МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

ФЕДЕРАЛЬНОЕ государственное БЮДЖЕТНОЕ

образовательное учреждение

высшего образования

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кафедра Автоматизированных систем управления

**ОТЧЁТ**

**по лабораторной работе №3**

**по дисциплине: «Методы анализа данных»**

Выполнили: Проверил:

Студенты гр. *АВТ-713*, *АВТФ* к.т.н.,

*Тягунова В.В. Ганелина Н.Д.*

*Ермаков И.В.*

*Майнгардт А.А.*

Новосибирск

2020

**3.1. Классификация**

Цель работы: изучение методов решения задачи классификации

**А) Описание массива данных**

В качестве исходного массива был выбран набор данных о клиентах банка. Целевая переменная, используемая для классификации заемщиков – оценка кредитоспособности лица, представленная в виде двух категорий: “good” или “bad”.

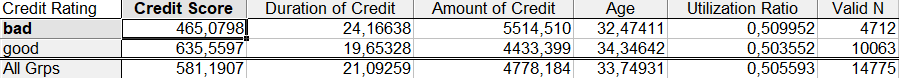
Набор данных содержит 15000 записей, при этом некоторые из записей подержат пропущенные данные (при последующем анализе такие наблюдения будут проигнорированы).

Признаки, используемые при анализе данных:

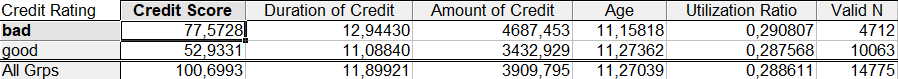
* Кредитный рейтинг (Credit Score) – интервальная переменная.
* Баланс банковского счета (Balance of Current Account) – порядковая переменная.
* Срок действия кредита (мес.) (Duration of Credit) – интервальная переменная.
* Статус погашения ранее оформленных кредитов (Payment of Previous Credits) – номинальная переменная.
* Цель кредита (Purpose of Credit) – номинальная переменная.
* Сумма кредита (Amount of Credit) – интервальная переменная.
* Ценность сбережений (Value of Savings) - порядковая переменная.
* Срок найма нынешним работодателем (Employed by Current Employer for) - порядковая переменная.
* Рассрочка от имеющегося дохода, % (Installment in % of Available Income) – порядковая переменная.
* Семейное положение (Marital Status) – номинальная переменная.
* Пол (Gender) - номинальная переменная.
* Время проживания по текущему месту жительства (Living in Current Household) – порядковая переменная.
* Наиболее ценный актив (Most Valuable Assets) – номинальная переменная.
* Возраст (Age) – интервальная переменная.
* Планы на дальнейший заем кредитов (Further Running Credits) - номинальная переменная.
* Тип недвижимости (Type of Apartment) – номинальная переменная.
* Число займов, осуществленном ранее в этом банке (Number of Previous Credits at This Bank) – порядковая переменная.
* Занятость (Occupation) – номинальная переменная.
* Тип населенного пункта (Home Area) – номинальная переменная.
* Наличие ранее просроченных кредитов (Prior Delinquency) – порядковая переменная.
* Коэффициент использования кредита (Utilization Ratio) – интервальная переменная.

**Б) Дескриптивный анализ данных**

1. Матрица средних для числовых категоризированных признаков выглядит следующим образом:

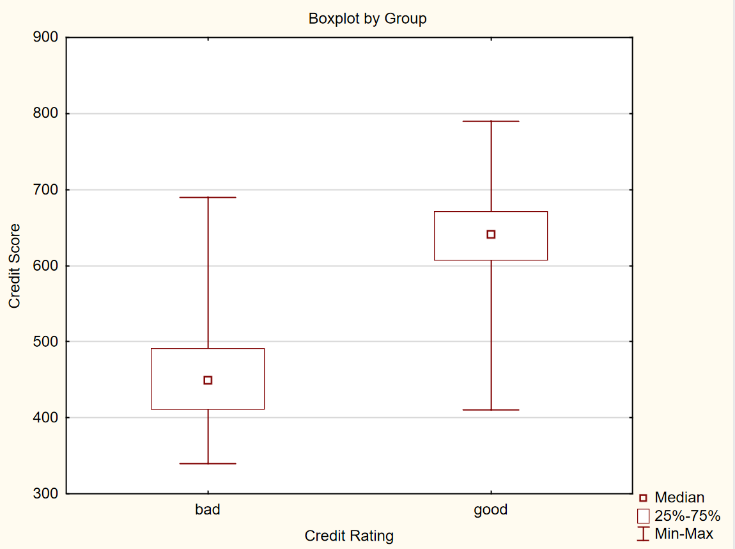
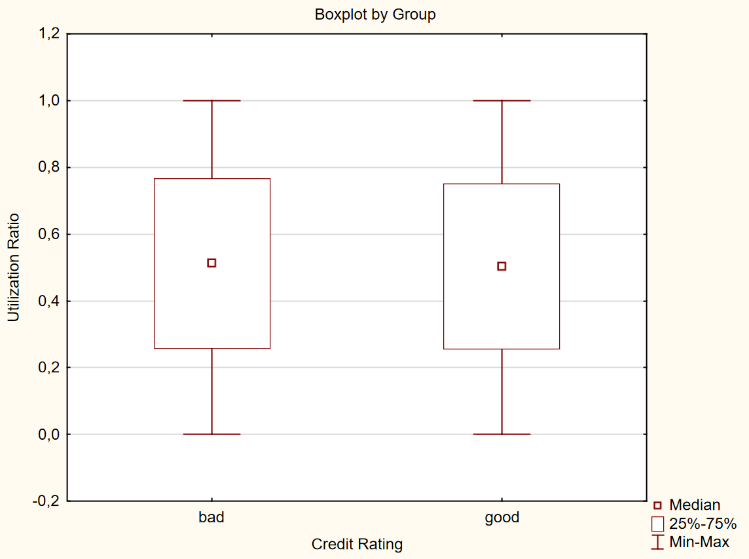


