МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОНИКИ И МАТЕМАТИКИ НИУ ВШЭ

Кафедра \_\_ИТАС\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

к дипломной работе

На тему:  Разработка и реализация методов оценки правдоподобия бинарной классификации

Студент  Ларионов Алексей Александрович \_

Руководитель работы Подлесных Валерий Григорьевич \_

Допущен к защите \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2013\_\_\_ г.

КОНСУЛЬТАНТ РАБОТЫ:

Специальная часть: Подлесных Валерий Григорьевич

*Зав. кафедрой* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

МОСКВА

**АННОТАЦИЯ**

В данной дипломной работе рассмотрены методы оценки правдоподобия бинарной классификации. В современном бизнесе, экономике, маркетинге, медицине, технике и других отраслях на настоящий момент ставится большое количество задач, связанных с отнесением объекта к одному из двух классов. Оценка правдоподобия бинарной классификации позволяет определить не сам класс, а вероятность того, что наблюдение относится к тому или иному классу, что также очень востребовано с учетом издержек ошибочной классификации.

В работе предложен новый подход к оценке правдоподобия логистической регрессии с использованием шансов вместо вероятностей для расчета коэффициентов логистической регрессии. Оценка шансов производится по базе данных анализируемой предметной области.

Для реализации предлагаемого подхода решены следующие задачи.

1. Выявлены наиболее целесообразные методы оценки достоверности бинарной классификации с точки зрения применимости к выборке «текучесть абонентской базы».
2. Модифицирован и реализован метод вычисления коэффициентов логистической регрессии с помощью метода оценки шансов по базе данных для дихотомической и полихотомической входных переменных.
3. Разработан алгоритм построения ROC-кривых для логистической регрессии с полихотомической входной переменной, позволяющий определить наиболее подходящий предсказывающий атрибут, минимизирующий ошибки при оценке правдоподобия классификации.

СОДЕРЖАНИЕ

[Введение 5](#_Toc358001581)

[Глава 1 Применение логистической регрессии в задачах бинарной классификации 7](#_Toc358001582)

[1.1 Обзор методов оценки достоверности и правдоподобия 7](#_Toc358001583)

[1.1.1 Линейная регрессия 7](#_Toc358001584)

[1.1.2 Логистическая регрессия 9](#_Toc358001585)

[1.1.3 Простой классификатор Байеса (Naive Bayes) 10](#_Toc358001586)

[1.1.4 Lift-диаграммы 11](#_Toc358001587)

[1.1.5 Gain-диаграммы 13](#_Toc358001588)

[1.1.6 ROC-анализ 15](#_Toc358001589)

[1.2 Пример простой бинарной классификации с числовой входной переменной 17](#_Toc358001590)

[1.3 Проблема ошибок 1 и 2 рода и издержек классификации 20](#_Toc358001591)

[1.4 Метод оценки шансов для нахождения коэффициентов логистической регрессии 22](#_Toc358001592)

[Глава 2 Постановка задач исследования 24](#_Toc358001593)

[Глава 3 Теория и практика логистической регрессии в задачах бинарной классификации 26](#_Toc358001594)

[3.1 Понятие логистической регрессии и сфера ее применения 26](#_Toc358001595)

[3.1.1 Уравнение логистической регрессии 27](#_Toc358001596)

[3.1.2 График функции логистической регрессии 28](#_Toc358001597)

[3.2 Метод максимального правдоподобия для нахождения коэффициентов логистической регрессии 28](#_Toc358001598)

[3.3 Разработка метода расчета параметров логистической регрессии на основе оценки шансов 33](#_Toc358001599)

[3.3.1 Разработка метода расчета параметров логистической регрессии на основе оценки шансов для дихотомической входной переменной 36](#_Toc358001600)

[3.3.2 Разработка метода расчета параметров логистической регрессии на основе оценки шансов для полихотомической (категориальной) входной переменной 39](#_Toc358001601)

[Глава 4 Разработка и реализация алгоритма для расчета коэффициентов логистической регрессии и ROC-кривых 45](#_Toc358001602)

[4.1 Разработка алгоритма построения ROC-кривых 45](#_Toc358001603)

[4.2 Реализация разработанного метода нахождения коэффициентов логистической регрессии и алгоритма построения ROC-кривых 49](#_Toc358001604)

[4.2.1 Исходные данные программы 49](#_Toc358001605)

[4.2.2 Реализация алгоритма нахождения коэффициентов логистической регрессии для дихотомической входной переменной 51](#_Toc358001606)

[4.2.3 Реализация алгоритма нахождения коэффициентов логистической регрессии для полихотомической входной переменной 57](#_Toc358001607)

[4.2.4 Реализация алгоритма построения ROC-кривых 65](#_Toc358001608)

[Заключение 73](#_Toc358001609)

[Список используемой литературы 74](#_Toc358001610)

[Приложения 76](#_Toc358001613)

[Приложение 1. Листинг программы 76](#_Toc358001614)

# Введение

В настоящий момент существует множество методов анализа и классификации, направленных на построение наиболее точных и эффективных математических моделей, которые широко используются в экономике, медицине, сфере телекоммуникационных услуг, маркетинге и других областях, в качестве инструмента для принятия решения. Существующие методы позволяют выявить имеющиеся в выборке взаимосвязи между входными и выходными переменными, построить модели, отражающие эти взаимосвязи, сравнить их с точки зрения точности, прибыльности и издержек и выбрать наиболее эффективную модель.

В данной дипломной работе рассмотрены методы оценки правдоподобия применительно к бинарной классификации, потому что в современном бизнесе, экономике, маркетинге, медицине, технике и других отраслях на сегодняшний день ставится большое количество задач, связанных с отнесением объекта к одному из двух классов. В медицине это может быть наличие или отсутствие заболевания, в маркетинге — наличие или отсутствие отклика респондента, в военной технике — наличие или отсутствие цели. Оценка правдоподобия бинарной классификации позволяет определить не сам класс, а вероятность того, что наблюдение относится к тому или иному классу, что также очень востребовано.

К методам оценки правдоподобия бинарной классификации относятся логистическая регрессия, простой классификатор Байеса, Lift, Gain и Roc диаграммы. К сожалению, данные методы имеют множество недостатков, не позволяющих им осуществить наиболее достоверную оценку.

Цель данной дипломной работы – разработка методов оценки правдоподобия бинарной классификации и их реализация на примере выборки «текучесть абонентской базы».

Для достижения этой цели необходимо решить следующие задачи.

1. Выявить наиболее целесообразные методы оценки достоверности бинарной классификации с точки зрения применимости к выборке «текучесть абонентской базы».
2. Разработать метод оценки шансов для нахождения коэффициентов логистической регрессии для дихотомической и полихотомической входной переменной.
3. Разработать алгоритм построения ROC-кривых для логистической регрессии с полихотомической входной переменной.
4. Реализовать разработанный метод оценки шансов для нахождения коэффициентов логистической регрессии для дихотомической и полихотомической входной переменной и разработанный алгоритм построения ROC-кривых для полихотомической входной переменной на примере выборки «текучесть абонентской базы» путем написания программы на языке Delphi.

Разработанные методы оценки правдоподобия бинарной классификации, реализованные в программном обеспечение, могут быть использованы в качестве надежного инструмента принятия решения аналитиками в различных предметных областях, включая медицину, телекоммуникационные слуги, экономику, маркетинг и д.р.

# Глава 1 Применение логистической регрессии в задачах бинарной классификации

## 1.1 Обзор методов оценки достоверности и правдоподобия

В настоящее время для решения задач, связанных с обработкой и анализом большого количества разнородных данных, разработано огромное количество методов и алгоритмов классификации и регрессии, включая их модификации. В данном обзоре рассматриваются следующие методы оценки достоверности и правдоподобия:

* логистическая регрессия;
* простой классификатор Байеса;
* Lift, Gain и Roc диаграммы.

Так же в обзоре рассматривается линейная регрессия в качестве инструмента построения математических моделей для лучшего понимания достоинств логистической регрессии.

### 1.1.1 Линейная регрессия

Линейная регрессия используется для моделирования линейных зависимостей между непрерывной выходной переменной и набором входных переменных. «При определенных условиях уравнение линейной регрессии служит незаменимым и очень качественным инструментом анализа и прогнозирования» [1]. Модель линейной регрессии является наиболее распространенным и простым уравнением зависимости между входными и выходными переменными. Кроме того, построенное уравнение линейной регрессии может быть начальной точкой анализа данных.

Теоретическая линейная регрессионная модель имеет вид:

,

Где:

– выходная (зависимая) переменная модели;

– входные (независимые) переменные;

– коэффициенты линейной регрессии;

– свободный член;

− оценка теоретического случайного отклонение.

«Для определения значений теоретических коэффициентов линейной регрессии необходимо знать и использовать все значения переменных *x* и *y* генеральной совокупности, что практически невозможно» [1].

По выборке ограниченного объема возможно построить так называемое эмпирическое уравнение регрессии:

*,*

Где:

– выходная (зависимая) переменная модели;

– входные (независимые) переменные;

– коэффициенты линейной регрессии;

– свободный член;

− оценка теоретического случайного отклонение.

Коэффициенты подбираются таким образом, чтобы на заданный входной вектор регрессионная модель формировала желаемое выходное значение . «Эмпирические коэффициенты регрессии и являются оценками теоретических коэффициентов и , а само уравнение отражает общую тенденцию в поведении рассматриваемых переменных. Индивидуальные значения переменных в силу различных причин могут отклоняться от модельных значений» [1].

Одним из наиболее привлекательных свойств линейной регрессии является то, что ее коэффициенты могут быть получены с помощью метода наименьших квадратов [5].

Однако, линейная регрессионная модель не универсальна. Когда для решения задачи строят модель линейной регрессии, на значения зависимой переменной обычно не налагают никаких ограничений. Но на практике такие ограничения часто встречаются и могут быть весьма существенными. Например, выходная переменная может быть категориальной или бинарной. В таких случаях используют логистическую регрессию.

### 1.1.2 Логистическая регрессия

При анализе данных часто встречаются задачи, где выходная переменная является категориальной, и тогда использование линейной регрессии затруднено. Поэтому при поиске связей между набором входных переменных и категориальной выход­ной переменной получила распространение логистическая регрессия. Логистическая регрессия является методом бинарной классификации. Она позволяет оценивать вероятность реализации (или не реализации) события в зависимости от значений некоторых независимых переменных. Линия логистической регрессии, в отличие от линейной, не является прямой.

Условное среднее для логической регрессии имеет вид:

,

Где:

*е* – основание натурального логарифма;

*ρ* - вероятность того, что произойдет интересующее событие;

*β0, β1* – коэффициенты логистической регрессии.

*х* – значение независимой (входной) переменной.

Эту функцию называют логистической. Значения изменяются в диапазоне от 0 до 1. Если предположить, что значение выходной переменной , равное 1, рассма­тривается как успех, а значения 0 — как неуспех, то можно интерпретировать как вероятность успеха, а – неуспеха.

Для оценки коэффициентов логистической регрессии метод наименьших квадратов не применим. Поэтому ее коэффициенты оцениваются на основе метода максимального правдоподобия. Логарифмическая функция правдоподобия имеет вид:

*.*

В книге Н. Паклина и В. Орешкова «Бизнес-аналитика: от данных к знаниям» оценки максимального правдоподобия были найдены путем дифференцирования относительно каждого параметра и приравниваем результирующих выражений к 0 [5]. Коэффициенты и в полученной системе уравнений могут быть найдены только вычислительным путем. Аналитическое решение данной системы уравнений авторами найдено не было.

### 1.1.3 Простой классификатор Байеса (Naive Bayes)

Байесовский подход объединяет группу алгоритмов классификации, основанных на принципе условной вероятности: для объекта с помощью формулы Байеса определяется апостериорная вероятность принадлежности к каждому классу и выбирается тот класс, для которого она максимальна. «Изначально байесовская классификация использовалась для формализации знаний экспертов в экспертных системах, сейчас байесовская классификация также применяется в качестве одного из методов Data Mining» [2].

