САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА МАТЕМАТИЧЕСКОЙ ТЕОРИИ ИГР И СТАТИСТИЧЕСКИХ РЕШЕНИЙ

Козыкин Алексей Алексеевич

Выпускная квалификационная работа бакалавра

Моделирование качества жизни в Санкт-Петербурге на основе данных социологических опросов жителей

Направление 010400

Прикладная математика, фундаментальная информатика
и основы программирования

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Тарашнина С. И.

Санкт-Петербург
2016

Содержание

Введение

Одним из этапов социологического исследования является анализ полученных данных и их интерпретация. Этот этап вызывает наибольшие затруднения в процессе исследования. Данная работа посвящена выявлению скрытых закономерностей в данных социологических опросов населения с целью получения наглядной структуры, способствующей качественной интерпретации исследуемых процессов.

В настоящее время, в связи с появлением большого объема количественных данных, все чаще возникает потребность для исследовательских работ применение методов Data Mining. Принципам и перспективам использования методов Data Mining посвящено множество научных статей с практической значимостью в сферах медицины, психологии, банковского сектора, криминалистики и других. В качестве примера социологических исследований можно привести исследования рынка занятости населения [3], а также задачу выявления предрасположенности к наркозависимости [8].

Санкт-Петербургским информационно-аналитическим центром, работающим в области информатизации и информационно-аналитического обеспечения органов государственной власти, ежегодно проводится социологический мониторинг [7] с целью выявления основных городских проблем (по мнению жителей) и их устранения органами государственной власти. В работе используются данные мониторинга за апрель 2013 года, полученные путем анкетирования с закрытым характером вопросов и соблюдением всех условий репрезентативности.

Постановка задачи

Пусть имеются данные социологического опроса населения, направленного на изучение определенной проблематики. В данном исследовании изучается отношение жителей к качеству жизни в Санкт-Петербурге. Проведен опрос населения Санкт-Петербурга: выборка 1197 человек в возрасте 18 лет и старше с квотированием по полу и возрасту.

Целью исследования является выявление скрытых закономерностей в данных социологических опросов населения, определение наиболее значимых факторов, оказывающих влияние на качество жизни петербуржцев.

Для достижения поставленных целей должны быть решены следующие задачи:

1. Изучить три метода «Data mining»: искусственные нейронные сети, логистическая регрессия и деревья решений.
2. Разработать модуль, реализующий рассматриваемые методы.
3. Подготовить данные социологических опросов к использованию в разработанном модуле.
4. С помощью разработанного модуля оценить качество построенных моделей и выявить статистически значимые факторы в каждой модели.
5. Сформировать отчет о результатах, содержащих модель наилучшего качества.

Теоретические аспекты исследования приведены в Главе 1.

Глава 2 посвящена программной реализации модуля в прикладном пакете MATLAB и сравнению промежуточных результатов.

Обзор литературы

На основе исследования, проведенного специалистами Санкт-Петербургского информационно-аналитического центра, о применении информационных технологий в целях обеспечения социально-экономического прогнозирования развития региона [2] сформулирована задача выявления скрытых закономерностей в данных социологических опросов.

Основные проблемы выявления закономерностей в данных социологических опросов, а также примеры и методы анализа и интерпретации данных в социологии описываются в [6]. Принципы и особенности применения методов Data Mining представлены в [3] и [9].

В работе рассматриваются модели искусственной нейронной сети, логистической регрессии и бинарного дерева решений. Алгоритмические структуры моделей описаны в учебных пособиях [4, 5], статьях по классификации данных на сайте [9] и виртуальных курсах вашингтонского университета по машинному обучению [10]. В ходе изучения методов классификации рассмотрены методы нахождения минимума ошибки, описанные в [1] и [11].

Использование данных социологических опросов в моделях требует перевода их в численный вид с определенными ограничениями. Необходимость этой процедуры описана в информационных источниках [9, 10].

Алгоритмы реализации рассмотренных методов можно найти на официальном сайте технической поддержки компании «MathWorks» [12] и статьях на сайте по машинному обучению [11].

Глава 1. Теоретическое описание модуля и методов классификации

Данные социологического опроса представляют из себя таблицу, где столбец соответствует номеру опрашиваемого респондента, а строка – номеру вопроса, задаваемого респонденту. В каждой ячейке хранится значение из номинальной шкалы ответов, что позволяет лишь проверять эквивалентность ответов и группировать их. Для корректной работы моделей данные необходимо преобразовать в численный вид, а также ранжировать по значимости. Полученная таблица с ранжированными числами в каждой ячейке используется в дальнейшем исследовании.

1.1. Формат данных

Обозначим через x_j – фактор, описывающий отношение респондентов к вопросу j, всего факторов m. Будем считать вектор $x^{(i)} = (x_1^{(i)},...,x_m^{(i)})$ ответами респондента i, y_i – ответом респондента i на контрольный вопрос, всего n респондентов.

Из полученной матрицы $X_{m\times n} = \left\{ x_j^{(i)} \right\}_{\substack{i=\overline{1,n} \\ j=\overline{1,m}}}$ необходимо выбрать непересекающиеся множества для обучения и тестирования моделей. Обозначим обучающее множество как $A_{m\times(n-k)}$ и тестовое множество как $B_{m\times k}$. Значения факторов каждого респондента и тестового множества $x_j^{(\tilde{i})} \in B$ должны быть ограничены максимальным и минимальным значениями соответствующих факторов обучающего множества $A_{m\times(n-k)}$, то есть для $x_j^{(\tilde{i})} \in B$ имеет место:

$$\min_{i, x_j^{(i)} \in A} x_j^{(i)} \le x_j^{(\tilde{i})} \le \max_{i, x_j^{(i)} \in A} x_j^{(i)}, \quad j = \overline{1,m}, \tilde{i} = \overline{1,k}.$$

Значение ответа респондента на контрольный вопрос может быть равно только $y_i = 1$ или $y_i = 0,$, что означает принадлежность респондента к классу «1» или классу «0» соответственно.

Данные, представленные в таком виде, можно использовать в классифицирующих моделях для выявления статистической зависимости:

$$y_i = f(x^{(i)}) + \varepsilon_i, \quad i = \overline{1,n},$$

где ε_i – ошибки.

