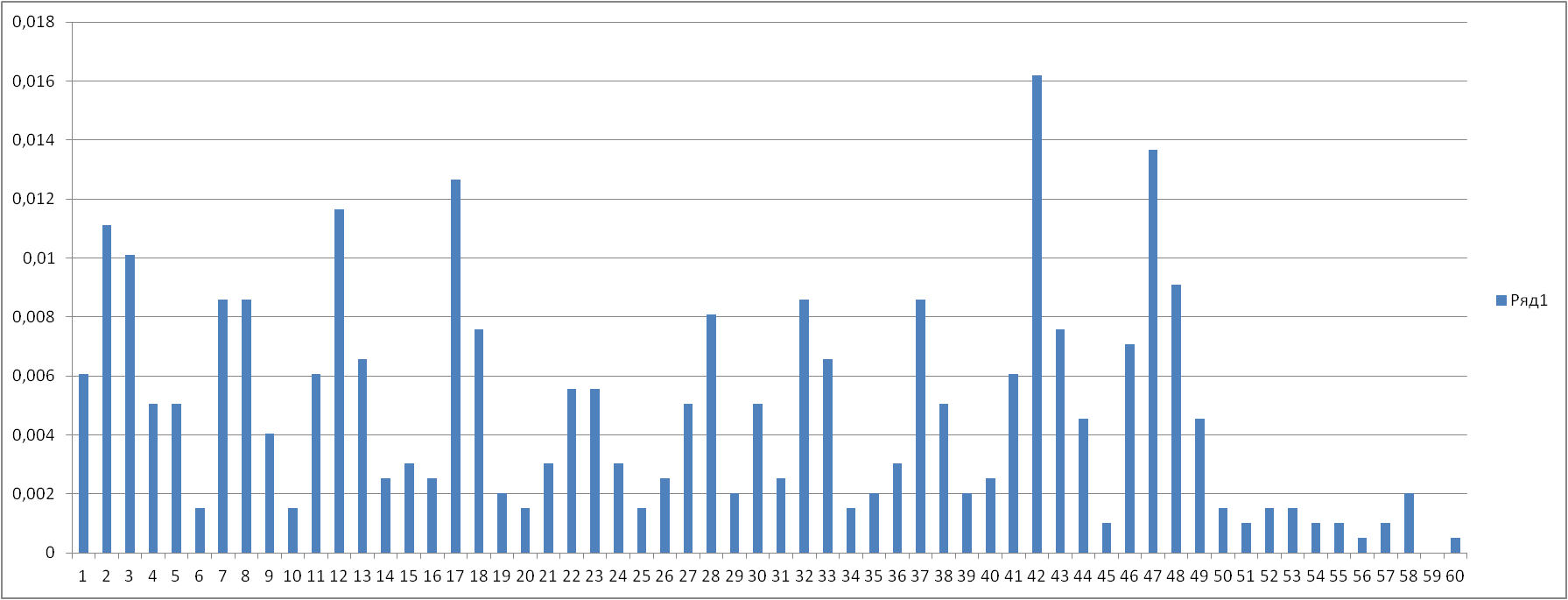


Рисунок 2.3. Гистограмма совместного распределения при X=0.



Рисунок 2.4. Гистограмма совместного распределения при X=1.