Матрица стандартных отклонений:

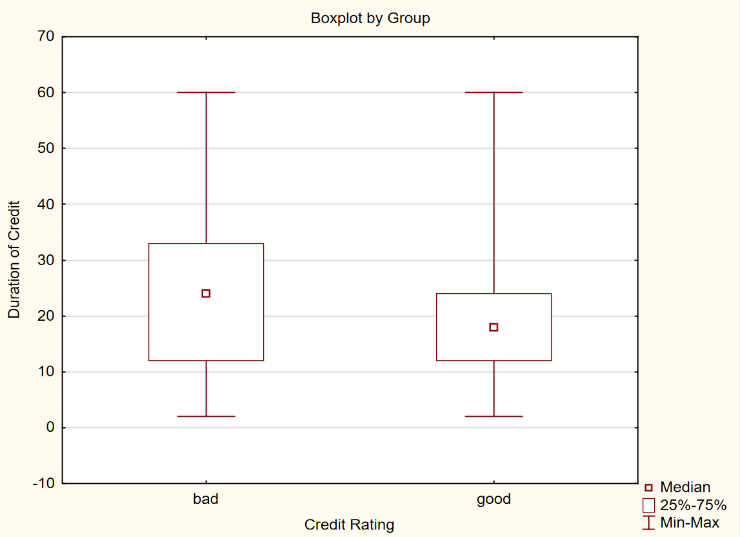
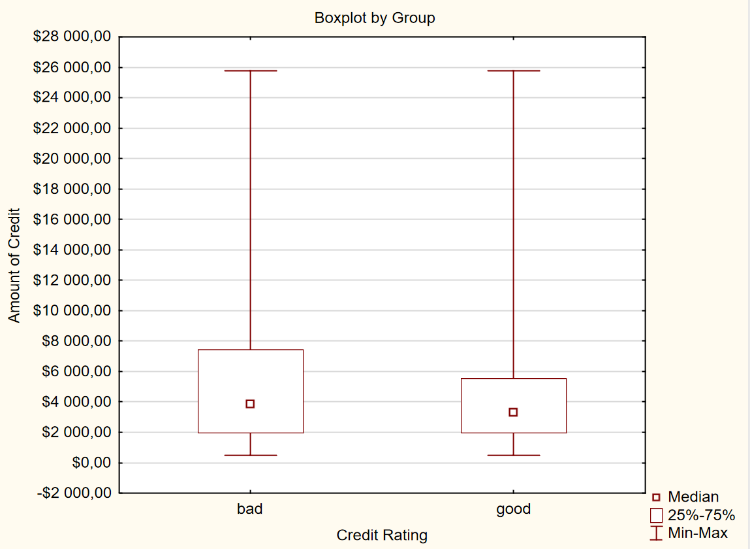


Помимо признаков, используемых в моделях классификации, в матрице произведен расчет числа записей по каждому классу (без учета записей, содержащих пропущенные данные) и записан в Valid N.

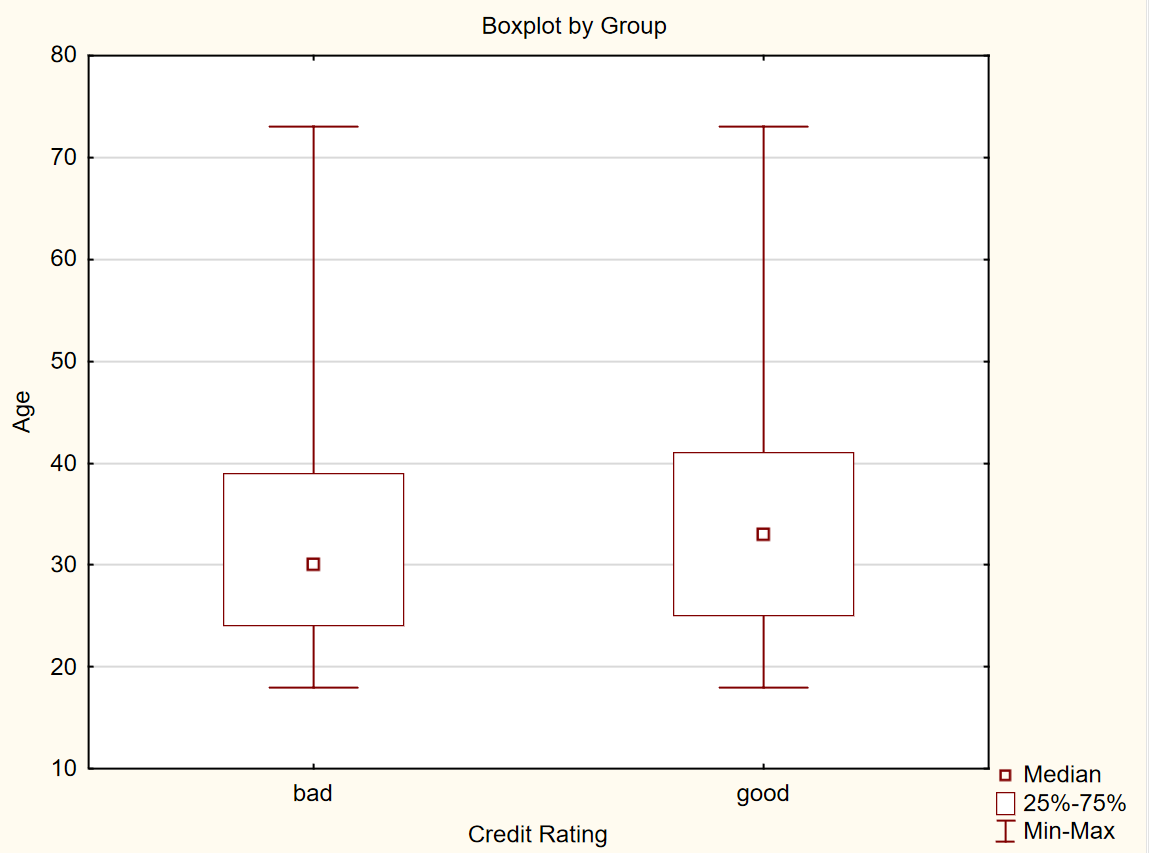
1. Диаграммы размаха (box plot) по каждому из параметров (здесь точкой обозначена медиана, нижняя и верхняя границы «ящика» отвечают за 25% и 75% квартиль соответственно, верхний и нижний отрезки вокруг прямоугольников характеризуют максимальное и минимальное значение по признаку соответственно):