1.2. Искусственная нейронная сеть

Искусственная нейронная сеть – это математическая модель, напоминающая по своей структуре нейронную сеть головного мозга и способная на основе обучающего множества создать обобщенную зависимость результирующего значения от факторов, для обоснованной классификации данных, не входивших в обучающее множество.

Архитектура нейронной сети характеризуется количеством входов, выходов, слоев, нейронов в каждом слое, функцией обучения и функцией активации.

Нейрон – единица информации нейронной сети, определяющаяся набором входных сигналов, весами, смещениями, сумматором и функцией активации (рис. 1).

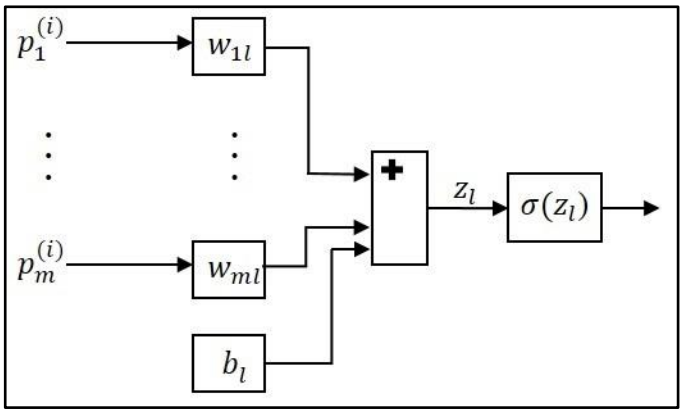


Рис. 1: Схема нейрона. l – номер нейрона, σ –функция активации, i – номер слоя

Каждый набор сигналов $p_1^{(i)},\dots,p_m^{(i)}$, входящий в нейрон, умножается на соответствующие нейрону l веса $w_{1l},\dots,w_{ml}$. Полученные значения суммируются со смещением b_l, представляя собой линейную комбинацию:

$$z_l = \sum_{j=1}^{m} p_j^{(i)} w_{jl} + b_l.$$

Значение z_l сжимается с помощью функции активации σ. Таким образом, функция активации ограничивает амплитуду выходного сигнала. В нашем случае используется сигмоидальная функция активации:

$$\sigma(z_l) = \frac{2}{1+e^{-2z_l}} - 1.$$

Обучение нейронной сети происходит по методу обратного распространения ошибки, в котором минимизируется среднеквадратическая ошибка E нейронной сети:

$$E = \frac{1}{2}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2,$$

где $\tilde{y}$ – результат нейронной сети.

Для минимизации используется итеративный метод градиентного спуска с адаптивным обучением. Весовые коэффициенты и смещения определяются по формулам:

$$w_{jl}(t+1) = w_{jl}(t) - h \cdot \frac{\partial E(l)}{\partial w_{jl}(t)}, \quad j = \overline{1,\tilde{m}},$$

$$b_l(t+1) = b_l(t) - h \cdot \frac{\partial E(l)}{\partial b_l(t)},$$

где l – номер нейрона, $E(l)$ – среднеквадратическая ошибка в нейроне, $\tilde{m}$ – число входных сигналов в нейроне, t – итерация. Адаптивный шаг обучения h определяется по формуле:

$$h(t) = \frac{1}{1 + \sum_{j=1}^{\tilde{m}} (p_j^{(i)}(t))^2}.$$

Используется двуслойный персептрон. Количество входов и выходов определяется количеством факторов и видом данных. Соответственно, если на вход подается вектор размерности $m \times 1$, а на выход – 1×1, то количество входов – m, количество выходов – 1. Для выбора количества нейронов необходимо рассмотреть множество альтернатив. Архитектуру нейронной сети можно представить в виде блок-схемы (рис. 2).

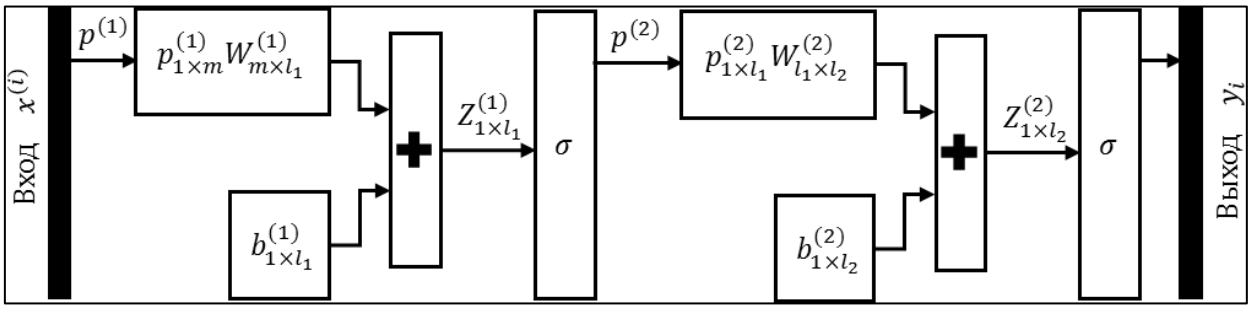


Рис. 2: Архитектура нейронной сети

1.3. Логистическая регрессия

Логистическая регрессия – модель, являющаяся частным случаем метода обобщенной линейной регрессии, использующаяся для бинарной классификации. Предсказывает вероятность попадания в класс «1». Архитектура данной модели значительно проще, чем архитектура нейронной сети. Она характеризуется количеством входов, количеством выходов и функцией вероятности попадания в класс «1».

Количество входов и выходов определяется количеством факторов и видом данных. Функция вероятности имеет вид:

$$f(x^{(i)}) = \sigma(z(x^{(i)})) = \frac{1}{1 + e^{-z(x^{(i)})}},$$

$$z(x^{(i)}) = \sum_{j=1}^{m} x_j^{(i)} w_j + b,$$

где w_j – веса функции, b – смещение.