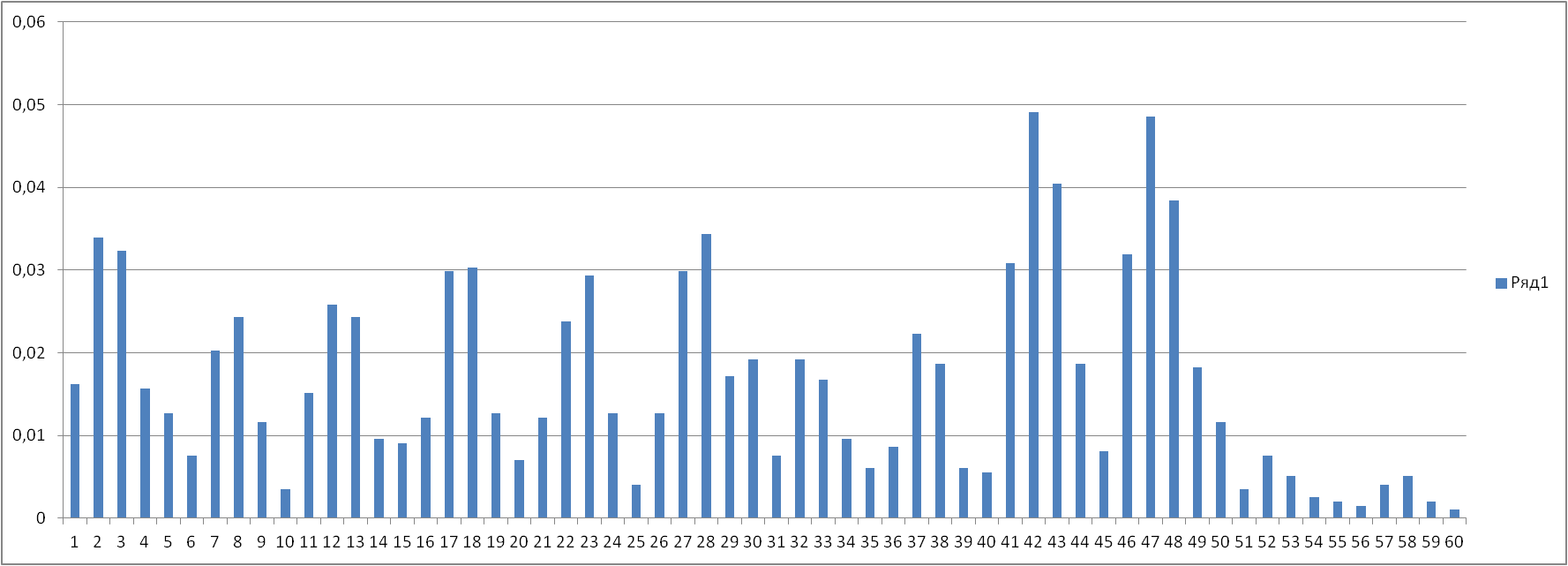


Найдём эмпирическое распределение вероятностей попадания в каждую из групп. Для этого разделим количество клиентов в каждой группе на общее количество клиентов. Общее количество клиентов . .

Таблица 2.4. Эмпирическое распределение вероятностей попадания в каждую группу.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № |  | № |  | № |  | № |  | № |  |
| 1 | 0,016186 | 13 | 0,024279 | 25 | 0,004047 | 37 | 0,022256 | 49 | 0,018209 |
| 2 | 0,03389 | 14 | 0,009611 | 26 | 0,012645 | 38 | 0,018715 | 50 | 0,011634 |
| 3 | 0,032372 | 15 | 0,009105 | 27 | 0,029843 | 39 | 0,00607 | 51 | 0,003541 |
| 4 | 0,01568 | 16 | 0,01214 | 28 | 0,034396 | 40 | 0,005564 | 52 | 0,007587 |
| 5 | 0,012645 | 17 | 0,029843 | 29 | 0,017198 | 41 | 0,030855 | 53 | 0,005058 |
| 6 | 0,007587 | 18 | 0,030349 | 30 | 0,019221 | 42 | 0,049064 | 54 | 0,002529 |
| 7 | 0,020233 | 19 | 0,012645 | 31 | 0,007587 | 43 | 0,040465 | 55 | 0,002023 |
| 8 | 0,024279 | 20 | 0,007081 | 32 | 0,019221 | 44 | 0,018715 | 56 | 0,001517 |
| 9 | 0,011634 | 21 | 0,01214 | 33 | 0,016692 | 45 | 0,008093 | 57 | 0,004047 |
| 10 | 0,003541 | 22 | 0,023773 | 34 | 0,009611 | 46 | 0,031866 | 58 | 0,005058 |
| 11 | 0,015175 | 23 | 0,029337 | 35 | 0,00607 | 47 | 0,048558 | 59 | 0,002023 |
| 12 | 0,025797 | 24 | 0,012645 | 36 | 0,008599 | 48 | 0,038442 | 60 | 0,001012 |

Рисунок 2.5. Гистограмма попадания в группу i.