*Кредитный рейтинг* *Коэфф. исп. кредита*



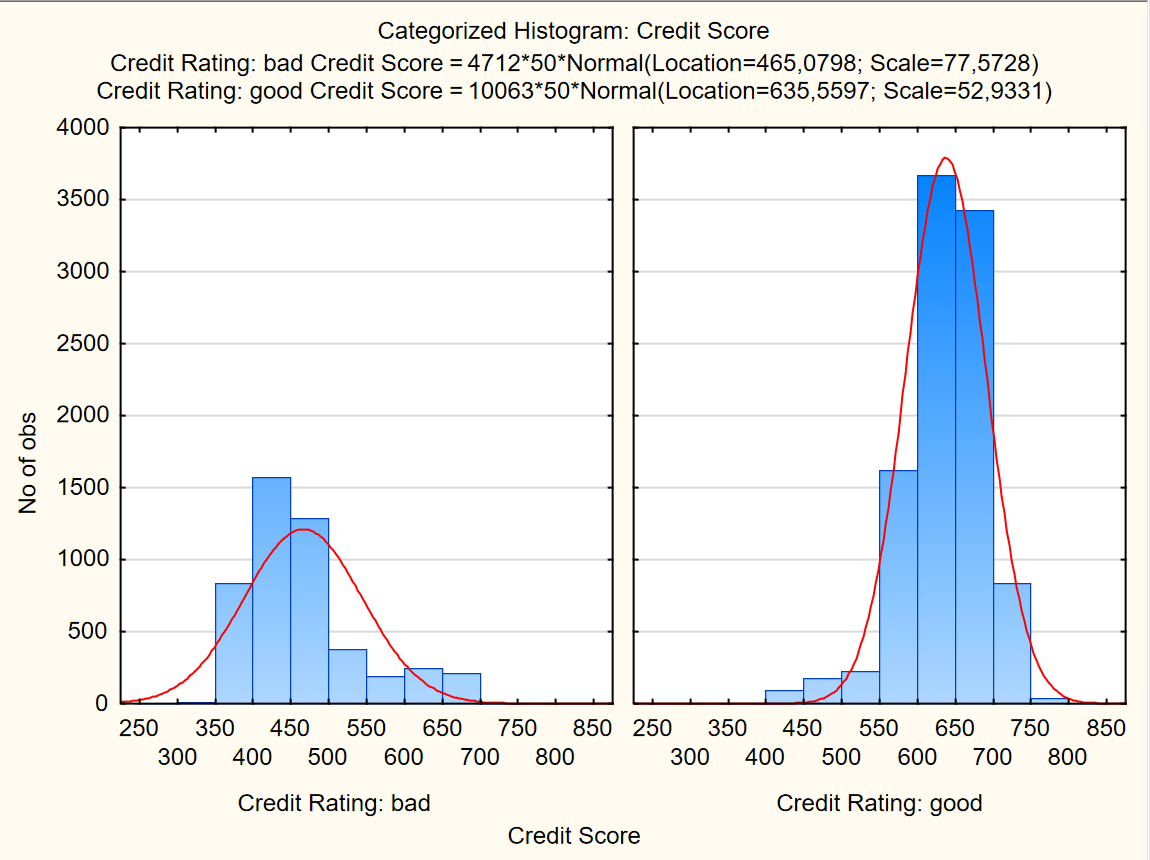
*Срок действия кредита Сумма кредита*



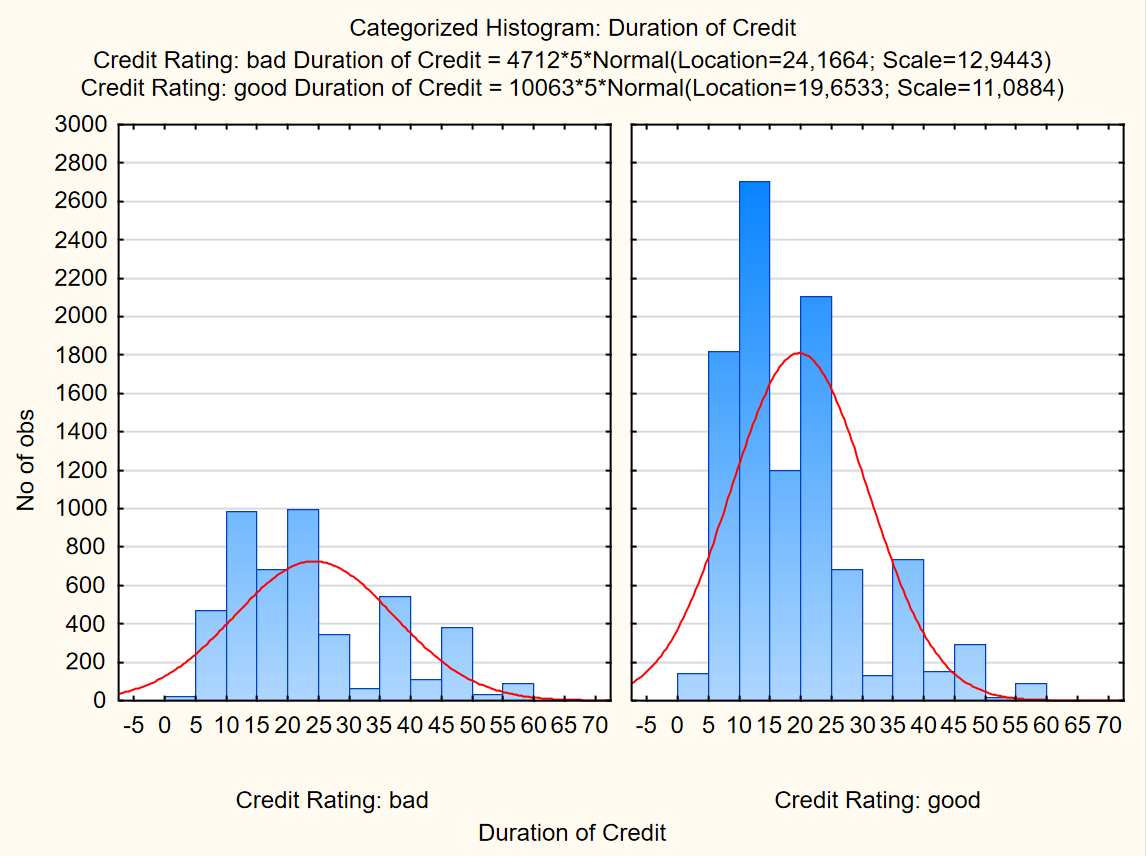
*Возраст*

Заметим, что для коэффициента использования кредита диаграммы рассеяния для имеющихся