С помощью метода наименьших квадратов с итеративным пересчетом весов можно определить веса $w = (w_1, \ldots, w_m)$. Для этого используется функция качества аппроксимации:

$$E(w) = \sum_{i=1}^{n} \ln(\sigma(-x^{(i)} w^T y_i),$$

где σ – сигмоидальная функция. По методу Ньютона-Рафсона итеративно вычисляется оптимальный вектор весовых коэффициентов:

$$w(t+1) = w(t) - h \cdot \left(\frac{\partial^2 E(w(t))}{\partial w^2(t)} \right)^{-1} \cdot \frac{\partial E(w(t))}{\partial w(t)}, \quad j = \overline{1, m},$$

где h определяется количеством итераций. Алгоритм останавливается при условии

$$\|w(t) - w(t+1)\| < w(t) \cdot 10^{-6}.$$

1.4. Дерево решений

Дерево решений – это модель, классифицирующая данные путем создания иерархической структуры правил типа «если..., то…». Выбрана модель бинарного дерева решений для мультиклассовой классификации, в каждом узле которого задано правило «если фактор $x_j^{(i)} \geq Q_{j_k}$, то перейти к правому узлу следующего уровня, иначе – к левому», где i – номер элемента, j – номер фактора, k – номер ограничения на фактор j. Конечные узлы дерева содержат классы «0» и «1». Для определения правил необходимо для всех факторов вычислить и минимизировать среднеквадратические ошибки:

$$E_j = \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \tilde{y}_i)^2, \quad j = \overline{1, m},$$

$$y_{ij} = \theta(x_j^{(i)} - Q_{j_k}) = \begin{cases} 0, & x_j^{(i)} < Q_{j_k}; \\ 1, & x_j^{(i)} \geq Q_{j_k}. \end{cases}$$

θ – функция Хэвисайда. Правила определяются в каждом узле с учетом предыдущих узлов текущей ветви. Таким образом, все данные описываются одной иерархической структурой (рис. 3).

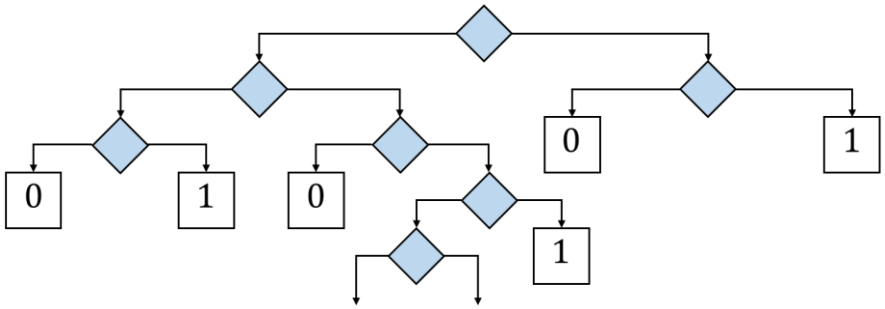


Рис. 3: Структура бинарного дерева решений

1.5. Критерий сравнения

Чтобы сравнить модели и выбрать «лучшую», необходимо оценить их качество. Для этого будем использовать выделенное ранее множество тестовых данных B. Сравнивая смоделированные результаты элементов тестового множества с истинными, выделим следующие характеристики моделей:

1. True Positives (TP) - верно классифицированные элементы с результатом моделирования «1».
2. True Negatives (TN) - верно классифицированные элементы с результатом моделирования «0».
3. False Positives (FP) – неверно классифицированные элементы с результатом моделирования «1».
4. False Negatives (FN) – неверно классифицированные элементы с результатом моделирования «0».

С помощью этих характеристик проведем ROC-анализ моделей, позволяющий оценить качество моделей бинарной классификации. ROC-кривая – это график, показывающий отношение чувствительности (TPR) алгоритма классификации к его не специфичности ($1-FPR$) (рис. 4).

Чувствительность (TPR) – отношение количества верно классифицированных элементов с результатом моделирования «1» к количеству истинных значений «1».

Специфичность (FPR) – отношение количества верно классифицированных элементов с результатом моделирования «0» к количеству истинных значений «0».

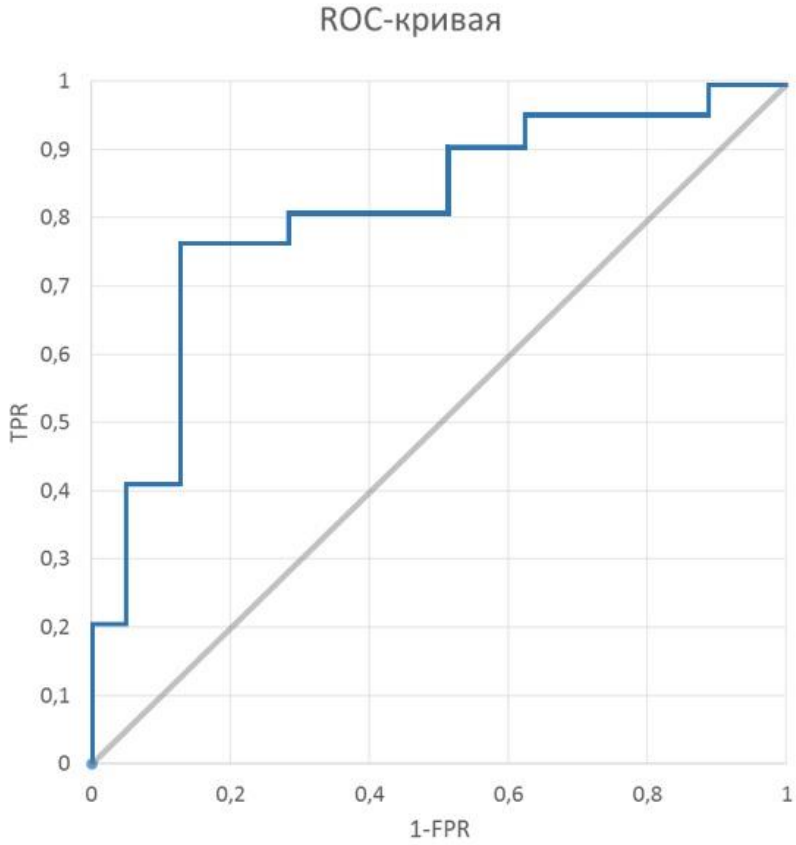


Рис. 4: Пример ROC-кривой

При визуальной оценке ROC-кривой расположение графика вдоль диагонали указывает на неэффективность модели.

Для числовой оценки модели используется значение AUC (Area Under Curve) – площадь фигуры, ограниченной сверху ROC-кривой. Значение $AUC \in [0,1]$, при этом $AUC = 0.5$ означает «гадание», 1 – правильную классификацию всех элементов, (-1) – неправильную классификацию всех элементов. В качестве критерия сравнения используется значение качества модели AUC.