Особое место в данной области занимает простой классификатор Байеса (Naive Bayes), в основе которого лежит предположение о независимости признаков, описывающих классифицируемые объекты. Это предположение значительно упрощает задачу, поскольку вместо сложной процедуры оценки многомерной плотности вероятности требуется оценка нескольких одномерных. К сожалению, на практике предположение о независимости признаков редко выполняется, является «наивным», что и дало название методу.

К основным преимуществам наивного байесовского классификатора можно отнести легкость программной реализации и низкие вычислительные затраты. В тех редких случаях, когда признаки действительно независимы (или близки к этому), он оптимален. Главный его недостаток — относительно низкое качество классификации в большинстве реальных задач. Поэтому чаще всего его используют как примитивный эталон для сравнения различных моделей, либо как блок для построения более сложных алгоритмов [7].

Хотя предположение о статистической независимости признаков на практике выполняется достаточно редко, в Data Mining существуют различные методы, которые позволяют отбирать наименее коррелированные из них. Использование таких методов позволяет повысить эффективность байесовских классификаторов.

### 1.1.4 Lift-диаграммы

Наличие издержек классификации делает актуальной не только оценку эффективности модели, но и оптимизацию ее с точки зрения издержек. Для решения данной задачи может быть построено несколько моделей, после чего потребуется проверить, какая из них более эффективна. Чтобы выбрать модель, которая позволит найти оптимальное соотношение между издержками и прибылью, для каждой модели-кандидата строится график, называемый Lift-кривой (Lift curve). Семейство Lift-кривых образует Lift-диаграмму (Lift chart). Поведение Lift-кривых и их взаимное расположение на диаграмме позволяют сделать вывод о том, какая из моделей работает лучше. С помощью Lift-диаграмм также можно определить размер выборки, который даст наилучшее соотношение издержек и прибыли.

Однако, при оценке эффективности моделей с помощью Lift-диаграмм можно столкнуться со следующей проблемой: одна модель лучше работает в одном диапазоне (например, при объемах выборки 70-100%), а вторая — в другом (0-50%). Поэтому Lift-диаграммы не всегда применимы и эффективны в решении отдельных задач.

1000

500

0

100

50

Размер выборки, %

Число откликов

Рисунок 1. Lift-диаграмма

Lift-диаграмма изображена на рис. 1. «По горизонтальной оси Lift-диаграммы откладывается размер выборки в процентах от общего числа наблюдений в упорядоченном списке, а по вертикальной — число полученных положительных исходов. Каждая точка Lift-кривой показывает ожидаемое число положительных исходов для выборки заданного размера. Точка с координатами (0; 0) соответствует нулевому числу наблюдений и, следовательно, положительных исходов» [5]. Точка с координатами (100; 1000) соответствует максимальному числу наблюдений с исключительно положительными исходами.

Диагональная линия на Lift-диаграмме отражает работу так называемого произвольного (случайного) классификатора. Модель, которой соответствует диагональная линия, бесполезна. В то же время, любая модель, Lift-кривая которой проходит выше диагональной прямой, дает прирост числа положительных исходов относительно случайной модели и обеспечивает лифт. Классификатор, проходящей ниже диагональной прямой, дает отрицательный лифт и будет хуже «бесполезного», поэтому применение таких моделей с точки зрения оптимизации издержек бессмысленно. Следовательно, чем выше проходит Lift-кривая, тем эффективнее модель. При прочих равных условиях предпочтение следует отдать той модели, для которой Lift-кривая проходит выше других или максимально приближается к точке с координатами (0; 1000).

Для сравнения эффективности моделей, описываемых Lift-кривыми, часто используется площадь под ними (area under curve, AUC). Чем выше проходит график, тем больше площадь под кривой. Этот числовой показатель удобен, когда Lift-кривые двух моделей проходят близко друг к другу и визуально оценить разницу между ними трудно.

### 1.1.5 Gain-диаграммы

Еще одна разновидность Lift-диаграмм — Gain-диаграмма (Gain chart), или кумулятивная Lift-диаграмма (cumulative lift chart), изображена на рис. 2.

Методика использования Gain-диаграмм в целом похожа на методику использования обычных Lift-диаграмм, однако, в отличие от Lift-диаграмм, с помощью Gain-диаграмм ищется оптимальное соотношение между размером выборки и количеством ошибок классификатора [5].

0

0,5

1

100

50

Размер выборки, %

Чувствительность, Se

Рисунок 2. Gain-диаграмма

По горизонтальной оси Gain-диаграммы также откладывается размер выборки, полученной из упорядоченного списка. Но по вертикальной оси фиксируется не количество (или процент) положительных наблюдений, попавших в выборку, а отношение числа истинноположительных наблюдений к числу всех наблюдений, классифицированных как положительные» [5]. Данная величина называется чувствительностью (*Se*). При увеличении размера выборки (смещении по диаграмме вправо), увеличивается количество ложноположительных наблюдений, соответственно, риск ошибочной классификации растет. При уменьшении выборки (смещении по диаграмме влево), понижается число ошибок. Задача заключается в подборе размера выборки таким образом, чтобы чувствительность модели была на достаточно высоком уровне. Если использовать Gain-диаграмму для сравнения моделей, то лучшей также будет та из них, Gain-кривая которой проходит выше.

Использовать Gain-кривые для сравнения моделей удобнее, чем обычные Lift-кривые, поскольку последние в большей степени применяются для оптимизации издержек ошибок классификации, а первые непосредственно указывают на точность модели. Однако, использование Gain-кривых сопряжено с тем же ограничением, что и использование Lift-диаграмм – они не универсальны и не могут быть применены для решения отдельных задач. Обе выше перечисленные диаграммы не позволяют численно измерить издержки или выгоды от использования модели.

### 1.1.6 ROC-анализ

ROС-анализ представляет собой графическую методику оценивания эффективности моделей с помощью двух показателей — чувствительности (*Se*) и специфичности (*Sp*).

В задачах бинарной классификации, когда модель предсказывает вероятность того, что наблюдение относится к одному из двух классов, очень важен выбор точки отсечения, то есть порога вероятности, разделяющего два класса. Такая точка отсечения показывает, после какого значения вероятности на выходе модели один класс сменяется другим. Выбирая точку отсечения, мы управляем вероятностью правильного распознавания положительных и отрицательных примеров. При уменьшении порога отсечения увеличивается вероятность ошибочного распознавания положительных наблюдений (ложноположительных исходов), а при увеличении возрастает вероятность неправильного распознавания отрицательных наблюдений (ложноотрицательных исходов).

Цель ROС-анализа заключается в том, чтобы подобрать такое значение точки отсечения, которое позволит модели с наибольшей точностью распознавать положительные или отрицательные исходы и выдавать наименьшее количество ложноположительных или ложноотрицательных ошибок, соответственно.

ROС-анализ широко используется в медицине, биологии, маркетинге, банковском деле и других областях, где применяется бинарная вероятностная классификация.

В основе ROC-анализа лежат ROC-кривые. График ROC-кривой изображен на рис. 3.

Для построения ROC-кривой требуется изменять порог отсечения в диапазоне от 0 до 1 с заданным шагом, например 0,01. Чем меньше шаг, тем больше будет точек на графике, что косвенно влияет на точность построения ROC-кривой. При каждом значении порога заново рассчитываются значения специфичности (Sp) и чувствительности (Se), т.е. меняется количество распознанных ошибок I и II рода. Чувствительность (Se) откладывается по оси ординат, а по оси абсцисс откладывается значение 100-Sp.

0

100

100

50

50

Se, %

100-Sp, %

Рисунок 3. ROC-кривая и идеальная ROC-кривая

Еще одно полезное свойство ROC-кривой заключается в том, что она позволяет оценить качество модели бинарной классификации. Определить степень качества модели можно по кривизне кривой: чем ближе она к идеальному классификатору, тем качественнее модель. Если же кривая близка к диагонали, то модель бесполезна.

Визуально сравнить ROC-кривые и выявить более эффективную модель не всегда представляется возможным. Поэтому их можно сравнивать с помощью подсчета площади под кривой (AUC). Площадь под кривой характеризует прогностическую силу модели, при этом AUC=1 соответствует идеальному классификатору, который не достижим на практике, а AUC=0,5 соответствует бесполезному классификатору.

В результате анализа существующих методов оценки достоверности бинарной классификации было выявлено, что модель, описывающая зависимость между входными переменными и выходной бинарной переменной выборки «текучесть абонентской базы» должна быть построена с помощью логистической регрессии, а оценка качества данной модели должна быть произведена с помощью построения ROC-кривых.

## 1.2 Пример простой бинарной классификации с числовой входной переменной

Рассмотрим пример простой логистической регрессии с числовой входной переменной, приведенный в книге Н.Б. Паклина [5].

Пусть врача интересует зависимость между возрастом пациента и наличием или отсутствием у пациента какого-либо заболевания. Данные, касающиеся 20-ти пациентов, представлены в табл. 1. В данной таблице присутствуют следующие поля: «№ пациента», «Возраст пациента, *х*» и «Наличие заболевания, *у*».

Поле «№ пациента», содержит порядковый номер пациента.

Поле «Возраст пациента, *х*», содержит значения входной числовой переменной «Возраст пациента» и обозначается, как переменная «*х*».

Поле «Наличие заболевания, *у*», содержит значения выходной бинарной переменной «Наличие заболевания» и обозначается, как переменная «*y*». Она может принимать только два значения 0 и 1.

Таблица 1

Данные о возрасте пациентов и наличии у них некоторого заболевания

|  |  |  |
| --- | --- | --- |
| № пациента | Возраст пациента, *х* | Наличие заболевания, *у* |
| 1 | 25 | 0 |
| 2 | 29 | 0 |
| 3 | 30 | 0 |
| 4 | 31 | 0 |
| 5 | 32 | 0 |
| 6 | 41 | 0 |
| 7 | 41 | 0 |
| 8 | 42 | 0 |
| 9 | 44 | 1 |
| 10 | 49 | 1 |
| 11 | 50 | 0 |
| 12 | 59 | 1 |
| 13 | 60 | 0 |
| 14 | 62 | 0 |
| 15 | 68 | 1 |
| 16 | 72 | 0 |
| 17 | 79 | 1 |
| 18 | 80 | 0 |
| 19 | 81 | 1 |
| 20 | 84 | 1 |

[5]

Диаграмма, изображенная на рис. 4, наглядно демонстрирует возможности логистической регрессии по отношению к линейной регрессии.

0

0,5

1

25

35

45

55

65

75

85

Вероятность наличия заболевания, доли единицы

Возраст пациента, лет

Рисунок 4. Диаграмма соответствия линий линейной и логистической регрессии на примере выборки данных о возрасте пациентов и наличии у них некоторого заболевания [5]

Сплошной линией на диаграмме изображена прямая простой линейной регрессии, а пунктиром - кривая логистической регрессии, построенные на основе данных таблицы 1. Для обеих кривых кружками обозначены ошибки оценивания заболевания у пациентов.

По данной диаграмме видно, что линия линейной регрессии равномерно возрастает и большое количество ее значений лежит рядом с вероятностью наличия заболевания, равной 50%. Однако, вероятность в 50% и близкие к ней вероятности не дают точно отнести событие к положительному или отрицательному исходу. Применительно к данной выборке это наличие или отсутствие заболевания. В то же время, у аналитиков возникают трудности с классификацией таких событий.

Линия логистической регрессии имеет порог в районе 50% вероятности, а значит, меньшее количество событий будет иметь вероятности в диапазоне 50%. Это позволит аналитику точно идентифицировать большее количество событий, как положительный или отрицательный исход.