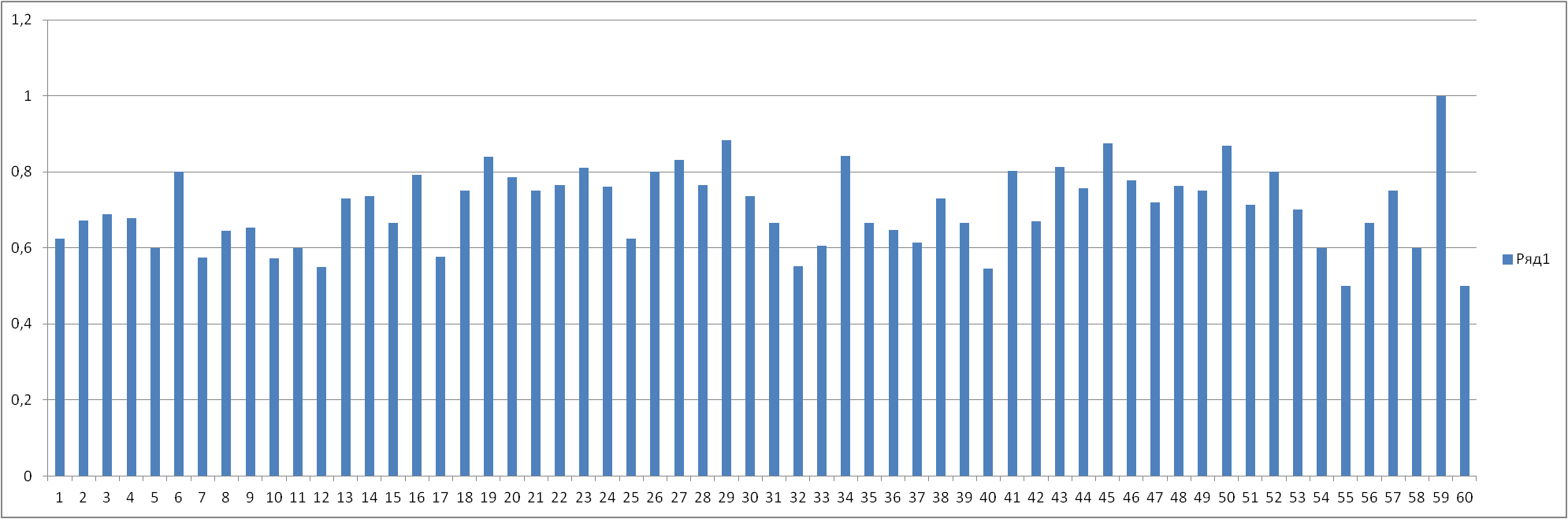


Найдём рейтинги клиентов как эмпирическую вероятность того, что клиент вернёт кредит при условии, что он принадлежит какой-либо группе. Для этого воспользуемся тем, что .