1.6. Значимость факторов

В каждой модели определяются незначимые и малозначимые факторы путем их исключения. Если, исключая фактор, качество модели возрастает или остается прежним, то исключенный фактор считается незначимым и далее не рассматривается. Для оставшихся $\hat{m}$ факторов будут определены коэффициенты значимости BF, рассчитанные по формуле:

$$BF_j = 2 \cdot (AUC_{\hat{m}} - AUC_j), \quad j = \overline{1, \hat{m}}, \tag{1}$$

где $AUC_{\hat{m}}$ – точность модели с $\hat{m}$ факторами, AUC_j – точность модели с исключенным фактором x_j.

1.7. Выводы

В этой главе исследованы методы классификации и методы нахождения минимума функций ошибок. Также для дальнейшей программной реализации модуля определены: вид данных, подающихся на вход модели, критерий сравнения моделей и значимость факторов.

Глава 2. Реализация модуля

2.1. Входные данные

В апреле 2013 года СПб ГУП «СПб ИАЦ» проводился социологический опрос, в котором респондентам задавались вопросы, касающиеся разных аспектов их жизни, с предложенными вариантами ответов. Анкета представлена в Приложении 1. Данные сгруппированы в виде таблицы, где номер строки соответствует номеру опрашиваемого респондента, а столбец – номеру вопроса. Таблица содержит 1197 строк и 28 столбцов. Для корректной работы с моделью данные были предварительно обработаны. Вариантам ответов приписаны числовые значения, ранжированные по значимости от 0 до 1. В 15 и 26 вопросах взяты средние значения от суммы их подпунктов. Таким образом перечисленные вопросы можно рассматривать как переменные модели, а именно:

- x_1 – пол;
- x_2 – возраст;
- x_3 – образование;
- x_4 – оценка качества жизни в будущем (через 6 месяцев);
- x_5 – оценка направления развития РФ;
- x_6 – готовность на какие-либо формы протеста;
- x_7 – отношение к проблемам межнациональных отношений;
- x_8 – оценка характера межнациональных отношений в РФ;
- x_9 – оценка угрозы распада России;
- x_{10} – отношение к приезжим из других регионов России;
- x_{11} – отношение к приезжим из стран ближнего зарубежья (СНГ и Прибалтика);
- x_{12} – отношение к приезжим из стран дальнего зарубежья;

- x_{13} – отношение к трудовым мигрантам;
- x_{14} – отношение к представителям различных национальностей;
- x_{15} – оценка услуг ЖКХ;
- x_{16} – оценка уровня преступности;
- x_{17} – отношение к банкам и подобным финансовым учреждениям;
- x_{18} – отношение к работе;
- x_{19} – отношение мужчин к детям по мнению респондента;
- x_{20} – отношение женщин к детям по мнению респондента;
- x_{21} – отношение к финансовой независимости женщин в семье;
- x_{22} – оценка высказывания о том, что семья – это главное предназначение женщины;
- x_{23} – семейное положение;
- x_{24} – количество детей в возрасте до 18 лет;
- x_{25} – оценка уровня дохода в семье;
- x_{26} – отношение к политике и власти в целом;
- x_{27} – наличие городских проблем, часто касающихся респондента;
- y – отношение респондента к уровню жизни в целом.

В результате мы имеем 27 переменных – факторов $x_1,...,x_{27}$, которые могут оказывать влияние на результирующую переменную – признак y.

Случайным образом из множества $X_{28\times1197}$ выделены множества для обучения $A_{m\times(n-k)}$ и тестирования $B_{m\times k}$, состоящие из $m=28$ факторов, $n-k=898$ элементов и $k=299$ элементов соответственно. Важной особенностью данных является отношение числа элементов класса «0» и «1». Если множество одного из классов намного больше другого, то ошибка минимизируется в пользу большего класса, то есть обобщение этого класса будет лучше, чем второго. В обучающем множестве 674 элемента класса «1»

и 224 элемента класса «0». В тестовом множестве 208 элементов класса «1» и 91 элемент класса «0». Это означает, что класс «0» будет обобщен хуже.

2.2. Настройка и реализация моделей

Все модели были реализована в прикладном пакете MATLAB.

Модель искусственной нейронной сети

В пакете MATLAB нейронная сеть создается с помощью функций newff:

```
[m,n] = size(A);
minputSet = A(1:m,1:n);
outputSet = A(m,1:n);
PR = minmax(inputSet);
net = newff(PR, [e q 1]);
net.trainFcn = 'traingda';
net.trainParam.epochs = 1000;
net = train(net, inputSet, outputSet);
```

PR – матрица максимальных и минимальных значений факторов, trainFcn – функция обучения, epochs – максимальное количество итераций обучения.

Точное количество нейронов для данных с нечетко выраженной структурой невозможно рассчитать аналитически, поэтому используем метод равномерного поиска оптимальной структуры. Количество связей нейронов не должно превышать количества элементов. Так же не должно наблюдаться эффекта переобучения, когда модель запоминает все примеры обучающего множества, а не обобщает их. Эффект переобучения будет проверен с помощью значения AUC_0 на обучающем множестве A, он не должен превышать значение 0.9. По следующему алгоритму вычислим значения

AUC и AUC_0 на тестовом и обучающем множествах соответственно для каждой структуры нейронной сети:

```
AUC0 = zeros(30,30);
AUC = zeros(30,30);
[m,n] = size(A);
inputSet = A(1:m-1,1:n);
outputSet = A(m,1:n);
[m,n] = size(B);
inputTestSet = B(1:m-1,1:n);
outputTestSet = B(m,1:n);
PR = minmax(inputSet);
for e = 1:30
    for q = 1:round(30/e)
        net = newff(PR, [e q 1]);
        net.trainFcn = 'traingda';
        net.trainParam.epochs = 1000;
        net = train(net, inputSet, outputSet);
        outputCode0 = sim(net, inputSet);
        outputCode = sim(net, inputTestSet);
        [~,~,~,AUC0(e,q)] = perfcurve(outputSet,outputCode0,1);
        [~,~,~,AUC(e,q)] = perfcurve(outputTestSet,outputCode,1);
    end
end
```

На основе результатов выбрана структура с 3 нейронами в первом слое и 3 нейронами во втором слое, при которой $AUC_0(3,3) = 0.7223$ и $AUC(3,3) = 0.6364$.