В результате анализа диаграммы выявлено, что логистическая регрессия является более точной моделью для оценивания вероятности наличия заболевания у пациента в зависимости от его возраста, чем линейная регрессия, и сделан вывод о том, что логистическая регрессия имеет более высокую предсказательную способность по сравнению с линейной регрессией.

## 1.3 Проблема ошибок 1 и 2 рода и издержек классификации

В процессе работы бинарной классификационной модели могут возникать ошибки – несоответствия выходных значений модели и реальных значений выборки. Предсказание, соответствующее положительному исходу, может быть распознано моделью как отрицательный исход. Такие ошибки называются ложноотрицательными или ошибками I рода. Предсказание, соответствующее отрицательному исходу, наоборот, может быть распознано моделью как положительный исход. Такие ошибки называются ложноположительными или ошибками II рода. Поэтому в бинарной классификации каждое предсказание может иметь следующие четыре исхода, которые приведены в табл. 2:

* истинноположительный (true positive, *TP*);
* истинноотрицательный (true negative, *TN*);
* ложноположительный (false positive, *FP*);
* ложноотрицательный (false negative, *FN*).

Таблица 2

Матрица классификации

|  |  |  |
| --- | --- | --- |
| Фактически класс | Предсказанный класс | |
| Да | Нет |
| Да | *TP* | *FN* |
| Нет | *FP* | *TN* |

В таблице 2 под фактическим классом понимается то, что конкретное событие по данным выборки действительно имеет положительный или отрицательный исход. В тоже время, предсказанный класс говорит о том, что событие было отнесено моделью к положительному или отрицательному исходу, что может противоречить или соответствовать реальным данным.

Точность бинарной классификационной модели можно оценить с помощью четырех показателей: общего показателя успеха, общего показателя ошибки, чувствительности и специфичности.

Общий показатель успеха (overall success rate, *OSR*) представляет собой число правильно классифицированных наблюдений, отнесенное к общему числу наблюдений:

.

Иногда данную величину называют точностью классификатора.

Общий показатель ошибки (overall error rate, *OVR*) определяется отношением числа неправильно классифицированных наблюдений к общему числу наблюдений:

.

Чувствительность (sensitivity, *Se*) определяется как отношение числа истинноположительных наблюдений к числу фактически положительных наблюдений:

.

Специфичность (specificity, *Sp*) определяется как отношение числа истинноотрицательных наблюдений к числу фактически отрицательных наблюдений:

.

В бинарной классификации два типа ошибок — ложноположительный и ложноотрицательный — будут давать различные издержки, а два типа корректной классификации — истинноположительный и истинноотрицательный — будут приносить различную прибыль.

Минимизировать издержки ошибок классификации можно с помощью выбора точки отсечения (cut-off point) — порогового значения вероятности, разделяющего классы. При увеличении порога (приближении его к 1), увеличивается вероятность ложноположительных исходов и уменьшается вероятность ложноотрицательных. При снижении порога, наоборот, уменьшается вероятность ложноположительных исходов и увеличивается вероятность ложноотрицательных. Если известно, какой из двух типов ошибок обходится дороже, то можно подобрать порог так, чтобы минимизировать связанные с ним издержки. Если событие, связанное с положительным исходом классификации, является редким, то издержки ложноотрицательных ошибок обычно больше, чем издержки ложноположительных. Для таких случаев порог отсечения должен задаваться существенно ниже, чем 0,5.

## 1.4 Метод оценки шансов для нахождения коэффициентов логистической регрессии

Модель логистической регрессии учитывает отношение шансов, поэтому для нахождения коэффициентов логистической регрессии может применяться метод оценки шансов.

«Шанс определяется как вероятность того, что событие произошло (шанс успеха), разделенная на вероятность того, что событие не произошло (шанс неуспеха). Шансы и вероятности содержат одну и ту же информацию, но по-разному выражают ее. Если вероятность того, что событие произойдет, обозначить как *ρ*, то шанс возникновения этого события будет равен *ρ/*(1- *ρ*)» [5].

В бинарной логистической регрессии с дихотомической входной переменной вероятность того, что выходная переменная примет значение *у*=1 (событие произошло) при *х*=1, может быть записана в виде:

,

Где:

*е* – основание натурального логарифма;

*β0, β1* – коэффициенты логистической регрессии.

*х* – значение независимой (входной) переменной.

Аналогично вероятность того, что выходная переменная примет значение *у*=1 (событие произошло) для наблюдений, в которых *х*=0, может быть записана как:

.

Отношение шанса того, что выходная переменная примет значение 1 (событие произошло), к шансу того, что переменная примет значение 0 (событие не произошло) имеет вид:

.

Благодаря последней формуле становится возможным расчет коэффициентов логистической регрессии для дихотомической и полихотомической входной переменной. Расчет коэффициентов логистической регрессии методом оценки шансов для дихотомической и полихотомической переменной приведен в третьей главе дипломной работы.

# Глава 2 Постановка задач исследования

В результате анализа существующих методов оценки достоверности бинарной классификации было выявлено, что модель, описывающая зависимость между входными и выходными переменными выборки «текучесть абонентской базы», должна быть построена с помощью логистической регрессии, а оценка качества данной модели должна быть произведена с помощью построения ROC-кривых.

Для реализации этой цели необходимо выполнить следующие задачи.

1) Разработать метод оценки шансов для нахождения коэффициентов логистической регрессии для дихотомической и полихотомической входной переменной.

3) Разработать алгоритм построения ROC-кривых для логистической регрессии с полихотомической входной переменной.

4) Реализовать разработанный метод оценки шансов для нахождения коэффициентов логистической регрессии для дихотомической и полихотомической входной переменной и разработанный алгоритм построения ROC-кривых для полихотомической входной переменной на примере выборки «текучесть абонентской базы» путем написания программы на языке Delphi.

* Рассчитать коэффициенты логистической регрессии для дихотомической входной переменной с помощью метода оценки шансов.
* Рассчитать отношения шансов для данной переменной.
* Рассчитать вероятности положительного исхода выходной переменной в зависимости от значений входной переменной.
* Рассчитать коэффициенты логистической регрессии для полихотомической входной переменной с помощью метода оценки шансов.
* Разбить входную переменную на категории «низкое», «среднее», «высокое».
* Рассчитать отношения шансов для категорий «среднее» и «высокое» по отношению к категории «низкое».
* Рассчитать вероятности положительного исхода выходной переменной для категорий «низкое», «среднее», «высокое» входной переменной.
* Построить ROC-кривые для категорий «низкое», «среднее», «высокое» входной полихотомической переменной.
* Рассчитать площади под кривыми (AUC) для каждой категории входной переменной.

В качестве исходных данных для реализации программы была выбрана выборка «текучесть абонентской базы» из-за репрезентативности данной выборки, наличия в ней ряда входных полихотомических и дихотомических переменных, независимых между собой, и выходной бинарной переменной «Уход» клиента, которая зависима от входных переменных.

# Глава 3 Теория и практика логистической регрессии в задачах бинарной классификации

## 3.1 Понятие логистической регрессии и сфера ее применения

В анализе данных часто встречаются задачи, где выходная переменная является категориальной, и использование линейной регрессии затруднено. Поэтому в поиске связей между набором входных переменных и категориальной выход­ной переменной, получила распространение логистическая регрессия.

«Логистическая регрессия или логит-регрессия - разновидность множественной регрессии, назначение которой состоит в анализе связи между несколькими независимыми (входными) переменными и зависимой (выходной) переменной» [6].

Логистическая регрессия применяется в случаях, когда зависимая (выходная) переменная является бинарной, то есть может принимать только два значения: 1 или 0 [9]. При этом объясняющие (входные) переменные должны быть независимы друг от друга.

Логит-регрессия является методом бинарной классификации. Она позволяет оценивать вероятность реализации (или не реализации) события в зависимости от значений некоторых независимых переменных.

Модель логистической регрессии используется в медицине при проведении клинических исследований, в банковском деле при построении рейтинга заемщика и управлении кредитными рисками, в маркетинге для моделирования поведения клиента и т. д. [7].

### 3.1.1 Уравнение логистической регрессии

Все регрессионные модели могут быть записаны в виде функции

. Например, в множественной линейной регрессии зависимая (выходная) переменная является линейной функцией независимых (входных) переменных:

.

Для построения модели логистической регрессии, уравнение регрессии преобразуют таким образом, чтобы значения переменной *y* никогда не выходили за интервал [0,1]. Такое преобразование называется логит*-*преобразованием [8].

В результате логит-преобразования уравнение логистической регрессии с одной входной переменной принимает вид:

,

Где:

*е* – основание натурального логарифма;

*ρ* - вероятность того, что произойдет интересующее событие;

*β0, β1* – коэффициенты логистической регрессии.

*х* – значение независимой (входной) переменной.

Уравнение логистической регрессии с несколькими входными переменными имеет вид:

,

Где:

*–* переменная logit*;*

*х –* значения независимых входных переменных.

### 3.1.2 График функции логистической регрессии

График функции логистической регрессии изображен на рис. 5.

0

0,5

1

25

35

45

55

65

75

85

Рисунок 5. График функции логистической регрессии

По оси абсцисс отложена переменная logit (), по оси ординат - .

Особенностью логистической функции является то, что она может принимать любые входные значения от минус бесконечности до плюс бесконечности, тогда как выходная переменная ограничена диапазоном [0; 1].

## 3.2 Метод максимального правдоподобия для нахождения коэффициентов логистической регрессии

«Метод максимального правдоподобия очень часто используется в статистике для получения оценок параметров генеральной совокупности по данным выборки. Он основан на функции правдоподобия» [5].Функция правдоподобия определяет вероятность появления значений параметров *β=β1,β2,β3,...,βn* для данного значения *х.* Задача метода максимального правдоподобия заключается в поиске таких значений этих параметров, которые максимизируют функцию правдоподобия. То есть метод максимального правдоподобия строит оценки параметров *β=β1,β2,β3,...,βn*, наилучшим образом описывающие данные выборки.

Рассмотрим функцию максимального правдоподобия. Для этого введем следующие условные обозначения.

Пусть вероятность того, что выходная переменная *y* приобретет значение 1 для заданного значения *x* (вероятность успеха), будет *ρ*(*x*) *= P*(*y=*1*|x*), а вероятность того, что *y*=0 при заданном *x*, будет 1-*ρ*(*x*) = *P*(*y=*0*|x*).

Тогда, при *yi*=0 или 1, вклад *i*-го наблюдения может быть выражен как

.

Предположение, что наблюдения выборки являются независимыми, позволяет представить функцию правдоподобия как произведение двух отдельных членов:

.

В вычислительном плане более удобна логарифмическая функция правдоподобия

,

Где:

;

.

Согласно утверждению Н.Б. Паклина, оценки максимального правдоподобия могут быть найдены путем дифференцирования относительно каждого параметра и приравнивания полученных выражений к 0 [5].

Проверим это утверждение и составим систему уравнений для нахождения коэффициентов и логистической регрессии по данным выборки «Данные о возрасте пациентов и наличии у них некоторого заболевания» из книги Н.Б. Паклина, рассмотренной ранее в первой главе данной дипломной работы [5].

Продифференцируем по и приравняем полученное выражение к нулю:

.

Возьмем производную левого слагаемого выражения:

,

.

Теперь продифференцируем правое слагаемое выражения:

,

,

.

С учетом двух продифференцированных слагаемых производная по исходного выражения примет вид:

.

Отбросив знаменатель, который не влияет на правую часть уравнения, получим первое уравнение системы:

.

Теперь продифференцируем по и приравняем полученное выражение к нулю:

.

Дифференцирование по происходит по схеме, аналогичной дифференцированию по , с той разницей, что:

.

Для первого слагаемого после дифференцирования по получим следующее выражение:

.

Для второго слагаемого после дифференцирования по получим следующее выражение:

.

С учетом двух продифференцированных слагаемых производная по исходного выражения примет вид:

.

Отбросив общий знаменатель, получим второе уравнение системы:

.

Итак, система уравнений для нахождения коэффициентов логистической регрессии и , приведена ниже:

.