двух групп практически совпадают. Вероятно, что такой признак будет впоследствии исключен по причине минимального влияния при классифицировании объектов.

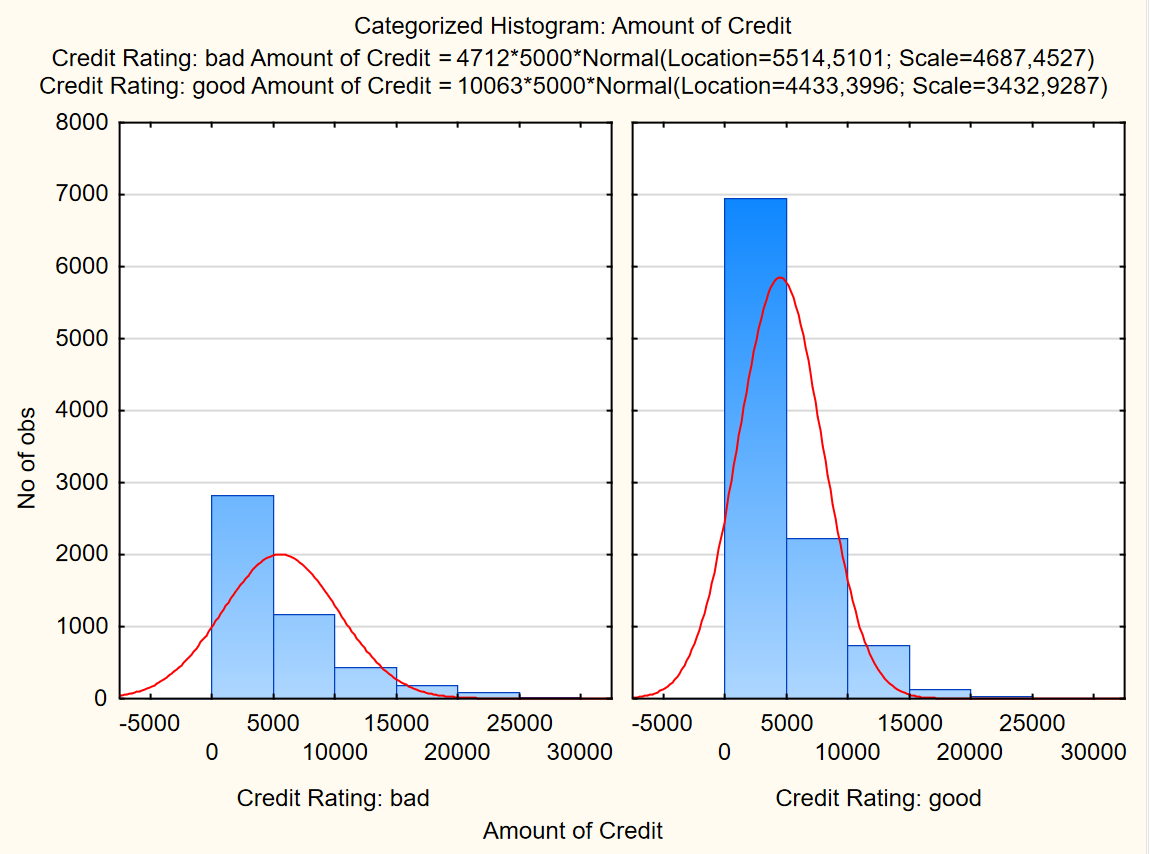
1. Гистограммы для числовых признаков будут иметь следующий вид:



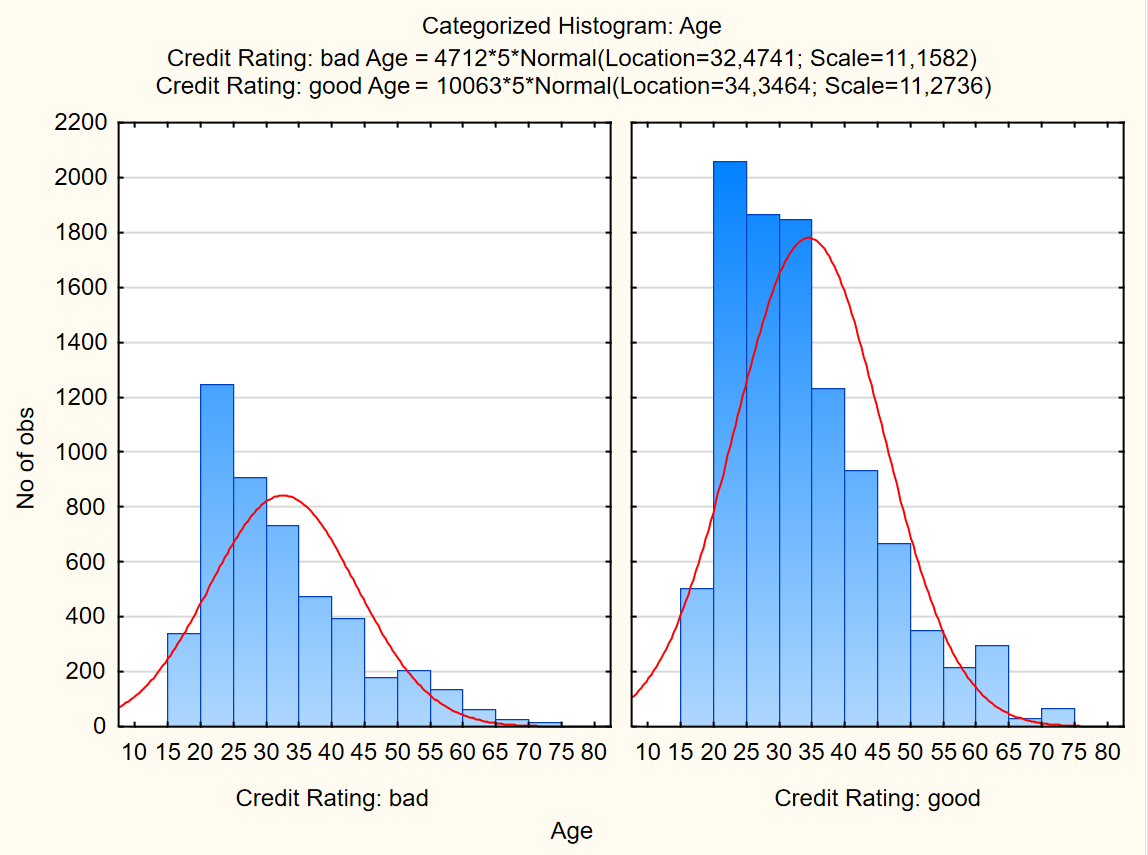
*Кредитный рейтинг*



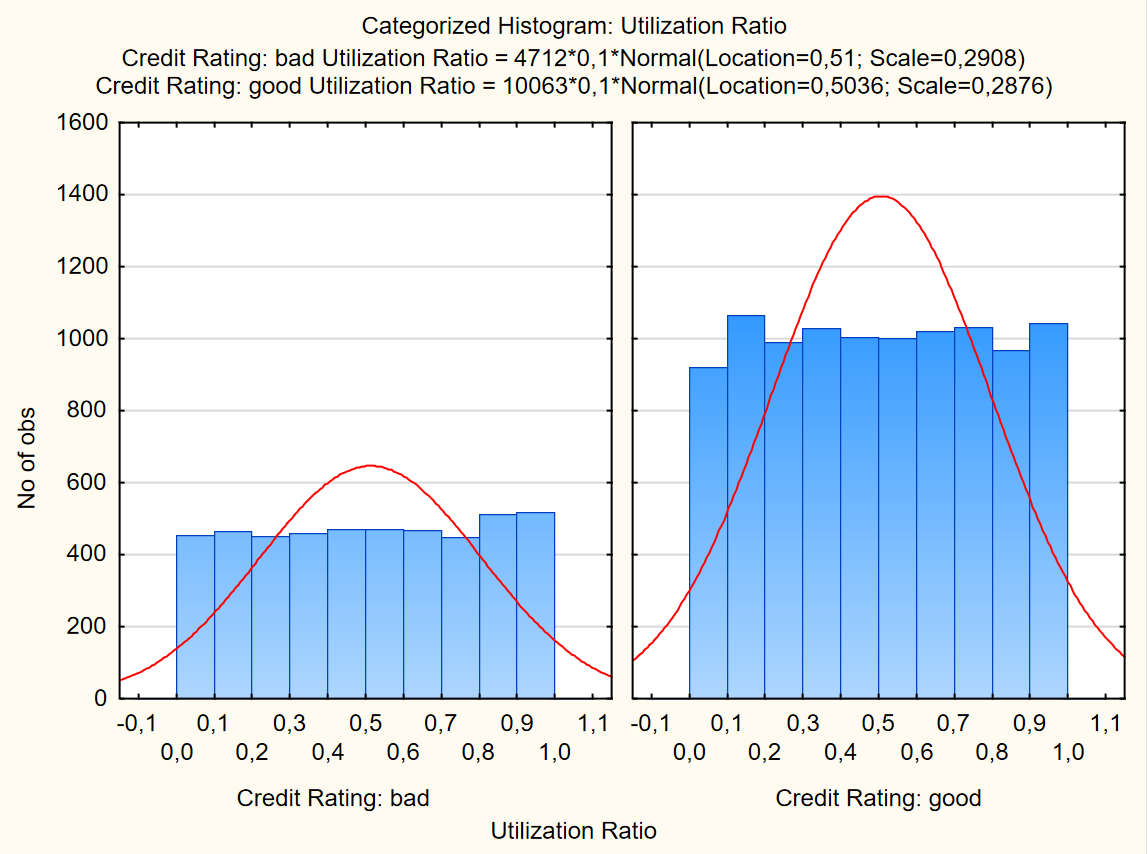
*Срок действия кредита*



*Сумма кредита*



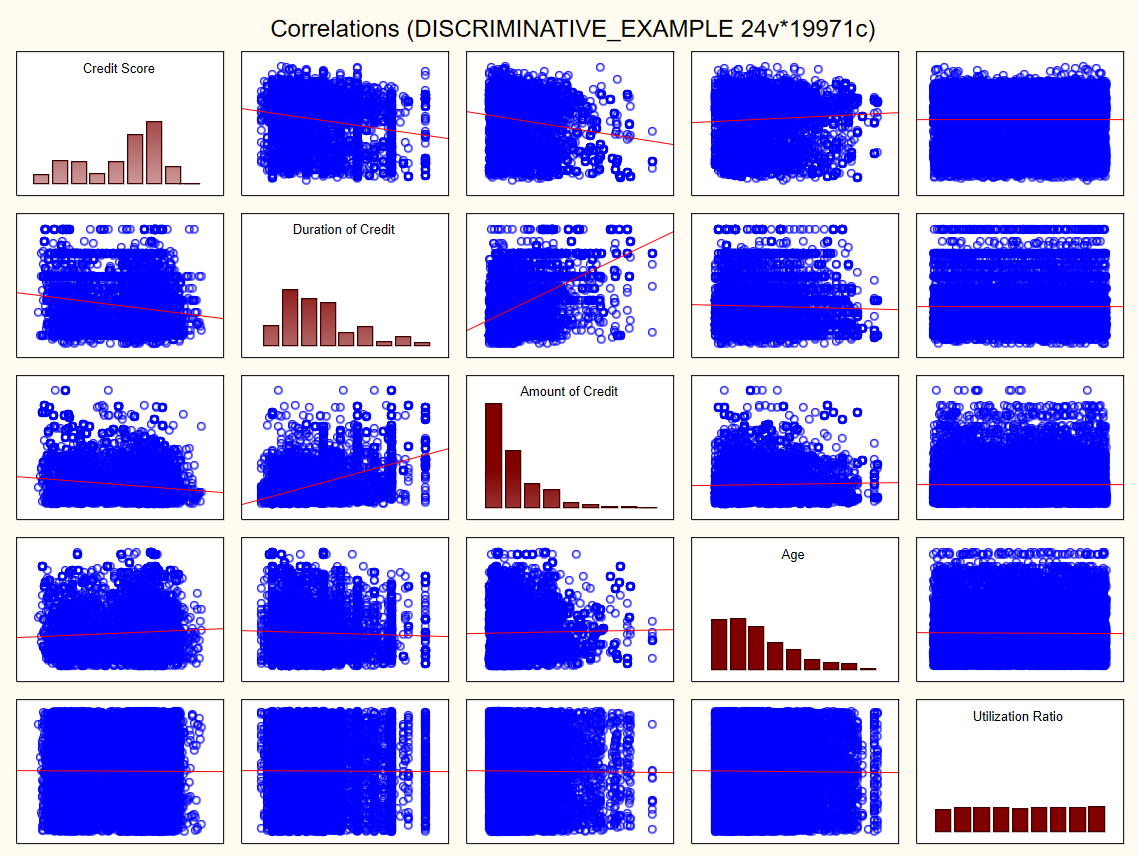
*Возраст*



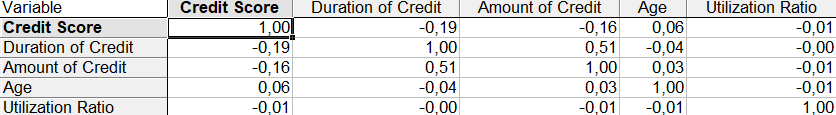
*Коэффициент использования кредита*

Нетрудно заметить, что большинство описанных ранее интервальных признаков не согласуются с нормальным распределением. Так, например, *Возраст*, *Сумма кредита* и *Срок действия кредита* имеют ярко выраженную правостороннюю асимметрию, распределение *Коэффициента использования кредита* во многом напоминает равномерное распределение. Наиболее согласован с распределением Гаусса *Кредитный рейтинг*, однако за счет ранее рассмотренных признаков на используемую модель будут накладываться соответствующие ограничения.