По причине сложности достижения глобального минимума функции ошибки модель нейронной сети имеет важный недостаток – обучение может

пройти не оптимально. Для дальнейших исследований, связанных с моделью нейронной сети, будем использовать среднее качество моделей от 15 испытаний. Количество испытаний определено итеративно, до момента, когда следующее испытание дает отклонение среднего значения менее, чем на 0.001.

Модель логистической регрессии

В пакете MATLAB модель обобщенной линейной регрессии создается с помощью функции glmfit:

```
[m,n] = size(A);
inputSet = A(1:m-1,1:n);
outputSet = A(m,1:n);
[m,n] = size(B);
inputTestSet = B(1:m-1,1:n);
outputTestSet = B(m,1:n);
bHat = glmfit(inputSet,outputSet,'binomial','logit');
outputCode = glmval(bHat, inputTestSet, 'logit');
```

Модель логистической регрессии является частным случаем обобщенной линейной регрессии, поэтому достаточно задать тип функции glmfit как 'binomial', 'logit'.

Для вычисления порога отсечения μ^* определим для него требование баланса чувствительности и специфичности:

$$\max_{\mu}(TPR(\mu) + FPR(\mu)).$$

Данный подход реализован в пакете MATLAB с шагом 0.001, результат проиллюстрирован на графике (рис. 5).

```
[m,n] = size(A);
```

```
inputSet = A(1:m-1,1:n);
outputSet = A(m,1:n);
[m,n] = size(B);
inputTestSet = B(1:m-1,1:n);
outputTestSet = B(m,1:n);
bHat = glmfit(inputSet, outputSet,'binomial','logit');
outputCode = glmval(bHat, inputTestSet, 'logit');
yHat = outputCode;
maxFcn = zeros(2);
for i = 1:1000
    for j = 1:299
        if yHat(j,1) < i/1000
            outputCode (j,1) = 0;
        else
            outputCode (j,1) = 1;
        end
    end
    [MTPR,MFPR,~,~] = perfcurve(outputTestSet, outputCode,1);
    TPR(j) = MTPR(2);
    FPR(j) = MFPR(2);
    if (TPR(j)+1-FPR(j))> maxFcn(1)
        maxFcn(1)= (TPR(j)+1-FPR(j));
        maxFcn(2)= i/1000;
    end
end
```

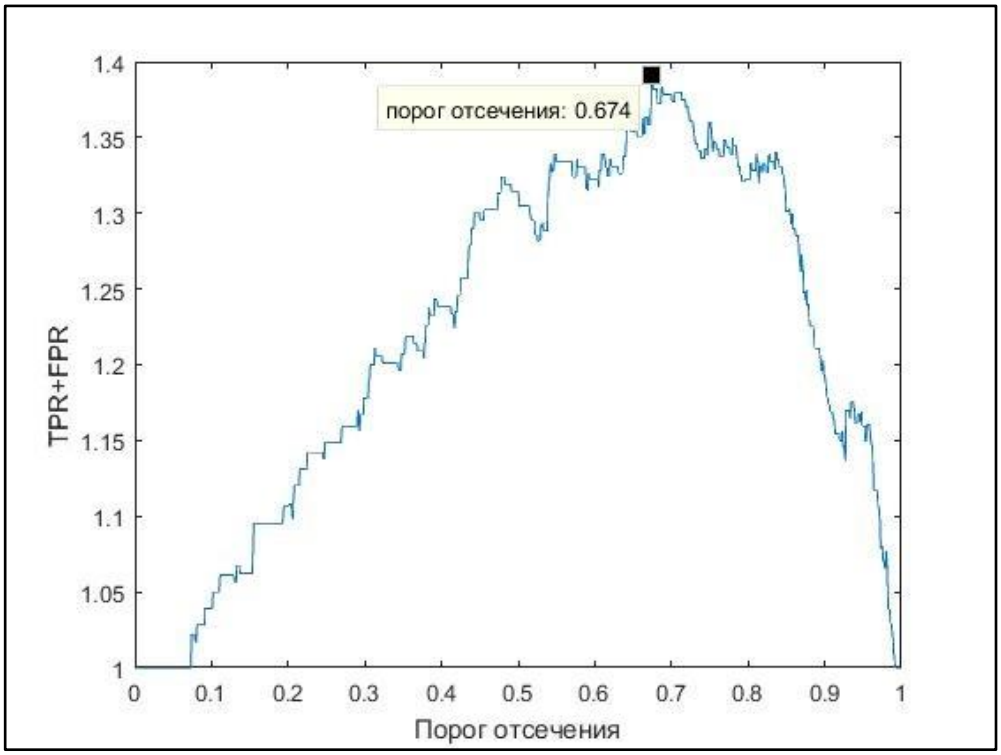


Рис. 5: График зависимости суммы чувствительности и специфичности от порога отсечения

При максимальном значении суммы $TPR(\mu) + FPR(\mu)$ порог отсечения $\mu^* = 0.674$.

Модель дерева решения

В пакете MATLAB модель бинарного дерева решений создается с помощью функции fitctree:

```
[m,n] = size(A);
inputSet = A(1:m-1,1:n);
outputSet = A(m,1:n);
[m,n] = size(B);
inputTestSet = B(1:m-1,1:n);
outputTestSet = B(m,1:n);
Tree = fitctree (inputSet, outputSet);
outputCode = predict(Tree, inputTestSet)
```

Данная модель не потребовала дополнительных настроек. Значения, установленные по умолчанию, удовлетворяют требованиям данной задачи.