Проверим правильность полученной системы уравнений, используя данные, приведенные в книге Н.Б. Паклина и в таблице 1 первой главы данной дипломной работы [5].

В табл. 1 присутствуют следующие поля: «№ пациента», «Возраст пациента, *х*» и «Наличие заболевания, *у*». Поле «Возраст пациента, *х*», содержит значения входной числовой переменной «Возраст пациента» и обозначается, как переменная «*х*». Поле «Наличие заболевания, *у*», содержит значения выходной бинарной переменной «Наличие заболевания» и обозначается, как переменная «*y*». Она может принимать только два значения 0 и 1.

Запишем первое уравнение из приведенной выше системы уравнений в следующем виде:

.

В левой части уравнения суммируются , которые равны единице. В правой части суммируются , которые равны нулю. В правую часть уравнения подставляются из таблицы 1 , которым соответствуют , равные нулю. Соответственно, в левой части, наоборот, суммируются такие , которых равны единице.

Подставив посчитанные в книге Н.Б. Паклина значения коэффициентов и для данной выборки, получим тождество, что свидетельствует о правильности выполненных преобразований [5].

Аналогичная проверка второго уравнения так же дает тождество, что говорит о правильности второго уравнения системы.

Из полученной системы уравнений нельзя выразить и аналитическим путем, решение системы может быть получено только итерационным численным методом.

Поэтому автор дипломной работы считает, что метод оценки максимального правдоподобия не позволяет найти коэффициенты уравнения регрессии аналитическим путем, и не подходит для реализации целей, поставленных в дипломной работе.

## 3.3 Разработка метода расчета параметров логистической регрессии на основе оценки шансов

Все расчеты, представленные в данной дипломной работе, выполнены по выборке «текучесть абонентской базы», взятой из CD приложения к учебному пособию Паклина Н.Б. [5] Фрагмент данной выборки приведен в табл. 3.

Таблица 3

Фрагмент выборки «текучесть абонентской базы»

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Штат | Длина счета | Код | № телефона | Международные звонки | Голосовая почта | Количество голосовых сообщений | Использовано дневных минут | Количество звонков днем | Абонентская плата 1 | Использовано вечерних минут | Количество звонков вечером | Абонентская плата 2 | Использование ночных минут | Количество ночных звонков | Абонентская плата 3 | Минут международных разговоров | Число международных звонков | Абонентская плата 4 | Число обращений в сервисную службу | Уход |
| KS | 128 | 415 | 382-4657 | нет | да | 25 | 265,10 | 110 | 45,07 | 197,40 | 99 | 16,78 | 244,70 | 91 | 11,01 | 10,00 | 3 | 2,70 | 1 | Нет |
| OH | 107 | 415 | 371-7191 | нет | да | 26 | 161,60 | 123 | 27,47 | 195,50 | 103 | 16,62 | 254,40 | 103 | 11,45 | 13,70 | 3 | 3,70 | 1 | Нет |
| NJ | 137 | 415 | 358-1921 | нет | нет | 0 | 243,40 | 114 | 41,38 | 121,20 | 110 | 10,30 | 162,60 | 104 | 7,32 | 12,20 | 5 | 3,29 | 0 | Нет |
| OH | 84 | 408 | 375-9999 | да | нет | 0 | 299,40 | 71 | 50,90 | 61,90 | 88 | 5,26 | 196,90 | 89 | 8,86 | 6,60 | 7 | 1,78 | 2 | Нет |
| OK | 75 | 415 | 330-6626 | да | нет | 0 | 166,70 | 113 | 28,34 | 148,30 | 122 | 12,61 | 186,90 | 121 | 8,41 | 10,10 | 3 | 2,73 | 3 | Нет |
| AL | 118 | 510 | 391-8027 | да | нет | 0 | 223,40 | 98 | 37,98 | 220,60 | 101 | 18,75 | 203,90 | 118 | 9,18 | 6,30 | 6 | 1,70 | 0 | Нет |

Рассмотрим поля представленной в таблице выборки:

* поле «Штат» выборки содержит название штата в котором проживает абонент;
* поле «Длина счета» показывает число номерных знаков в счете абонента;
* поле «Код» указывает на используемый префикс телефонного номера абонента;
* поле «№ телефона» содержит номер телефона абонента без префикса;
* поле «Международные звонки» показывает, пользовался ли абонент данной услугой;
* поле «Голосовая почта» говорит о том, что пользователь голосовой почты использовал или не использовал данную функцию;
* поле «Количество голосовых сообщений» показывает, какое количество голосовых сообщений было у данного пользователя;
* поле «Использовано дневных минут» указывает на то, сколько минут использовал абонент в дневное время.
* поле «Количество звонков днем» показывает, сколько звонков совершил абонент днем.
* поле «Абонентская плата 1» содержит информацию о том, сколько абонент должен заплатить за потраченные дневные минуты.
* поле «Использовано вечерних минут» указывает на то, сколько минут использовал абонент в вечернее время.
* поле «Количество звонков вечером» показывает, сколько звонков совершил абонент вечером.
* поле «Абонентская плата 2» содержит информацию о том, сколько абонент должен заплатить за потраченные вечерние минуты.
* поле «Использовано ночных минут» указывает на то, сколько минут использовал абонент в ночное время.
* поле «Количество ночных звонков» показывает, сколько звонков совершил абонент ночью.
* поле «Абонентская плата 3» содержит информацию о том, сколько абонент должен заплатить за потраченные ночные минуты.
* поле «Минут международных разговоров» указывает на то, сколько минут использовал абонент на международные разговоры.
* поле «Число международных звонков» показывает, сколько абонент совершил международных звонков.
* поле «Абонентская плата 4» содержит информацию о том, сколько абонент должен заплатить за потраченные на международные звонки минуты.
* поле «Число обращений в сервисную службу» содержит информацию о том, сколько абонент раз обращался в сервисную службу для устранения каких-либо неисправностей.
* поле «Уход» содержит информацию, о том, прекратил ли абонент пользоваться всеми видами услуг.

Поля «Штат», «Длина счета», «Код», «№ телефона», «Международные звонки», «Голосовая почта», «Количество голосовых сообщений», «Использовано дневных минут», «Количество звонков днем», «Абонентская плата 1», «Использовано вечерних минут», «Количество звонков вечером», «Абонентская плата 2», «Использовано ночных минут», «Количество ночных звонков», «Абонентская плата 3», «Минут международных разговоров», «Число международных звонков», «Абонентская плата 4», «Число обращений в сервисную службу» являются объясняющими (входными) переменными. Поле «Уход» клиента является бинарной зависимой (выходной) переменной. Поля «Голосовая почта» и «Международные звонки» - дихотомические входные переменные. Поля «Количество голосовых сообщений», «Количество звонков днем», «Количество звонков вечером», «Количество ночных звонков», «Число международных звонков», «Число обращений в сервисную службу» – полихотомические входные переменные.

### 

### 3.3.1 Разработка метода расчета параметров логистической регрессии на основе оценки шансов для дихотомической входной переменной

Автор дипломной работы анализирует влияние дихотомической входной переменной «*x*» «Голосовая почта» на дихотомическую выходную переменную «*y*» «Уход» клиента. Положительным исходом будет считаться уход клиента (*y* = 1). Отрицательным исходом будет случай, когда клиент продолжил пользоваться услугами компании (*y* = 0). Так как обе переменные дихотомические, можно подсчитать шанс ухода клиента, использующего и не использующего голосовую почту, по данным таблицы 4.

Следует найти коэффициенты логистической регрессии методом оценки шансов для дихотомической входной переменной «Голосовая почта» и определить вероятность отказа клиента от услуг компании в зависимости от того, пользовался ли он голосовой почтой.

Таблица 4

Сводная таблица ухода клиентов в зависимости от пользования голосовой почтой

|  |  |  |  |
| --- | --- | --- | --- |
|  | Голосовая почта = нет, *x*=0 | Голосовая почта = да, *x*=1 | Всего |
| *С*=нет, *y*=0 | 2008 | 842 | 2850 |
| *С*=да, *y*=1 | 403 | 80 | 483 |
| Всего | 2411 | 922 | 3333 |

В таблице 4 указано, какое количество клиентов, использующих и не использующих голосовую почту, ушло и осталось.

Переменная *C* в таблице 4 обозначает выходную переменную «Уход» и показывает, ушел ли клиент в другую компанию.

Из таблицы 4 видно, что из 2411 клиентов, не использующих голосовую почту, 2008 клиентов остались, а 403 ушли в другую компанию. Из 922 клиентов, использующих голосовую почту, 80 клиентов ушли, а 842 остались.

Рассмотрим порядок расчета коэффициентов логистической регрессии и вероятностей отказа клиента от услуг компании в зависимости от того, пользовался ли он голосовой почтой.

Шанс ухода клиента, использующего голосовую почту, определяется по формуле:

.

Для данной выборки равен: 80/842 = 0,095.

Шанс ухода клиента, не использующего голосовую почту, определяется по формуле:

.

Для данной выборки равен: 403/2008 = 0,201.

Тогда отношение шансов будет равно:

.

Вероятность *ρ*(*х*) ухода клиента, использующего (*х* = 1) или не использующего (*х* = 0) голосовую почту, определяется по формуле:

,

Где:

- переменная logit (см. пункт 3.1.1).

Расчет коэффициентов логистической регрессии *β0* и *β1* вызывает приведенные в пункте 3.2 затруднения. Их можно обойти для дихотомической входной переменной следующим образом.

Известно, что для бинарной логистической регрессии отношение шансов (*ОR*) равно :

.

Из этой формулы можно выразить коэффициент :

.

В данной выборке отношение шансов равно 0,473. Отсюда находится значение коэффициента :

.

Найденное значение коэффициента подставляется в уравнение . Тогда для клиентов, использующих голосовую почту (x = 1), вероятность ухода составит:

.

В полученном уравнении искомым является значение коэффициента .

Вероятность также может быть рассчитана по данным таблицы 4:

P(y = 1|x = 1) = 80/922 = 0,087.

Это означает, что:

.

Для дальнейших расчетов обозначается как *t*. Тогда необходимо решить следующее уравнение:

относительно.

Из полученной ранее формулы может быть выражено как:

.

При .

Отсюда .

Так как ; .

Вероятность того, что клиент, использовавший голосовую почту (*х* = 1), откажется от услуг компании, находится по формуле:

.

Она будет равна:

.

Вероятность того, что клиент, не использовавший голосовую почту (*х* = 0), откажется от услуг компании, находится по формуле:

.

Она будет равна:

.

Таким образом, автор дипломной работы нашел коэффициенты логистической регрессии методом оценки шансом для дихотомической входной переменной «Голосовая почта» и , равные (-1,602) и (-0,748) соответственно. Также он определил вероятность того, что клиент, использовавший голосовую почту, откажется от услуг компании, равную 0,087, и вероятность того, что клиент, не использовавший голосовую почту, откажется от услуг компании, равную 0,167.

Отметим, что значения коэффициентов логистической регрессии были получены аналитическим методом, не прибегая к численному итерационному решению.

### 3.3.2 Разработка метода расчета параметров логистической регрессии на основе оценки шансов для полихотомической (категориальной) входной переменной

В дипломной работе рассматривается влияние полихотомической входной переменной «Количество обращений в сервисную службу» на дихотомическую выходную переменную «Уход» клиента. Положительным исходом будет считаться уход клиента (*y* = 1). Отрицательным исходом будет случай, когда клиент продолжил пользоваться услугами компании (*y* = 0).

Полихотомическая входная переменная является дискретной. Значения полихотомической входной переменной можно разделить на интервалы и присвоить полученным интервалам категории: «низкое», «среднее» и «высокое». Назовем такую операцию квантованием.