1. Для графического обзора зависимостей между интервальными (числовыми) признаками представим график полных корреляций. Поскольку набор данных содержит достаточно большое число записей, на первый взгляд будет сложно оценить, по каким группам будут распределены клиенты банка. В то же время, можно заметить, что между признаками не наблюдается линейной корреляции (что в будущем будет подтверждено матрицей корреляций).



1. Матрица, описывающая линейную (корреляционную) связь между признаками:



Действительно, линейной связи между рассматриваемыми признаками не наблюдается. Следовательно, все ранее исследованные признаки будут участвовать в составлении классифицирующей модели прогнозирования.

**В) Описание параметров метода (сети)**

Для решения задачи классификации и последующего сравнения результатов было решено использовать следующие методы:

* Классифицирующее дерево (дерево решений)
* Обобщенный дискриминантный анализ
* Линейный дискриминантный анализ

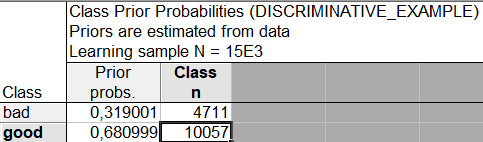
Для оценки точности классификации в каждом из методов будет использоваться матрица ошибок (Classification Matrix).

# Дерево классификации

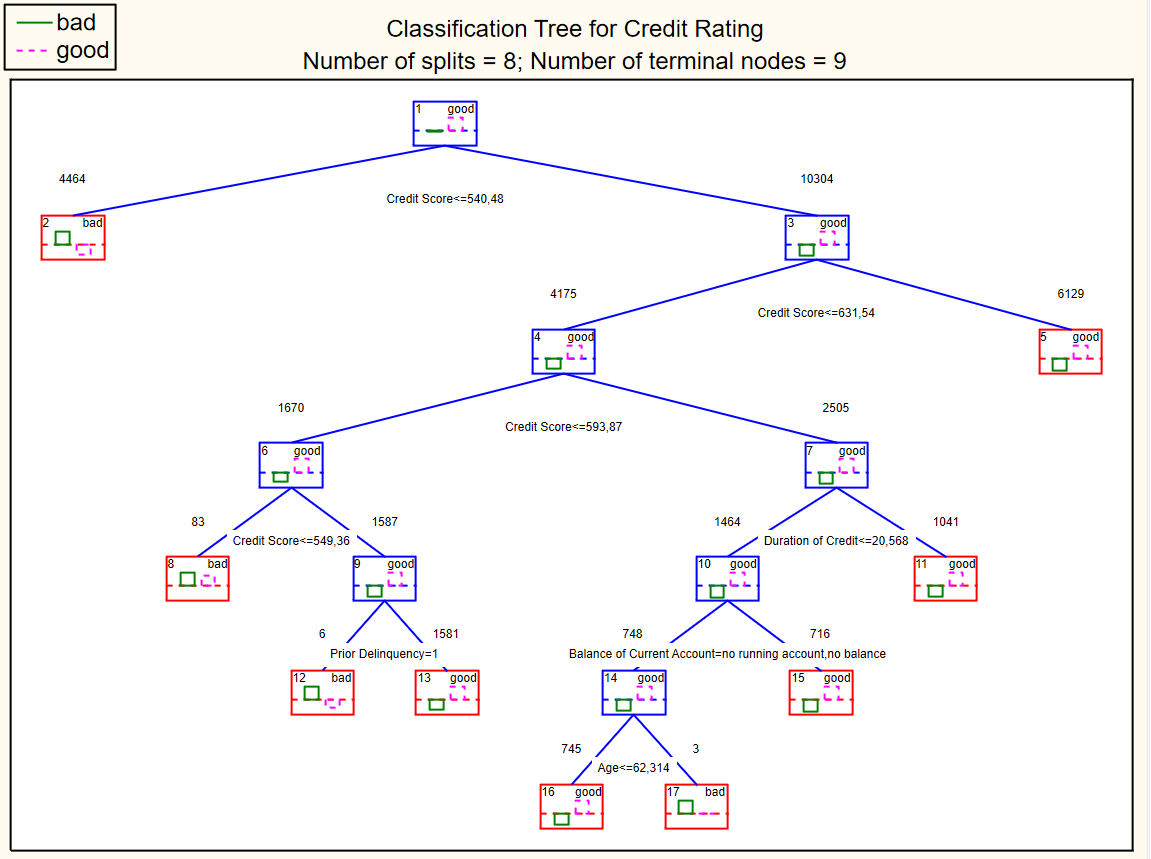
В качестве метода ветвления было выбрано одномерное дискриминантное ветвление QUEST. Особенность такого метода заключается в возможности использования категориальных и порядковых независимых переменных. Кроме того, под порядковыми переменными могут подразумеваться интервальные переменные переменные.

В качестве входного параметра метода будет приниматься критерий остановки по методу FACT. Здесь пользователем задается доля неклассифицированных объектов (Fraction of Objects) – параметр, прямым образом ограничивающий рост дерева (если не ограничивать рост дерева, то мощность дерева, представленная числом узлов, может выйти за программный лимит в 999 узлов). Было принято взять *Долю неклассифицированных объектов* равной 0,01. Теперь терминальные узлы (листья) дерева будут содержать не более 1% неправильно классифицированных объектов.