2.3. Результат работы модуля

Исключение факторов и выбор модели

На основе анализа значимости факторов исключена часть факторов и вычислена точность моделей.

```
[m,n] = size(A);
inputSet = A(1:m-1,1:n);
outputSet = A(m,1:n);
[m,n] = size(B);
inputTestSet = B(1:m-1,1:n);
outputTestSet = B(m,1:n);
excludedFctr = -1;
result = zeros(m-1,1);
factors = [1:m-1]';
while excludedFctr ~= 0
    excludedFctr = 0;
    [m,n] = size(inputSet);
    AUC = zeros(m);
    TPR = zeros(m);
    FPR = zeros(m);
    for j = 1:m
        if j == 1
            inputSet1 = inputSet(2:m, 1:898);
            inputTestSet1 = inputTestSet(2:m, 1:299);
        end
        if j == m
            inputSet1 = inputSet(1:m-1, 1:898);
            inputTestSet1 = inputTestSet(1:m-1, 1:299);
        end
        if j > 1 && j < m
```

```
                inputSet1 = inputSet([1:j-1, j+1:m], 1:n);
                inputTestSet1 = inputTestSet([1:j-1, j+1:m], 1:299);
            end
%/...алгоритм модели, с входом inputSet1, inputTestSet1, outputSet, outputTestSet .../
        [TPRM,FPRM,~,AUC(j)]=perfcurve(outputTestSet,outputCode,1);
            TPR(j)= TPRM;
            FPR(j)= FPRM;
            if AUC(j) >= AUCmax
                AUCmax = AUC(j);
                excludedFctr = j;
            end
        end
        if j ~= 0
            inputSet(excludedFctr,:) = [];
            inputTestSet (excludedFctr,:) = [];
            factors(excludedFctr) = [];
        end
        if result == 0
            result = [AUC(:),TPR(:),FPR(:),factors];
        else
            result = [AUC(:),TPR(:),FPR(:),factors];
        end
end
```

В таблицах 1 – 4 приведены результаты построения моделей при исключении незначимых факторов. Процедура исключения проводится итерационно. На каждой итерации исключается фактор с наименьшим показателем значимости и далее не рассматривается. Для каждой модели построен график ROC-кривых для конечных результатов (рис. 6–8).

Таблица 1: Результаты исключения незначимых факторов модели нейронной сети

Номер удаленного фактора	*AUC*	*TP*	*TN*	*FP*	*FN*
Без удаления	0.6364	171	41	50	37
13	0.6546	173.6	43.2	47.8	34.4
19	0.6619	168.867	46,6	44.4	39.133

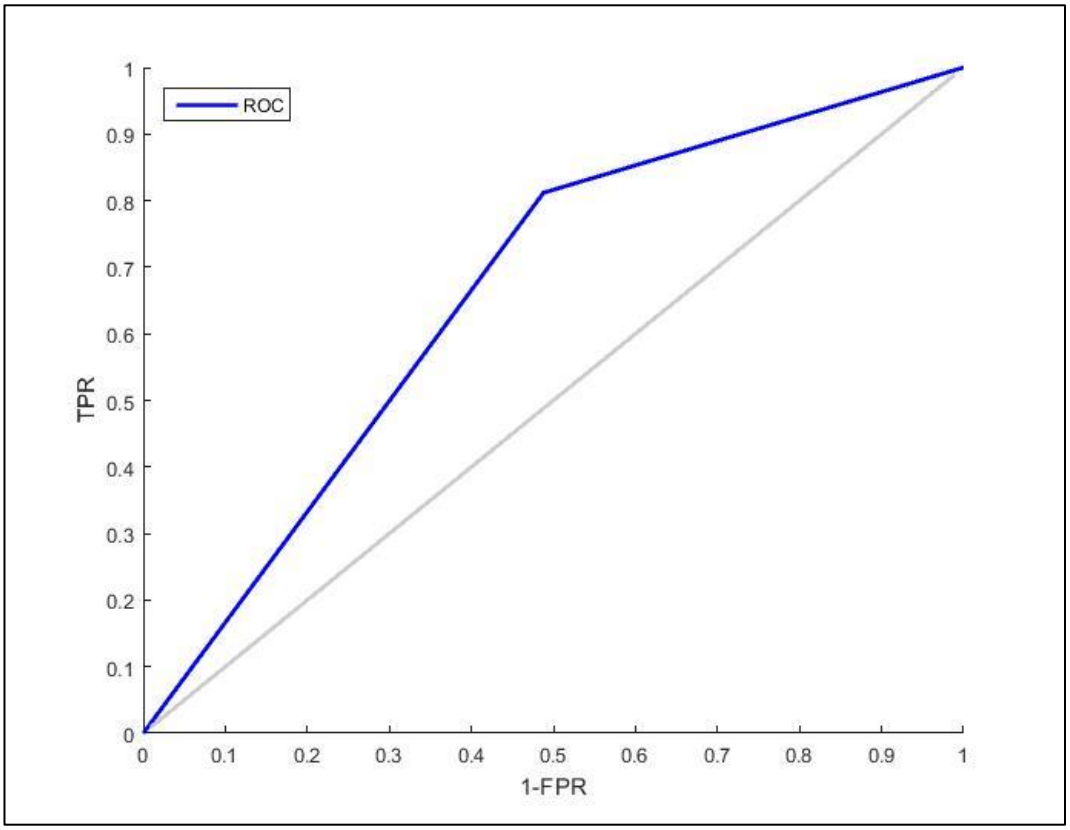


Рис. 6: ROC-кривая модели нейронной сети

Таблица 2: Результаты исключения незначимых факторов модели логистической регрессии

Номер удаленного фактора	*AUC*	*TP*	*TN*	*FP*	*FN*
Без удаления	0.6957	150	61	30	58
19	0.7053	154	61	30	54
20	0.7053	154	61	30	54
1	0.7108	154	62	29	54
23	0.7108	154	62	29	54