Квантованию подлежит переменная «Количество обращений в сервисную службу», свидетельствующая о числе возникших у клиентов абонентской базы проблем. Предполагается, что чем чаще клиент обращался в сервисную службу, тем больше у него проблем, следовательно, у него больше причин отказаться от услуг компании, а значит вероятность ухода клиента больше.

Данные о текучести абонентской базы приведены в таблице 3. Эти данные необходимо проанализировать на наличие связи между числом обращений клиента в сервисную службу и вероятностью ухода клиента из компании.

Для последующих расчетов необходимо ввести понятие «фиктивная (вспомогательная) переменная (CSC)». Фиктивная переменная указывает на количество обращений клиентов в абонентскую сервисную службу. В зависимости от числа обращений клиентов в сервисную службу фиктивной переменной экспертом предметной области присваивается соответствующая категория.

Для данной выборки, как видно из таблицы 5, числу обращений клиентов в сервисную службу, равному 1, либо отсутствию обращений, присваивается категория «низкое». Этой категории соответствуют значения фиктивных входных переменных , . Числу обращений, равному 2 и 3, присваивается категория «среднее». Этой категории соответствуют следующие значения фиктивных переменных: , . Категории «высокое» соответствует диапазон обращений клиентов в сервисную службу от 4 и более. Значения фиктивных переменных этой категории клиентов будут равны 1.

Следует найти коэффициенты логистической регрессии методом оценки шансов для полихотомической входной переменной «Количество обращений в сервисную службу» и определить вероятности отказа различных категорий клиентов от услуг компании в зависимости от того, насколько часто они обращались в сервисную службу.

Таблица 5

Квантование переменной CSC для числа обращений в сервисную службу

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Низкое (0 или 1 вызов), CSC = Низкое | 0 | 0 |
| Среднее (2 или 3 вызова), CSC = Среднее | 1 | 0 |
| Высокое ( вызовов), CSC = Высокое | 1 | 1 |

Таблица 6

Сводная таблица ухода в зависимости от числа обращений в сервисную службу для переменной CSC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CSC=Низкое | CSC=Среднее | CSC=Высокое | Всего |
| *C*=Нет, *y*=0 | 1664 | 1057 | 129 | 2850 |
| *C*=Да, *y*=1 | 214 | 131 | 138 | 483 |
| Всего | 1878 | 1188 | 267 | 3333 |

В таблице 6 приведены данные о том, какое количество клиентов каждой категории ушли и остались, а также указано общее количество клиентов, отказавшихся или не отказавшихся от услуг компании. Переменная *C* обозначает выходную переменную «Уход» и показывает, ушел ли клиент в другую компанию или остался.

Из данных таблицы 6 видно, что из 1878 клиентов, входящих в категорию «низкое», 1664 клиента остались, а 214 обратились к услугам другой компании. Из 1188 клиентов, входящих в категорию «среднее», 1057 абонентов продолжили пользоваться услугами компании, а 131 абонент ушел. Сто двадцать девять клиентов из 267, относящихся к категории «высокое», остались, а 138 обратились к услугам другой компании. Также видно, что всего абонентов, в базе данных компании 3333, из которых 2850 продолжили пользоваться ее услугами, а 483 ушли.

Рассмотрим порядок расчета коэффициентов логистической регрессии и вероятности отказа клиента от услуг компании в зависимости от того, насколько часто он обращался в сервисную службу.

В данной выборке для категориальной (полихотомической) входной переменной можно обойтись без расчета коэффициентов функции логит-преобразования , и подобно тому, как это было сделано ранее для дихотомической входной переменной. Для этого необходимо по данным таблицы 6 вначале вычислить отношения шансов ухода клиентов для категорий «среднее» и «высокое» по отношению к категории «низкое».

Шанс ухода клиента для категории «низкое» определяется по формуле:

.

Для данной выборки равен: 214/1664=0,129.

Шанс ухода клиента для категории «среднее» определяется по формуле:

.

Для данной выборки равен: 131/1057=0,124.

Шанс ухода клиента для категории «высокое» определяется по формуле:

.

Для данной выборки равен: 138/129=8,821.

Отношение шансов для категории «среднее»:

.

Отношение шансов для категории «высокое»:

.

Вероятность *ρ*(*х*) ухода клиента для категории «низкое» рассчитывается по формуле:

,

Где:

,

Где:

и - значения фиктивной переменной, которые содержатся в таблице 5.

Функция после подстановки в нее значений фиктивных переменных для категории «низкое» и примет вид:

.

Тогда *ρ*(*х*) равняется:

.

Так же известно, что вероятность можно найти, поделив число положительных исходов на общее количество исходов, 214/1878=0,114. То есть:

.

Отсюда можно найти коэффициент по следующей формуле:

.

После подстановки значений, в данном примере будет равен: (-2,05).

Для категории «среднее» , а , тогда :

.

находится по формуле:

.

После подстановки значений, в данном примере будет равен: (-0,04).

Для категории «высокое» , тогда :

.

можно найти по формуле:

.

После подстановки значений, в данном примере будет равен: 2,14.

Таким образом, в дипломной работе методом оценки шансом для полихотомической входной переменной «Количество обращений в сервисную службу» найдены коэффициенты логистической регрессии , и , равные (-2,05), (-0,04) и 2,14 соответственно. Также определена вероятность того, что клиенты, количество обращений в сервисную службу которых ноль или единица, т.е. принадлежащие к категории «низкое», откажутся от услуг компании с вероятностью 0,114. Вероятность того, что клиенты из категории «среднее» откажутся от услуг компании составляет 0,110, а вероятность ухода клиентов, принадлежащих к категории «высокое», равна 0,512.

# Глава 4 Разработка и реализация алгоритма для расчета коэффициентов логистической регрессии и ROC-кривых

## 4.1 Разработка алгоритма построения ROC-кривых

ROС-анализ представляет собой графическую методику оценивания эффективности моделей с помощью двух показателей — чувствительности (*Se*) и специфичности (*Sp*).

Чувствительность (sensitivity, *Se*) определяется как отношение числа истинноположительных наблюдений к числу фактически положительных наблюдений:

.

Специфичность (specificity, *Sp*) определяется как отношение числа истинноотрицательных наблюдений к числу фактически отрицательных наблюдений:

.

Если отсутствуют ложноотрицательные исходы, чувствительность равна 1; если отсутствуют ложноположительные исходы, специфичность равна 1. Модель, которая способна идеально точно классифицировать как положительные, так и отрицательные примеры, будет иметь 100% чувствительность и специфичность. Чтобы минимизировать ошибки I рода, нужно использовать модель с высокой чувствительностью. Чтобы минимизировать ошибки II рода, необходимо использовать модель с высокой специфичностью.

В задачах бинарной классификации, когда модель предсказывает вероятность того, что наблюдение относится к одному из двух классов, очень важен выбор точки отсечения, то есть порога вероятности, разделяющего два класса. Такая точка отсечения показывает, после какого значения вероятности на выходе модели один класс сменяется другим.

В общем случае задача ROC-анализа заключается в том, чтобы выбрать такую точку отсечения, которая обеспечит максимум чувствительности и специфичности. Но на практике это не всегда возможно, поскольку модели, дающие одинаковое количество ложноположительных и ложноотрицательных исходов встречаются очень редко. Чаще всего встречаются модели, которые выдают большее количество ошибок первого или второго рода.

Цель ROС-анализа заключается в том, чтобы подобрать такое значение точки отсечения, которое позволит модели с наибольшей точностью распознавать положительные или отрицательные исходы и выдавать наименьшее количество ложноположительных или ложноотрицательных ошибок.

ROС-анализ широко используется в медицине, биологии, маркетинге, банковском деле и других областях, где применяется бинарная вероятностная классификация.

В основе ROC-анализа лежат ROC-кривые. График ROC-кривой изображен на рис. 6.

Для построения ROC-кривой требуется изменять порог отсечения в диапазоне от 0 до 1 с заданным шагом, например 0,01. Чем меньше шаг, тем больше будет точек на графике, что косвенно влияет на точность построения ROC-кривой. При каждом значении порога заново рассчитываются значения специфичности (*Sp*) и чувствительности (*Se*), т.е. меняется количество распознанных ошибок I и II рода. Чувствительность (*Se*) откладывается по оси ординат, а по оси абсцисс откладывается значение 100-*Sp*.

0

100

100

50

50

Se, %

100-Sp, %

Рисунок 6. ROC-кривая и идеальная ROC-кривая

Рассмотрим процесс построения ROC-кривой более детально.

Зададим значение порога равное, например, 0,2. Вероятности для каждого примера выборки сравниваются с пороговым значением. В зависимости от исхода сравнения вероятности с порогом, рассчитываются значения *TP*, *TN*, *FN*, *FP*, следующим образом.

Если вероятность больше порога и пример положительный, то событие будет распознано, как истинноположительное (*TP*). В случае, если вероятность меньше порога и пример отрицательный, событие будет распознано, как истинноотрицательное (*TN*). Если вероятность больше порога и пример отрицательный, событие будет распознано, как ложноположительное (*FP*). В случае, если вероятность меньше порога и пример положительный, событие будет распознано, как ложноотрицательное (*FN*).

По полученным значениям *TP*, *TN*, *FN*, *FP* рассчитываются специфичность (*Sp*) и чувствительность (*Se*) выборки. На графике откладывается точка с координатами: по оси ординат - *Se*\*100%, по оси абсцисс - (100-*Sp*)\*100%.

При следующем значении порога, равному 0,21 (шаг увеличения порога 0,01), снова будут рассчитаны значения *TP, TN, FN, FP*, по которым на графике будет отложена следующая точка. Так, меняя порог от 0 до 1 с шагом 0,01, строится ROC-кривая.

Еще одно полезное свойство ROC-кривой заключается в том, что она позволяет оценить качество модели бинарной классификации. Определить степень качества модели можно по кривизне кривой: чем ближе она к идеальному классификатору, тем качественнее модель. Если же кривая близка к диагонали, то модель бесполезна. На рис. 6. пунктирной линией обозначен идеальный классификатор, а диагональной линией – бесполезный классификатор.

Так же стоит отметить, что на рис. 6 изображена сглаженная кривая. На практике, она имеет ступенчато-изрезанный вид, зависящий от однородности и репрезентативности выборки.

Визуально сравнить ROC-кривые и выявить более эффективную модель не всегда представляется возможным. Поэтому их можно сравнивать с помощью подсчета площади под кривой (AUC).

Площадь под кривой характеризует прогностическую силу модели, при этом AUC=1 соответствует идеальному классификатору, который не достижим на практике, а AUC=0,5 соответствует бесполезному классификатору.

Площадь под кривой может быть рассчитана, например, методом трапеций:

,

Где:

*n* - количество точек;

*x* – координата точки по оси абсцисс;

*y* – координата точки по оси ординат.

О том, насколько хорошей прогностической силой обладает модель можно судить исходя из таблицы 7. В ней представлена примерная экспертная шкала оценки качества модели в зависимости от площади под ROC-кривой.

Таблица 7

Качество модели в зависимости от площади под кривой

|  |  |
| --- | --- |
| Интервал AUC | Качество модели |
| 0,9-1 | Отличное |
| 0,8-0,9 | Очень хорошее |
| 0,7-0,8 | Хорошее |
| 0,6-0,7 | Среднее |
| 0,5-0,6 | Неудовлетворительное |

[9]

Идеальная модель имеет 100% чувствительность и специфичность. Но добиться этого на практике невозможно, т.к. одновременно повысить чувствительность и специфичность модели нельзя. Сто процентные чувствительность и специфичность означают, что все примеры как положительные, так и отрицательные распознаны, верно. Нужно находить компромисс с помощью порога отсечения, т.к. порог влияет на соотношение чувствительности и специфичности.

## 4.2 Реализация разработанного метода нахождения коэффициентов логистической регрессии и алгоритма построения ROC-кривых

### 4.2.1 Исходные данные программы

Исходными данными для построения модели логистической регрессии и нахождения ее коэффициентов, а так же построения ROC-кривых, послужила выборка «текучесть абонентской базы», взятая из CD приложения к учебному пособию Паклина Н.Б. [5] Фрагмент данной выборки приведен в таблице 3 третьей главы данной дипломной работы.