Таблица 2.5. Рейтинги клиентов.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № |  | № |  | № |  |
| 1 | 0,625 | 21 | 0,75 | 41 | 0,803279 |
| 2 | 0,671642 | 22 | 0,765957 | 42 | 0,670103 |
| 3 | 0,6875 | 23 | 0,810345 | 43 | 0,8125 |
| 4 | 0,677419 | 24 | 0,76 | 44 | 0,756757 |
| 5 | 0,6 | 25 | 0,625 | 45 | 0,875 |
| 6 | 0,8 | 26 | 0,8 | 46 | 0,777778 |
| 7 | 0,575 | 27 | 0,830508 | 47 | 0,71875 |
| 8 | 0,645833 | 28 | 0,764706 | 48 | 0,763158 |
| 9 | 0,652174 | 29 | 0,882353 | 49 | 0,75 |
| 10 | 0,571429 | 30 | 0,736842 | 50 | 0,869565 |
| 11 | 0,6 | 31 | 0,666667 | 51 | 0,714286 |
| 12 | 0,54902 | 32 | 0,552632 | 52 | 0,8 |
| 13 | 0,729167 | 33 | 0,606061 | 53 | 0,7 |
| 14 | 0,736842 | 34 | 0,842105 | 54 | 0,6 |
| 15 | 0,666667 | 35 | 0,666667 | 55 | 0,5 |
| 16 | 0,791667 | 36 | 0,647059 | 56 | 0,666667 |
| 17 | 0,576271 | 37 | 0,613636 | 57 | 0,75 |
| 18 | 0,75 | 38 | 0,72973 | 58 | 0,6 |
| 19 | 0,84 | 39 | 0,666667 | 59 | 1 |
| 20 | 0,785714 | 40 | 0,545455 | 60 | 0,5 |

Рисунок 2.6. Гистограмма рейтингов.



На основе полученных рейтингов можем вынести решение о выдаче кредита.

Выделим 3 варианта:

1. Рейтинг клиента лежит в полуинтервале (0.7, 1] 🡪 клиент считается надёжным. Ему можно выдать кредит.
2. Рейтинг клиента лежит в полуинтервале (0.5, 0.7] 🡪клиент считается среднерисковым. Если пересмотреть условия кредитования, то клиенту можно будет выдать кредит.
3. Рейтинг клиента 🡪 клиент считается рисковым. Ему не стоит выдавать кредит.

Исходя из этого, получаем классификацию:

* Клиенты из группы 60 рисковые
* Клиенты из групп 1-5, 7-12, 15, 17, 25, 31, 32, 33, 35-37, 39, 40, 42, 54-56, 58 – среднерисковые
* Клиенты из групп 6, 13, 14, 16, 18-24, 26-30, 34, 38, 41, 43-53, 57, 59 – надежные

# Глава 3. Подход, основанный на задаче Монжа-Канторовича.

## 3.1. Задача Монжа-Канторовича.

Задача Монжа.

Даны два вероятностных пространства  и  и неотрицательная измеримая функция  на , называемая функцией стоимости. Требуется найти отображение , измеримое относительно пары , т. е.  при всех , переводящее меру  в  и доставляющее минимум выражению



среди всех таких отображений. Условие, что  переходит в , означает, что  совпадает с образом меры  при отображении , который задается формулой . Если некоторое  дает минимум, то это  называют оптимальным отображением меры  в  или оптимальной транспортировкой.

В интерпретации самого Монжа речь шла о переносе грунта для ведения строительных работ, т.е. обе меры были обычными объемами, функция – обычным расстоянием, а минимизировалась выполненная работа. Внешне задача выглядит как прикладная, однако она таковой не является. Данная формулировка не является единственной “экономической” формулировкой транспортной задачи. Дискретный вариант задач Монжа и Канторовича возникает во многих современных областях.

В 1942 году майор Л.В. Канторович сдаёт в печать свою в заметку, в которой ставится задача, близкая к проблеме Монжа, но с одним принципиально важным нюансом: в задаче Канторовича вместо поиска отображения T (“оптимальной транспортировки”) предлагается нахождение лишь “оптимального плана транспортировки”, т. е. такой вероятностной меры  на , что её проекция на  и  суть  и  соответственно, причём  доставляет минимум выражению



по всем вероятностным мерам из класса  вероятностных мер на произведении , дающих  и  при проектировании на  и .