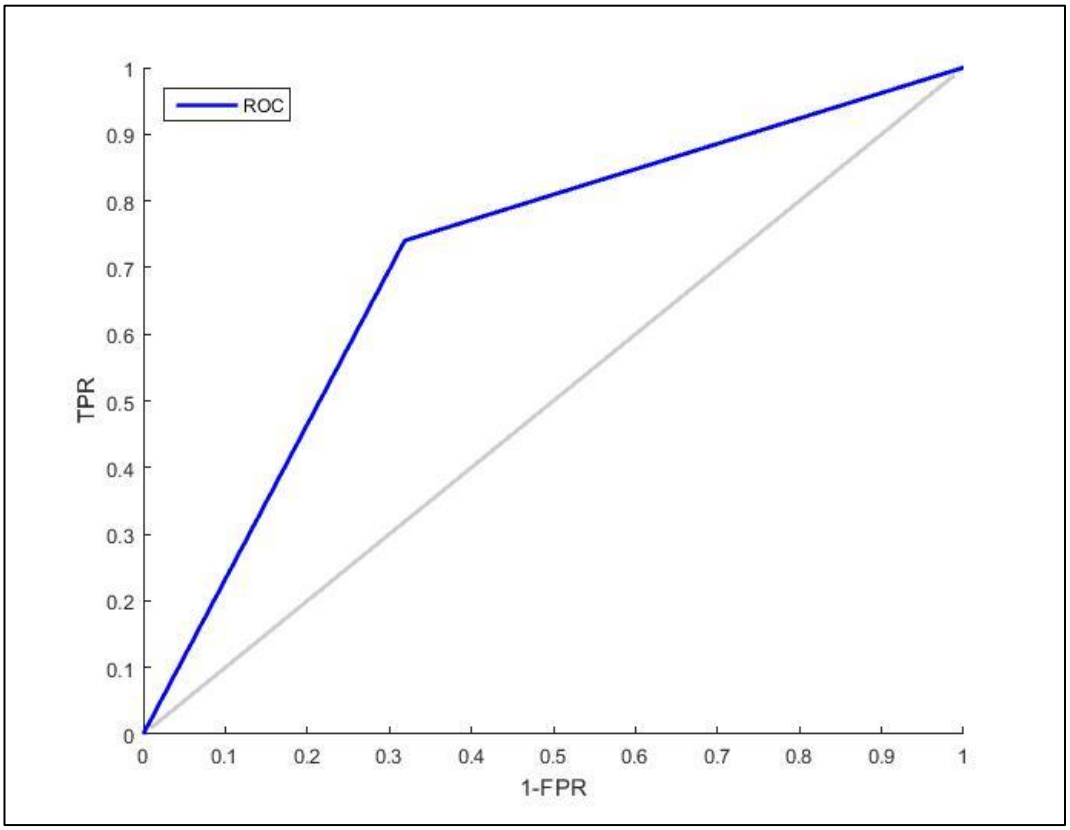


Рис. 7: ROC-кривая модели логистической регрессии

Таблица 3: Результаты исключения незначимых факторов модели дерева решений

Номер удаленного фактора	*AUC*	*TP*	*TN*	*FP*	*FN*
Без удаления	0.6473	171	43	48	37
5	0.6497	172	43	48	36
18	0.6545	174	43	48	34
27	0.6569	175	43	48	33
13	0.6569	175	43	48	33
16	0.6569	175	43	48	33

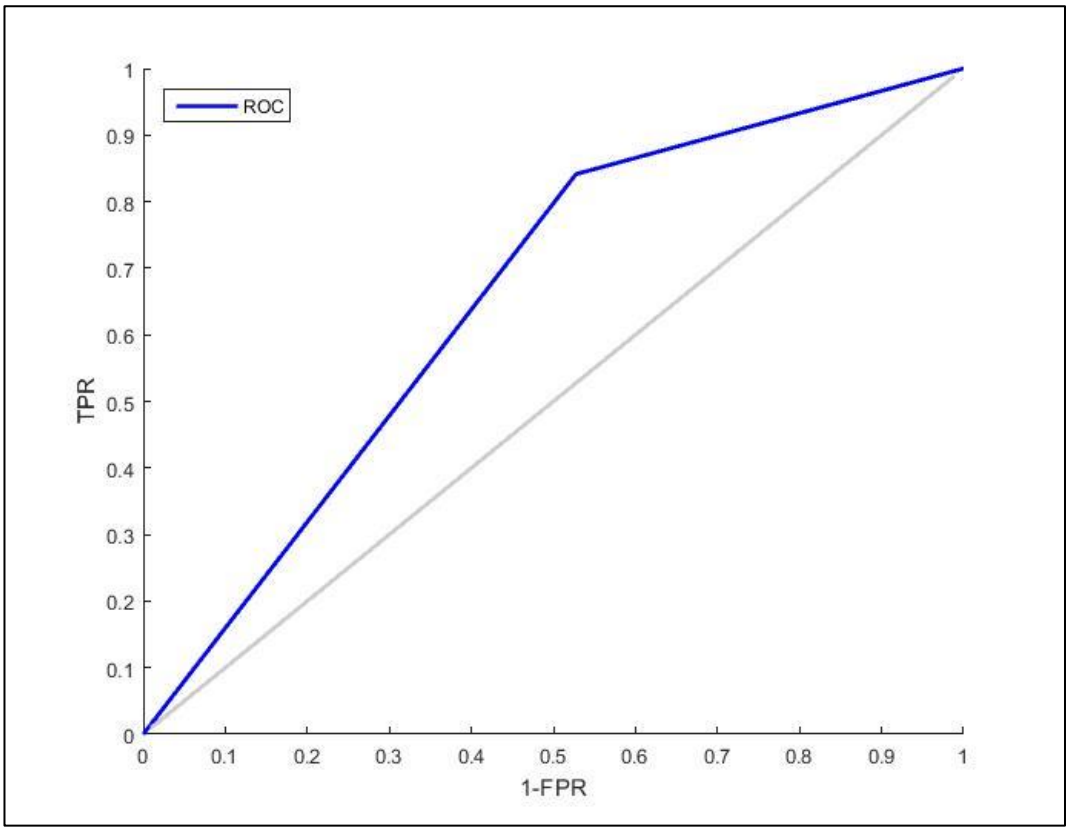


Рис. 8: ROC-кривая модели дерева решений

Таблица 4: Конечные результаты по всем моделям

Модель	*AUC*	*TP*	*TN*	*FP*	*FN*
Нейронная сеть	0.6619	168.86	46.6	44.4	39.13
Логистическая регрессия	0.71085	154	62	29	54
Дерево решений	0.6655	174	45	46	34

Модель логистической регрессии имеет наибольшее значение качества *AUC* .

Выделение значимых факторов и создание отчета

По формуле (1) рассчитывается коэффициент значимости факторов и создается отчет результатов модуля на основе модели логистической регрессии.