Сначала выше упомянутая выборка была проанализирована на наличие зависимости между входными переменными «Международные звонки», «Голосовая почта», «Количество голосовых сообщений», «Количество звонков днем», «Количество звонков вечером», «Количество ночных звонков», «Число международных звонков», «Число обращений в сервисную службу», и выходной переменной «Уход». Входные переменные «Международные звонки», «Голосовая почта» являются бинарными. Входные переменные «Количество голосовых сообщений», «Количество звонков днем», «Количество звонков вечером», «Количество ночных звонков», «Число международных звонков», «Число обращений в сервисную службу» являются дискретными т.е. могут быть преобразованы в категориальный вид. Выходная переменная «Уход» (*y*) является бинарной, т.е. может принимать значения 0 или 1.

В данном случае положительным исходом будет считаться уход клиента (*y* = 1). Отрицательным исходом будет случай, когда клиент продолжил пользоваться услугами компании (*y* = 0).

В результате анализа было выявлено, что между выше упомянутыми переменными имеется зависимость, которая может быть описана с помощью логистической регрессии. В качестве модели была выбрана логистическая регрессия потому, что она дает наиболее точную оценку отнесения примера к положительному (клиент ушел из компании) или отрицательному исходу (клиент продолжил пользоваться услугами компании).

### 4.2.2 Реализация алгоритма нахождения коэффициентов логистической регрессии для дихотомической входной переменной

Рассмотрим алгоритм нахождения параметров логистической регрессии для входных дихотомических переменных. В общем виде он изображен на рис. 7.

Фрагмент исходного кода программы, рассчитывающий коэффициенты логистической регрессии и для дихотомической входной переменной «Голосовая почта» с помощью метода оценки шансов, приведен ниже на рис. 8.

Для реализации этой задачи алгоритм рассчитывает отношения шансов для переменной «Голосовая почта» и данные для последующих расчетов. Так же алгоритм рассчитывает вероятности положительного исхода выходной переменной, т.е. уход клиента в зависимости от того, пользовался ли он голосовой почтой. Опишем его.

В 1 и 2 строках алгоритма активируется подключенная ранее база данных с исследуемой выборкой «текучесть абонентской базы», а так же устанавливается указатель на первую запись.

С 3 по 14 строчку описан цикл, который перебирает всех клиентов выборки и подсчитывает: сколько клиентов не используют голосовую почту и не отказались от услуг компании (в алгоритме обозначается как *ff*), сколько клиентов не пользовались голосовой почтой и отказались от услуг компании (*ft*); сколько клиентов пользовались голосовой почтой и не отказались от услуг компании (*tf*), количество клиентов, которые пользовались голосовой почтой и отказались от услуг компании (*tt*).



Рисунок 7. Блок-схема расчета параметров логистической регрессии для входных дихотомических переменных

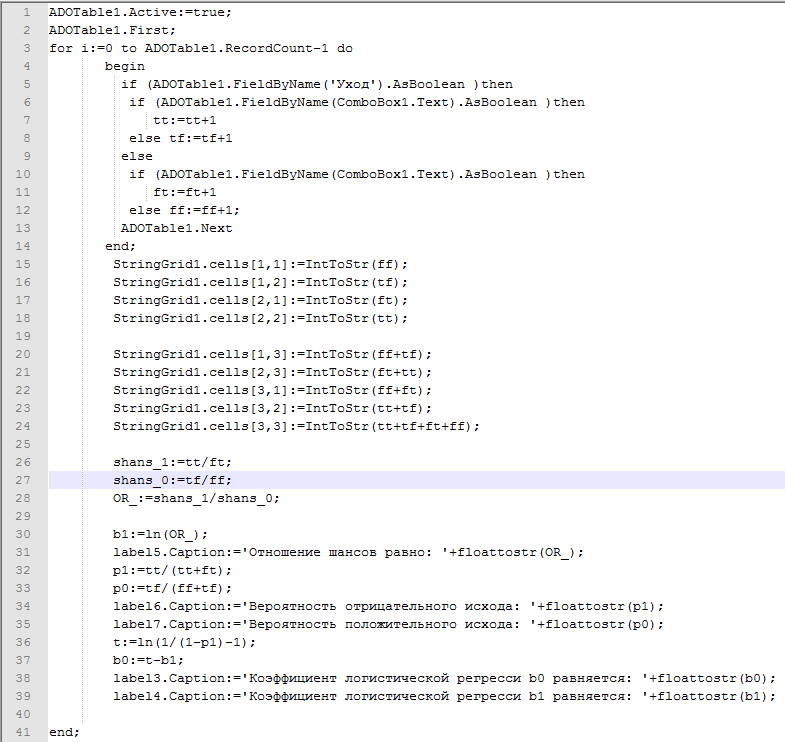


Рисунок 8. Фрагмент исходного кода «Алгоритм нахождения параметров логистической регрессии для входных дихотомических переменных»

В строчках 15-18 ранее подсчитанные значения записываются в таблицу, аналогичную таблице 8.

Таблица 8

Уход клиентов в зависимости от пользования голосовой почтой

|  |  |  |  |
| --- | --- | --- | --- |
|  | Голосовая почта = 0 | Голосовая почта = 1 | Всего |
| Уход=0 | 2008 | 842 | 2850 |
| Уход=1 | 403 | 80 | 483 |
| Всего | 2411 | 922 | 3333 |

В таблице 8 указано, какое количество клиентов, использующих и не использующих голосовую почту, ушло и осталось. В столбце «Голосовая почта=0» посчитаны клиенты, которые не пользовались голосовой почтой, ушедшие (Уход =1) или продолжившие пользоваться услугами компании (Уход = 0). В столбце «Голосовая почта=1» приведено количество клиентов, которые пользовались голосовой почтой, и ушли или не ушли из компании.

По данным таблицы видно, что из 2411 клиентов, не использующих голосовую почту, 2008 клиентов остались, а 403 ушли в другую компанию. Из 922 клиентов, использующих голосовую почту, 80 клиентов ушли, а 842 остались.

Значения для строчки и столбца «Всего», рассчитываются путем суммирования полученных ранее значений, а потом записываются в таблицу в строчках алгоритма 20-24.

В строках 26 и 27 алгоритма рассчитываются шансы ухода клиента, пользовавшегося или не пользовавшегося голосовой почтой, по следующим формулам.

,

Где:

– шанс ухода клиента, использовавшего голосовую почту;

*tt* - количество клиентов, которые пользовались голосовой почтой и отказались от услуг компании;

*ft* - количество клиентов, которые не пользовались голосовой почтой и отказались от услуг компании.

,

Где:

– шанс ухода клиента, не использовавшего голосовую почту;

*ff* - количество клиентов, которые не пользовались голосовой почтой и не отказались от услуг компании;

*tf* - количество клиентов, которые пользовались голосовой почтой и не отказались от услуг компании.

В строке 28 рассчитываются отношения шансов по формуле:

.

В строках 32 и 33 рассчитываются вероятности положительного исхода выходной переменной для значений входной дихотомической переменной по следующим формулам.

,

Где:

– вероятность того, что клиент, пользовавшийся голосовой почтой, уйдет.

,

Где:

– вероятность того, что клиент, не использовавший голосовую почту, уйдет.

В строках 30 и 37 рассчитываются коэффициенты логистической регрессии по следующим формулам.

.

,

Где:

.

В строках 31, 34, 35, 38 и 39 выводятся полученные параметры логистической регрессии на экран.

Полный исходный код программы представлен в приложении 1.

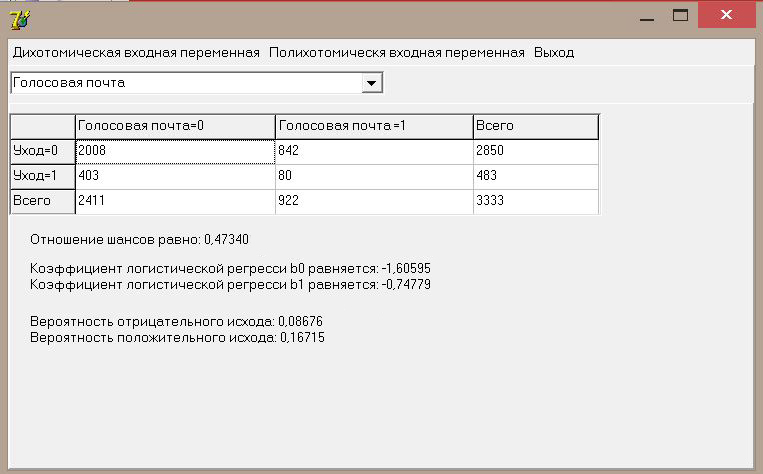


Рисунок 9. Окно расчета параметров логистической регрессии для входной дихотомической переменной

Окно программы с результатами расчета, на примере входной дихотомической переменной «Голосовая почта», по представленному выше алгоритму, приведено на рис. 9.

Как видно из рис. 9, программа выдает отношение шансов, равное 0,47, что говорит о том, что шанс благоприятного исхода больше, чем шанс не благоприятного. Применительно к данной выборке это означает, что клиенты скорее откажутся от услуг компании, чем останутся в рядах ее клиентов.

Также программа выдает значения вероятностей для положительного и отрицательного исходов, равные 16,715 и 8,676% соответственно. Это значит, что клиенты, не использовавшие голосовую почту, уйдут из компании с 16,715% вероятностью. В то же время клиенты, пользовавшиеся голосовой почтой, покинут ряды клиентов компании с меньшей вероятностью, равной 8,676%, что логично и соответствует действительности.

Из выше сказанного можно сделать вывод о том, что программа работает правильно и выдает адекватные значения искомых величин.

### 4.2.3 Реализация алгоритма нахождения коэффициентов логистической регрессии для полихотомической входной переменной

Рассмотрим алгоритм нахождения параметров логистической регрессии для входных полихотомических переменных. В общем виде он изображен на рис. 10.

Фрагмент исходного кода программы, рассчитывающий коэффициенты логистической регрессии , и для полихотомической входной переменной «Число международных звонков» с помощью метода оценки шансов, приведен ниже на рис. 11. Для реализации этой задачи алгоритм разбивает входную переменную на категории «низкое», «среднее» и «высокое» по числу международных звонков, рассчитывает отношения шансов для категорий «среднее», «высокое» по отношению к категории «низкое» и данные для последующих расчетов. Так же алгоритм рассчитывает вероятности положительного исхода выходной переменной, т.е. уход клиента для категорий «низкое», «среднее», «высокое» переменной «Число международных звонков». Опишем его.

В 1 и 2 строках алгоритма активируется подключенная ранее база данных с исследуемой выборкой «текучесть абонентской базы», а так же устанавливается указатель на первую запись.

С 3 по 10 строчку описан цикл, который перебирает всех клиентов выборки и находит максимальное и минимальное значение переменной «Число международных звонков», для того, чтобы впоследствии, в строках с 11 по12, используя найденные максимум и минимум переменной «Число международных звонков», вычислить и записать в таблицу, аналогичную таблице 9, значения диапазонов для категорий клиентов «низкое», «среднее», «высокое».



Рисунок 10. Блок-схема расчета параметров логистической регрессии для входных полихотомических переменных

Таблица 9

Разбиение на категории входной переменной «Число международных звонков»

|  |  |  |  |
| --- | --- | --- | --- |
|  | Диапазон |  |  |
| Низкое | [0;7] | 0 | 0 |
| Среднее | [8;12] | 1 | 0 |
| Высокое | [13;20] | 1 | 1 |

В таблице 9 представлены результаты разделения клиентов на категории «низкое», «среднее», «высокое» в зависимости от того,  
насколько активно они пользовались функцией международные звонки.