Задача Канторовича является линейной в отличие от сложной задачи Монжа: требуется найти минимум линейного функционала на выпуклом множестве.

В случае если задача Монжа имеет решение, то и задача Канторовича имеет решение. В некоторых частных случаях удается по решению задачи Канторовича построить решение задачи Монжа, но это довольно большая редкость.

Тем не менее, существует тесная связь между задачами Монжа и Канторовича. В связи с этим общепринятым стал термин “задача Монжа-Канторовича”

С задачи Канторовича начинается линейное программирование.

Рассмотрим дискретный вариант этой задачи. Он носит название *транспортной задачи* линейного программирования. Сформулируем её.

Даны пункты отправления  в каждом из которых имеется  единиц груза, и пункты назначения  с потребностями единиц груза. Известна стоимость перевозки единицы груза  из  в . Требуется составить план перевозок. При этом суммарная потребность в грузе совпадает с суммарным количеством груза в пунктах отправления, т.е.

.

Построим математическую модель данной задачи. Необходимо найти  - количество груза, перевозимое из пункта отправления  в пункт назначения . Очевидно, что должны выполняться следующие условия:

,



.

При этом должна минимизироваться суммарная стоимость перевозок

.

Получившаяся задача является типичной для линейного программирования. Она решается симплекс методом линейного программирования. Не будем приводить здесь алгоритм её решения, однако отметим, что существует процедура в среде MATLAB, находящая решение транспортной задачи.

## 3.2. Применение задачи Монжа-Канторовича к скорингу.

Постановка задачи та же, что и в главе 2. Для применения метода требуется выполнение тех же условий :

* независимость – клиенты не имеют сговора по выплате или не выплате кредита;
* однородность – данные взяты из одной генеральной совокупности;
* равновероятность – клиенты равновероятно распределены по группам.

Проверка данных гипотез приведена в главе 2.

Задача нахождения распределения рейтингов, сводится к нахождению совместного распределения вероятностей того, что клиент вернёт кредит и вероятностей того, что он попадёт в конкретную группу. Сводится она следующим образом: зная совместное распределение и вероятность попадания в каждую группу -  и , мы можем найти рейтинг

.

Сведём задачу нахождения совместного распределения к транспортной задаче.

Обозначим . При этом маргинальное распределение вероятностей  определено в главе 2. Укажем маргинальное распределение , рассчитываемое как отношение количества вернувших (не вернувших) кредит клиентов к общему количеству клиентов.

, .  – количество вернувших кредит,  – количество не вернувших кредит.

Требуется найти совместное распределение , такое что:

1. Минимизирует линейную форму ;
2. ;
3. ;
4. ;
5. 

В качестве функции стоимости  выберем модуль разницы между  и : .

Таким образом, мы свели задачу нахождения совместного распределения к транспортной задаче, решаемой методами линейного программирования. Однако в данном случае задачу можно решить более изящно. При этом сразу перейдём к нахождению рейтингов.

Рассмотрим 1-е условие к . Так как необходимо минимизировать сумму неотрицательных элементов, мы можем перейти к минимизации суммы по , зафиксировав при этом конкретное :

.

Воспользовавшись тем, что , перейдём к .

Что эквивалентно минимизации

.

От минимизации мы можем сделать эквивалентный переход к максимизации:

.

Обозначим , , , .

Таким образом, задача сводится к максимизации скалярного произведения .

Воспользуемся неравенством Коши-Буняковского, которое гласит, что . Равенство достигается в случае, когда вектора коллинеарные. Отсюда получаем, что .

Отсюда получаем решение нашей задачи: .