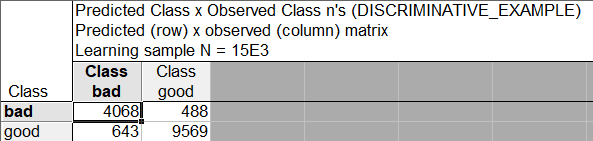
Следующий параметр, рассчитываемый при построении классифицирующего дерева – априорная вероятность. Данный параметр позволяет минимизировать потери, тем самым влияя на точность прогноза. Значения априорных вероятностей, при которых достигается максимальная точность, должны быть пропорциональны размерам классов.



Результат работы алгоритма:



*Графическое представление дерева классификации*



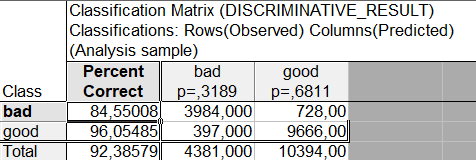
*Матрица ошибок*

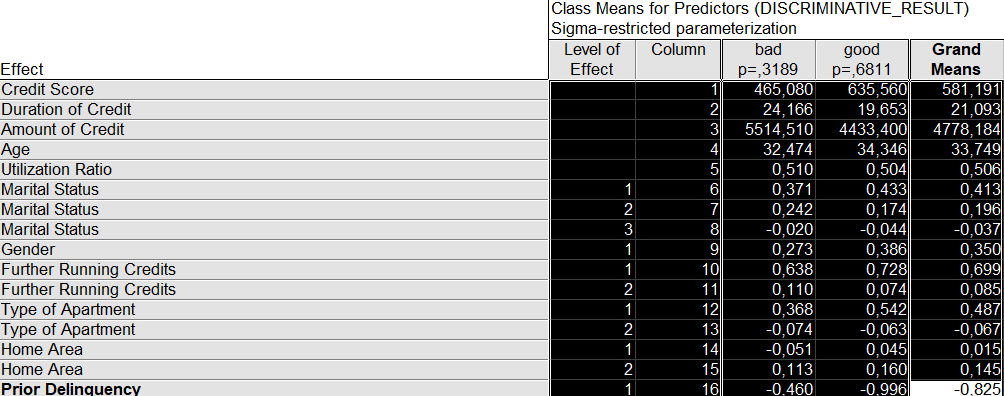
Таким образом, точность классификации (%) для классов bad = precision(bad) = 89,289% и good = precision(good) = 93,703%. Общая точность (accuracy) = 91,496%.

1. Обобщенный дискриминантный анализ

Точно так же, как и для деревьев классификации, данный метод способен принимать в качестве предикторов как порядковые/интервальные переменные, так и категориальные (номинальные). Однако, в силу высокой затраты ресурсов метода, было решено использовать лишь часть от всей группы номинальных переменных (с минимальным числом кодов группирующей переменной).

Параметр, ограничивающий включаемые переменные: p=0,05 (уровень значимости). Используется для модификации метода путем введения пошагового включения переменных, чей уровень значимости не превышает установленный пользователем p и исключения из дискриминантной функции оставшихся переменных. С той же целью используется критерий F, значение которого по умолчанию устанавливается равным 1. Другим параметром, используемым в качестве критерия остановки пошагового анализа, служит толерантность T = (1 - R2) – мера избыточности переменной (по умолчанию принимается равной 0.01).

Результат работы алгоритма:

Средние значения предикторов по классам:

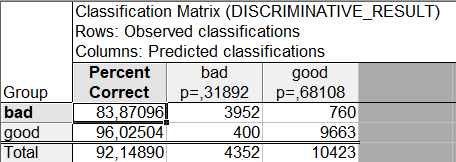
Заметим, что средние по группам в большинстве признаков значительно различаются.

1. Линейный дискриминантный анализ

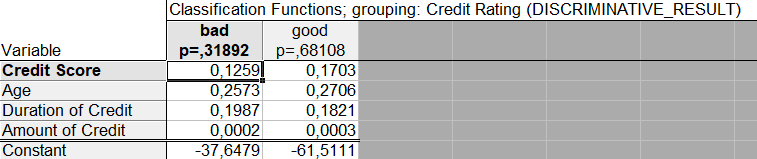
Данный метод используется для построения дискриминантной функции исключительно по интервальным (числовым) переменным. Еще одной отличительной особенностью является обязательность соответствия распределения каждой из предикторной переменной нормальному распределению. Поскольку ранее было сказано, что для исходной выборки это условие нарушается, метод будет выполнен с целью наглядно показать слабые стороны метода и сравнить результаты по ошибкам классификации.

Как и в случае с обобщенным дискриминантным анализом, для пошагового включения переменных будет использоваться ранее рассмотренный уровень значимости p, F и Tolerance (T).

Результат работы алгоритма:



Функции классификации:



Как можно заметить, ранее выдвинутая гипотеза о том, что срок действия кредита (Utilization Ratio) будет исключен из классифицирующей дискриминантной функции, подтверждается.

**В) Сравнительный анализ результатов**

Для классификации клиентов банка на их кредитоспособность наиболее точным оказался метод обобщенного дискриминантного анализа. Однако, несмотря на это, он оказался наиболее затратным, в то время как прирост точности был незначительным. Деревья классификации дали более плохой результат в сравнении с ЛДА. Тем не менее, в ходе нескольких последующих экспериментов было определено, что точность можно повысить таким образом, чтобы превысить соответствующую для обобщенного ДА (до 93% без потери времени). Таким образом, в конечном итоге наименее эффективным оказалась линейная дискриминантная функция (наименьшая точность связана с ограничениями, накладываемыми на предикторы).