Из данных таблицы 9 видно, что числу международных звонков, клиентов в диапазоне от 0 до 7, присваивается категория «низкое». Этой категории соответствуют значения фиктивных входных переменных , .

Числу международных звонков в диапазоне от 8 до 12 присваивается категория «среднее». Этой категории соответствуют следующие значения фиктивных переменных: , .

Категории «высокое» соответствует диапазон обращений клиентов в сервисную службу от 13 до 20. Значения фиктивных переменных этой категории будут равны единице.

В 15 и 16 строках алгоритма заново устанавливается указатель на первую запись.



Рисунок 11. Фрагмент исходного кода «Алгоритм нахождения параметров логистической регрессии для входных полихотомических переменных»

С 17 по 36 строчки описан цикл, который перебирает всех клиентов выборки и подсчитывает следующие значения: сколько клиентов из категории «низкое» не отказались от услуг компании (в алгоритме обозначается как *fb*), сколько клиентов из категории «низкое» отказались от услуг компании (*tb*); сколько клиентов из категории «среднее» не отказались от услуг компании (*fs*), сколько клиентов из категории «среднее» отказались от услуг компании (*ts*); сколько клиентов из категории «высокое» не отказались от услуг компании (*ft*), сколько клиентов из категории «высокое» отказались от услуг компании (*tt*). В строчках с 37 по 42 эти значения записываются в таблицу, аналогичную таблице 10.

Таблица 10

Уход клиента в зависимости от числа совершенных международных звонков для разных категорий клиентов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Число международных звонков=Низкое | Число международных звонков =Среднее | Число международных звонков =Высокое | Всего |
| Уход=Нет | 2541 | 280 | 29 | 2850 |
| Уход=Да | 439 | 38 | 6 | 483 |
| Всего | 2980 | 318 | 35 | 3333 |

В таблице 10 приведено, какое количество клиентов каждой категории ушло и осталось, а также указано общее количество клиентов, отказавшихся или не отказавшихся от услуг компании.

В столбце «Число международных звонков=Низкое» посчитаны клиенты, у которых число международных звонков входит в диапазон для категории «низкое», ушедшие (Уход =1) или продолжившие пользоваться услугами компании (Уход = 0). В столбце «Число международных звонков=Среднее» приведено количество клиентов, у которых число международных звонков входит в диапазон для категории «среднее», ушедшие (Уход = 1) или продолжившие пользоваться услугами компании (Уход = 0). В столбце «Число международных звонков=Высокое» приведено количество клиентов, у которых число международных звонков входит в диапазон для категории «высокое», ушедшие (Уход = 1) или продолжившие пользоваться услугами компании (Уход = 0).

Из данных таблицы 10 видно, что из 1878 клиентов, входящих в категорию «низкое», 1664 клиента остались, а 214 обратились к услугам другой компании. Из 1188 клиентов, входящих в категорию «среднее», 1057 абонентов продолжили пользоваться услугами компании, а 131 абонент ушел. Сто двадцать девять клиентов из 267, относящихся к категории «высокое», остались, а 138 обратились к услугам другой компании. Также видно, что всего абонентов, в базе данных компании 3333, из которых 2850 продолжили пользоваться ее услугами, а 483 ушли.

Значения для строчки и столбца «Всего», рассчитываются путем суммирования полученных ранее значений, а потом записываются в таблицу в строках алгоритма 43-48.

В строках 49 и 50 идет расчет отношения шансов для категорий «среднее» и «высокое» соответственно, по отношению к категории «низкое» по следующим формулам.

,

Где:

– отношение шансов для категории «среднее», по отношению к категории «низкое»;

*ts –* число клиентов из категории «среднее», которые отказались от услуг компании;

*fs –* числоклиентов из категории «среднее», которые не отказались от услуг компании;

*tb –* число клиентов из категории «низкое», которые отказались от услуг компании;

*fb –* число клиентов из категории «низкое», которые не отказались от услуг компании.

,

Где:

– отношение шансов для категории «высокое», по отношению к категории «низкое»;

*tt* – число клиентов из категории «высокое», которые отказались от услуг компании;

*ft* – число клиентов из категории «высокое», которые не отказались от услуг компании.

В строчках 53, 57 и 61 рассчитываются вероятности положительного исхода выходной переменной в зависимости от категорий клиентов «низкое», «среднее» и «высокое» входной переменной по следующим формулам.

,

Где:

– вероятность того, что клиенты из категории «низкое» перестанут пользоваться услугами компании.

,

Где:

– вероятность того, что клиенты из категории «среднее» перестанут пользоваться услугами компании.

,

Где:

– вероятность того, что клиенты из категории «высокое» перестанут пользоваться услугами компании.

Коэффициенты логистической регрессии , и рассчитываются в строках 55, 59 и 63 по следующим формулам.

,

,

.

Вывод всех выше описанных и подсчитанных значений осуществляется в строчках алгоритма 50, 52, 54, 56, 58, 60, 62 и 64.

Полный исходный код программы представлен в приложении 1.

Окно приложения с результатами расчетов на примере входной полихотомической переменной «Число международных звонков» представлено на рис. 12, ниже.

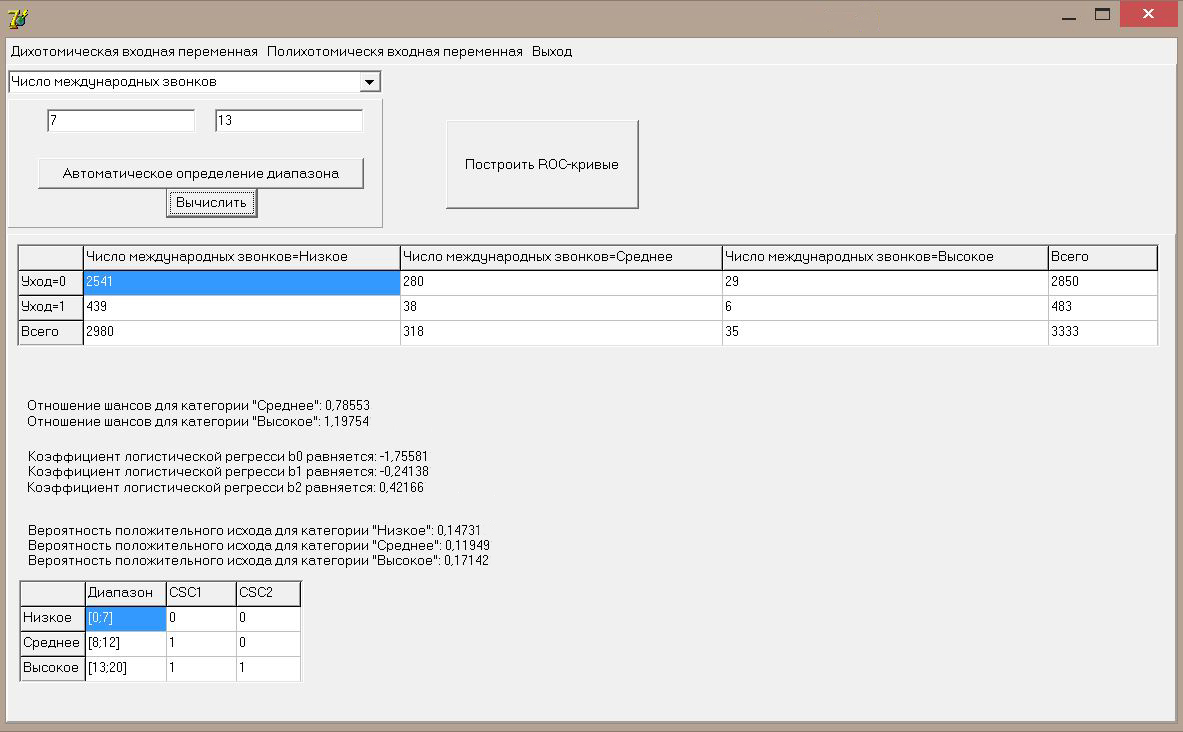


Рисунок 12. Окно расчета параметров логистической регрессии для полихотомической входной переменной

Как видно из рис. 12, программа выдает отношение шансов, равное 0,785, для клиентов, отнесенных по переменной «Число международных звонков» к категории «среднее», т.е. шанс благоприятного исхода больше, чем шанс не благоприятного. Применительно к данной выборке это означает, что клиент, пользовавшийся функцией международные звонки от 8 до 12 раз, скорее всего, останется клиентом компании. Отношение шансов для категории «высокое», равное 1,197, говорит о том, что клиенты данной категории скорее всего уйдут из компании.

Также программа выдает значения вероятностей для каждой категории клиентов. Вероятность того, что клиенты, относящиеся к категорий «низкое», «среднее», «высокое», перестанут пользоваться услугами компании составляет 14,731%, 11,949% и 17,142% соответственно. Это говорит о том, что с наименьшей вероятностью уйдут клиенты, редко пользующиеся услугой международных звонков, т.е. чем больше клиент пользуется функцией международных звонков, тем скорее он уйдет. По мнению автора дипломной работы, данная тенденция связана с неудовлетворительным качеством предоставляемой услуги.

### 4.2.4 Реализация алгоритма построения ROC-кривых

Рассмотрим алгоритм построения ROC-кривых для входных полихотомических переменных.

Блок схема данного алгоритма изображена на рис. 13.

Фрагмент исходного кода для построения ROC-кривой и расчета площади под ней, приведен на рис. 14.

Описанный в пункте 4.2.3 алгоритм разбивает входную переменную «Число международных звонков» на категории «низкое», «среднее» и «высокое» по числу международных звонков клиента.

В данном алгоритме ROC-кривые строятся для категорий клиентов «низкое», «среднее» и «высокое». Рассмотрим алгоритм построения ROC-кривых на примере категории «низкое».



Рисунок 13. Блок-схема построения ROC-кривой

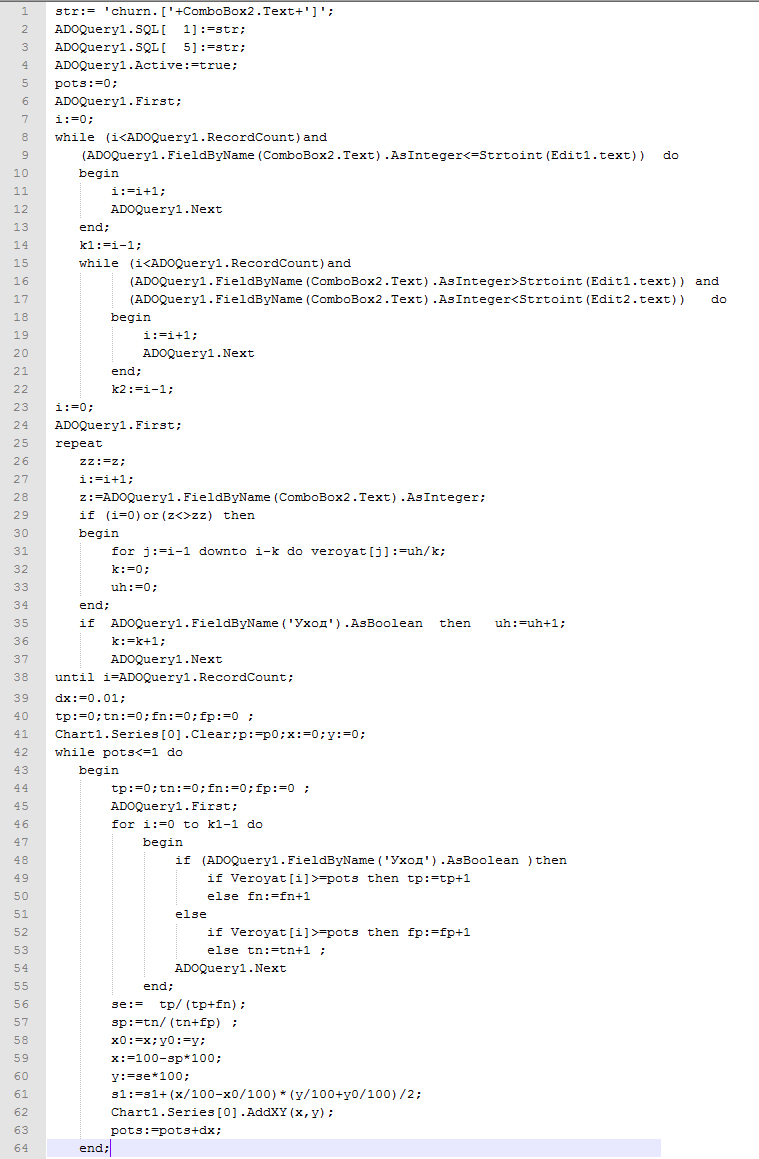


Рисунок 14. Фрагмент исходного кода. «Алгоритм построения ROC-кривых для категорий «низкое», «среднее» и «высокое», на примере категории «низкое»

В первых семи строках алгоритма производится необходимая подготовка к дальнейшей работе, которая включает в себя следующие действия.