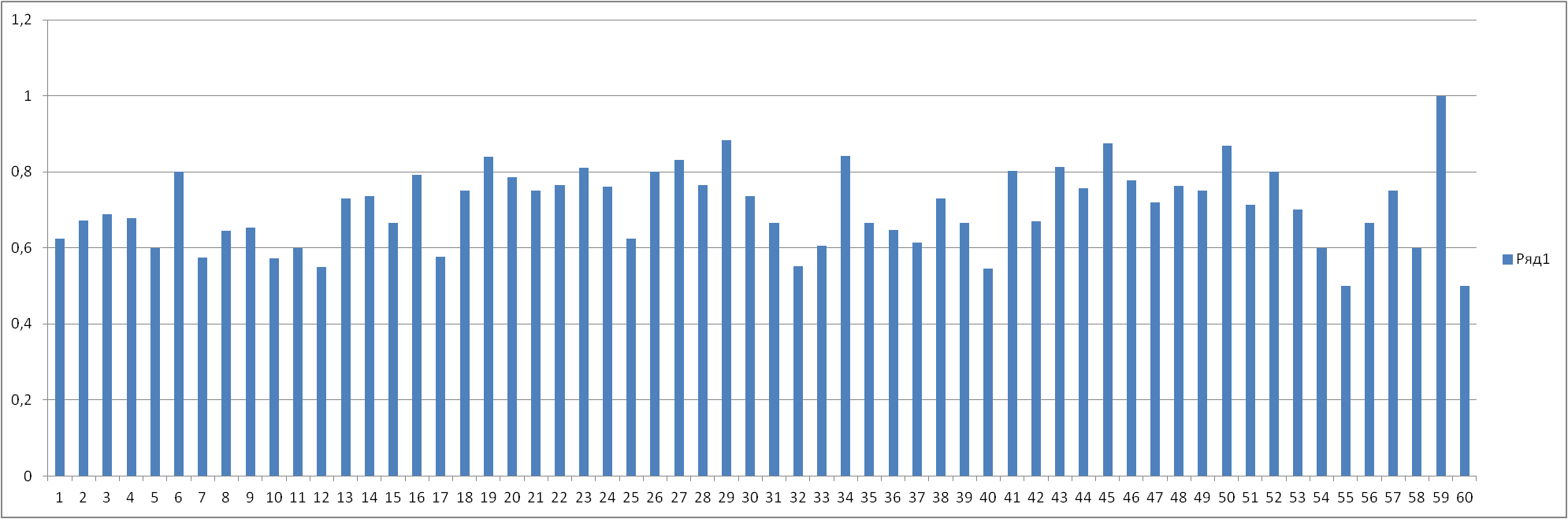
Таким образом, транспортная задача решилась явно. Мы получили рейтинги клиентов. Как видно из формулы они совпадают с рейтингами, найденными в эмпирическом Байесовском подходе.

Приведём посчитанные значения рейтингов.

Таблица 3.1. Рейтинги клиентов.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № |  | № |  | № |  |
| 1 | 0,625 | 21 | 0,75 | 41 | 0,803279 |
| 2 | 0,671642 | 22 | 0,765957 | 42 | 0,670103 |
| 3 | 0,6875 | 23 | 0,810345 | 43 | 0,8125 |
| 4 | 0,677419 | 24 | 0,76 | 44 | 0,756757 |
| 5 | 0,6 | 25 | 0,625 | 45 | 0,875 |
| 6 | 0,8 | 26 | 0,8 | 46 | 0,777778 |
| 7 | 0,575 | 27 | 0,830508 | 47 | 0,71875 |
| 8 | 0,645833 | 28 | 0,764706 | 48 | 0,763158 |
| 9 | 0,652174 | 29 | 0,882353 | 49 | 0,75 |
| 10 | 0,571429 | 30 | 0,736842 | 50 | 0,869565 |
| 11 | 0,6 | 31 | 0,666667 | 51 | 0,714286 |
| 12 | 0,54902 | 32 | 0,552632 | 52 | 0,8 |
| 13 | 0,729167 | 33 | 0,606061 | 53 | 0,7 |
| 14 | 0,736842 | 34 | 0,842105 | 54 | 0,6 |
| 15 | 0,666667 | 35 | 0,666667 | 55 | 0,5 |
| 16 | 0,791667 | 36 | 0,647059 | 56 | 0,666667 |
| 17 | 0,576271 | 37 | 0,613636 | 57 | 0,75 |
| 18 | 0,75 | 38 | 0,72973 | 58 | 0,6 |
| 19 | 0,84 | 39 | 0,666667 | 59 | 1 |
| 20 | 0,785714 | 40 | 0,545455 | 60 | 0,5 |

Рисунок 3.1. Гистограмма рейтингов.



На основе рейтингов можем классифицировать клиентов, по эмпирической вероятности возвращения ими кредита.

Выделим 3 варианта:

1. Рейтинг клиента лежит в полуинтервале (0.7, 1] 🡪 клиент считается надёжным. Ему можно выдать кредит.
2. Рейтинг клиента лежит в полуинтервале (0.5, 0.7] 🡪 клиент считается среднерисковым. Если пересмотреть условия кредитования, то этому клиенту можно будет выдать кредит.
3. Рейтинг клиента 🡪 клиент считается рисковым. Такому клиенту не стоит выдавать кредит.

Исходя из этого, получаем классификацию:

* Клиенты группы 60 – рисковые
* Клиент из групп 1-5, 7-12, 15, 17, 25, 31, 32, 33, 35-37, 39, 40, 42, 54-56, 58 – среднерисковые
* Клиенты из групп 6, 13, 14, 16, 18-24, 26-30, 34, 38, 41, 43-53, 57, 59 – надежные

# Заключение

В данной работе были описаны наиболее применяемые методы для построения скоринговой моделей. Приведено теоретическое обоснование эмпирического Байесовского метода. Метод был успешно применён к реальным кредитным данным. В результате были получены рейтинги клиентов.

Предложен новый метод, основывающийся на задаче Монжа-Канторовича. Описано его применение к тем же финансовым данным. Получены рейтинги клиентов, совпадающие с найденными Байесовским методом.

Из этого мы можем сделать вывод о том, что представленный в данной работе метод применим к построению скоринговых моделей. При этом метод основанный на задаче Монжа-Канторовича выделяется тем, что математически обоснован и определены условия его применения.

В настоящее время, скоринг получил широкое распространение в России и показал себя эффективным финансовым инструментом. В качестве примера, можно назвать область розничного кредитования. В данной области отказались от использования экспертной оценки в пользу скоринговых систем.

Однако, не смотря на широкое распространение, скоринг слабо освещён в российской литературе, хотя ему посвящено множество зарубежных работ. Скоринг обладает большим потенциалом применения, но все ещё является для людей использующих его “черным ящиком”. Следует продолжать изучать и совершенствовать скоринговые системы.

# Список литературы

1. Christine Bolton. Logistic regressions and its application in credit scoring. University of Pretoria, 2009, 240 c.
2. Samuel Glasson. Censored Regression Techniques for Credit Scoring. RMIT University, 2007, 196 c.
3. Georg Krempl. Adaptive Prediction Models and their Application to Credit Scoring. University of Graz, Austria, 2011, 163 c.
4. В. И. Богачев, А. В. Колесников. Задача Монжа-Канторовича: достижения, связи и перспективы // Успехи математических наук, 67: 5(407) (2012), 3-110.
5. Мудрецова Е. К. Дипломная работа на тему: Оценка рейтингов клиентов на основе математической модели скоринга. – М.: МИЭМ, 2013, 81 с.
6. Рачев С.Т. Задача Монжа-Канторовича о перемещении масс и её применение в стохастике // Теория вероятностей и ее применения, 29: 4 (1984), 625-653.
7. Руководство по кредитному скорингу. Под редакцией Элизабет Мэйз – М.: Гревцов Паблишер, 2008, 464 с.