* Сортировка входной переменной по возрастанию с «привязанной» к ней выходной переменной (реализовано в 1-3 строчке алгоритма).

Применительно к рассматриваемой переменной сортировка производится по возрастанию числа международных звонков с привязанной к ней переменной «Уход» клиента.

* Активация базы данных (4 строчка алгоритма).
* Обнуление счетчиков (5 и 7 строчки алгоритма).
* Установка указателя на первую запись базы данных (6 строчка алгоритма).

С 8 по 22 строчку алгоритма производится расчет двух чисел, характеризующих конец диапазона категории «низкое» и начало диапазона категории «высокое». С помощью этих чисел, в последствии, можно получить полные диапазоны для каждой категории клиентов.

В строке 23 обнуляется счетчик, нужный для последующего расчета. А в строке 24 идет установка указателя на первую запись базы данных.

С 25 по 38 строку кода описан алгоритм расчета вероятностей для каждого примера выборки с последующей записью в массив, т.е. вероятность ухода определяется для каждого клиента. Поскольку входная переменная «Число международных звонков» дискретна и клиентов в выборке много, становится возможным расчет вероятности для каждого клиента путем деления количества одинаковых значений переменной «число международных звонков» на число ушедших клиентов из этих примеров. Этот массив понадобится непосредственно для построения ROC-кривых.

С 39 по 41 строчку алгоритма производится необходимая подготовка к дальнейшей работе, которая включает в себя следующие действия.

* Задается шаг порога отсечения (реализовано в 39 строчке алгоритма).
* Обнуляются счетчики истиноположительных, истиноотрицательных, ложноотрицательных и ложноположительных примеров (40 строчка алгоритма).
* Очищается будущий график от возможного мусора (41 строчка алгоритма).
* Обнуляются другие счетчики (41 строчка алгоритма).

С 42 по 64 строки описан непосредственно алгоритм построения ROC-кривой для категории «низкое». Рассмотрим его подробней.

На 42 строчке начинается внешний цикл, условием выполнения которого является то, что порог отсечения должен быть меньше или равен единице. Сам порог меняется в этом цикле на 63 строке, с шагом в 0,01, установленным на 39 строке. На 44 и 45 строках идет обнуление всех счетчиков *tp*, *tn*, *fn*, *fp* и установка указателя базы данных на начало, что будет выполняться каждую итерацию цикла.

Далее, начиная с 46 по 55 строки, выполняется внутренний цикл, границы выполнения которого соответствуют диапазону категории «низкое». В цикле происходит распознавание и подсчет истиноположительных (*tp*), истиноотрицательных (*tn*), ложноотрицательных (*fn*) и ложноположительных (*fp*) примеров в зависимости от значения порога отсечения.

Каждой итерации внешнего цикла соответствует полный проход внутреннего цикла. Поскольку порог отсечения меняется во внешнем цикле, при каждом проходе внутреннего цикла меняется соотношение распознанных примеров.

В 56 и 57 строках идет расчет специфичности (Sp) и чувствительности (Se), а в строках 59 и 60 - вычисление координат точки. В строке 62 происходит добавление точки на график. Т.е. после каждого прохода внутреннего цикла строится точка графика, а количество точек зависит от величины шага точки отсечения.

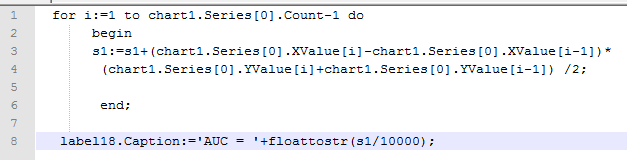


Рисунок 15. Фрагмент кода, вычисляющий площадь под ROC-кривой для категории «низкое»

На рис. 15 представлен фрагмент исходного кода, считающий площадь под кривой (AUC). Рассмотрим его.

В строках 1-6 рассчитывается площадь под кривой численным методом трапеций, по формуле:

,

Где:

*n* - количество точек;

*x* – координата точки по оси абсцисс;

*y* – координата точки по оси ординат.

Полученная площадь выводится под графиком с помощью строки 8 алгоритма.

Полный исходный код программы представлен в приложении 1.

Окно приложения с построенными ROC-кривыми для категорий «низкое», «среднее», «высокое» на примере входной полихотомической переменной «Число международных звонков» представлено на рис. 10, ниже.

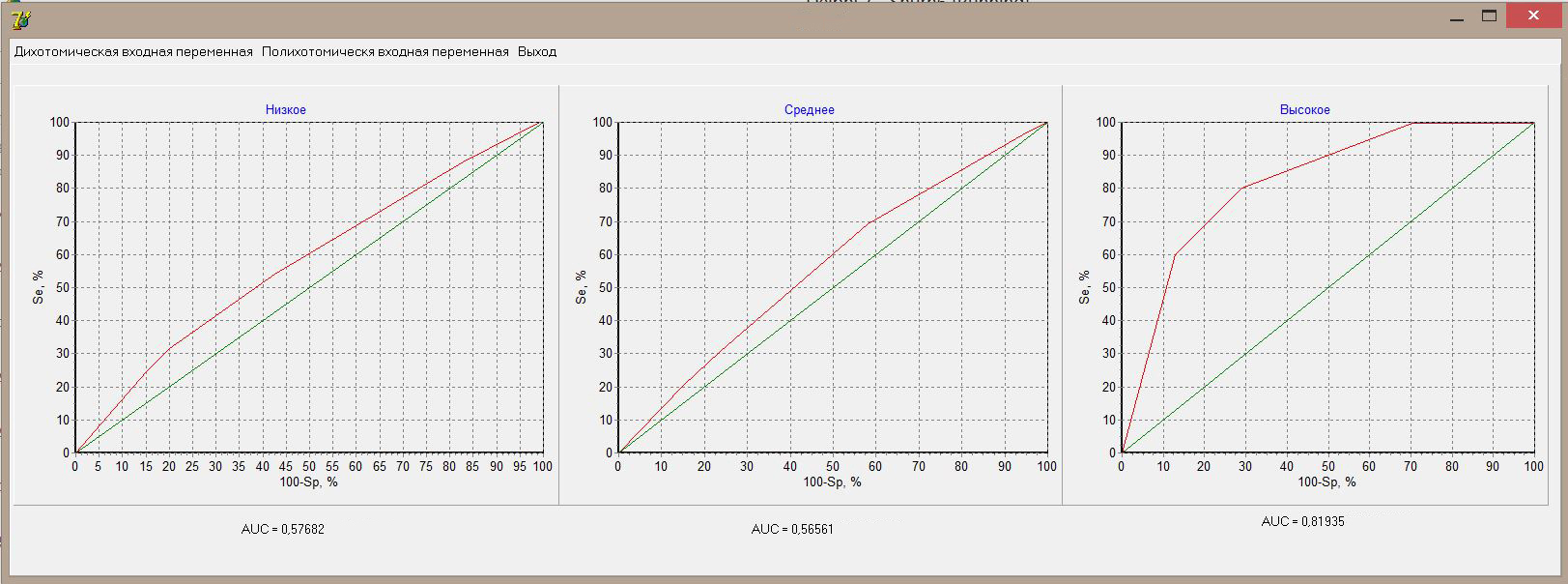


Рисунок 16. Окно расчета параметров логистической регрессии для полихотомической входной переменной

Как видно из рис. 16, программа строит ROC-кривые для трех категорий клиентов «низкое», «среднее», «высокое» по входной переменной «Число международных звонков» и выводит значения площадей под кривыми (AUC) для каждой кривой. Наиболее близка к идеальному классификатору ROC-кривая для категории «высокое», что говорит о том, что модель зависимости между входной переменной «Число международных звонков» и выходной переменной «Уход» качественна, точна и выдает небольшое количество ошибок I и II рода.

На это также указывает значение AUC для этой кривой, равное 81,935%. Согласно классификатору моделей в зависимости от величины AUC, данная модель может быть идентифицирована как «очень хорошая».

Остальные две модели для категорий клиентов «низкое» и «среднее» почти бесполезны, т.к. ROC-кривые для них близки к диагональной линии и площади под кривыми, согласно классификатору, соответствуют «неудовлетворительному» качеству моделей. Это означает, что прямой зависимости между количеством международных звонков и уходом клиентов в категориях «низкое» и «среднее» нет.

# 

# Заключение

Данная дипломная работа представляет собой разработку и реализацию методов оценки правдоподобия бинарной классификации на основе оценки шансов.

В качестве базы для построения модели, описывающей выборку «текучесть абонентской базы», взята логистическая регрессия. В ходе исследования существующих методов для нахождения коэффициентов логистической регрессии, был сделан вывод о том, что они могут быть найдены с помощью метода максимального правдоподобия только вычислительным итерационным путем, поэтому в дипломной работе был разработан метод оценки шансов, который позволяет аналитически найти данные коэффициенты логистической регрессии для дихотомической и полихотомической входной переменной.

Разработанный в дипломной работе метод был реализован в программном обеспечении на данных выборки «текучесть абонентской базы». С помощью этого метода была построена модель логистической регрессии, качество которой проверяется с помощью построения ROC-кривых. В результате анализа эффективности построенной модели было выявлено, что она работает достоверно и точно отражает существующие зависимости, дает адекватную оценку изучаемой проблемы и позволяет выработать соответствующие рекомендации по принятию решения.

Программное обеспечение может быть использовано аналитиками для решения задач бинарной классификации, построения соответствующих моделей и оценки их качества. Оно выгодно отличается от существующих программных продуктов простотой в использовании, универсальностью и некоммерческой основой распространения.

# Список используемой литературы

**Учебная и монографическая литература**

1. Бардасов С. А. Эконометрика: учебное пособие. 2-е изд., перераб. и доп. Тюмень: Издательство Тюменского государственного университета, 2010. 264 с.;
2. Барсегян, А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP/А.А. Барсегян. М.С. Куприянов, В.В. Степаненко, И.И. Холод. – 2-е изд., перераб. И доп. – СПб.: БХВ-Петербург, 2007. – 384 с.;
3. Бородич С.А. Вводный курс эконометрики: Учебное пособие – Мн.: БГУ, 2000. – 354 с.;
4. Люгер, Джордж, Ф. Искусственный интеллект: стратегии и методы решения сложных проблем, 4-е издание.: Пер. с англ. – М.: Издательский дом «Вильямс», 2003. – 865с.;
5. Паклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие. 2-е изд., испр. - СПб.: Питер, 2013. - 704 с.;
6. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA М., МедиаСфера, 2002, 312с.;
7. Суслов В.И, Ибрагимов Н.М. и др. Эконометрия. Учебник -2005 г. -742с.;

**Интернет-источники**

1. <http://www.statistica.ru/home/portal/taskboards/logitregression.htm> Дата обращения: 01.05.2013
2. <http://www.basegroup.ru/library/analysis/regression/logistic/> Дата обращения: 12.05.2013

# Приложения

## Приложение 1. Листинг программы