Формат отчета реализован в отдельном m-файле:

```
%% Reports
%% The result of the module operation
```

```
if checkConst==1
bestmodel = 'Artificial neural network';
end
if checkConst==2
bestmodel = 'Logistic regression';
end
if checkConst==3
bestmodel = 'Binary decision tree';
end
fprintf('Model with the best quality: %s ',bestmodel);
%% *Significant factors*
significance1 = (AUCmax-result(:,1))*2;
factors1= result(:,4);
nameFactors2 = [{'sex of the person'};{'age'};{'education'};{'evaluation of the quality of life in the future'};{'evaluation of the Russian Federation'};{'readiness on any form of protest'};{'attitude to the problems of international relations'};{'evaluation of the nature of international relations in the Russian Federation'};{'assess the threat of Russias disintegration'};{'relation to visitors from other regions of Russia'};{'attitude toward visitors from the CIS countries'};{'relevant to visitors from foreign countries'};{'relevant to migrant workers'};{'attitude to the representatives of different nationalities'};{'assessment of housing services'};{'assessment of the level of crime'};{'ratio for banks and similar financial institutions'};{'attitude to work'};{'male attitudes towards children in the opinion of the respondent'};{'female ratio of children in the opinion of the respondent'};{'relation to the financial independence of women in the family'};{'assessment statement that the family is the main purpose of a woman'};{'family status'};{'the number of children under the age of 18 years'};{'evaluation of the level of income in the family'};{'attitude toward
```

```
politics and government as a whole'};{'existence of urban problems, often related to the respondent'}];
nameFactors = [];
factors2 = [1:27]';
for j = size(factors1,1):-1:1
   for i=size(factors2,1):-1:1
      if factors1(j) == factors2(i)
         factors2(i) = [];
         nameFactors = [nameFactors2(factors1(j)); nameFactors];
         nameFactors2(factors1(j)) = [];
      end
   end
end
clear i;
clear j;
porog=sum(significance1)*sum(significance1)/size(significance1,1);
for i=size(significance1,1):-1:1
        if significance1(i) < porog
                nameFactors(i) = [];
                factors1(i) = [];
                significance1(i) = [];
end
end
table(factors1,significance1,nameFactors)
%% *Insignificant factors*
table(factors2, nameFactors2)
h = figure; hold on
plotroc(outputTestSet,maxoutputCode');
datetick('x','m');
```

```
axis tight
xlabel('1-FPR');
ylabel('TPR');
saveas(h,'report/img/roc','jpg');
close(h);
%% Roc-analysis
% <<img/roc.jpg>>
```

Шаблон включает в себя таблицы результатов модуля и график ROC-кривой, соответствующие модели с наибольшим показателем *AUC*. Вызывается командой publish с указанием директории хранения отчета ‘report’ и формата отчета ‘pdf’:

```
opts.outputDir = 'report';
opts.format = 'pdf';
opts.showCode = false;
publish ('report',opts);
```

Приведен пример отчета о результатах моделирования (рис. 9).

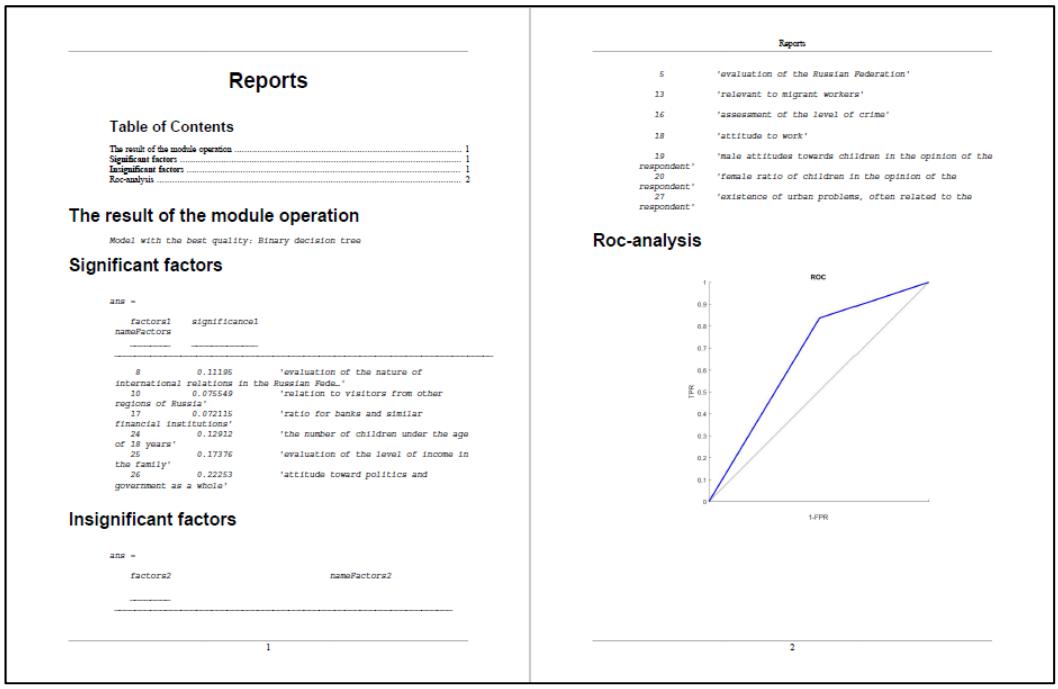


Рис. 9: Пример отчета о результатах моделирования на основе модели бинарного дерева решений

Отчет о результатах моделирования содержит две таблицы и график ROC-кривой (Приложение 2).

Таблица незначимых факторов содержит:

- x_1 – пол;
- x_{19} – отношение мужчин к детям по мнению респондента;
- x_{20} – отношение женщин к детям по мнению респондента;
- x_{23} – семейное положение.

Таблица значимых факторов содержит:

- x_2 – возраст;
- x_4 – оценка качества жизни в будущем (через 6 месяцев);
- x_5 – оценка направления развития РФ;
- x_{14} – отношение к представителям различных национальностей;
- x_{16} – оценка уровня преступности;
- x_{24} – количество детей в возрасте до 18 лет;
- x_{25} – оценка уровня дохода в семье;
- x_{27} – наличие городских проблем, часто касающихся респондента.

Среди значимых факторов наибольшее влияние на оценку качества жизни оказывают возраст, направление развития РФ и уровень дохода в семье. Исследование показывает, что старшее поколение в большей степени недовольно качеством жизни. Исключение факторов номер 19, 20, 23, и выделение фактора номер 24 дает понять, что для большинства людей требования к качеству жизни возрастают с появлением детей, но не меняются в зависимости от семейного положения. Отношение к ЖКХ мало повлияло на оценку качества жизни, что может означать удовлетворительную работу этого сервиса, как и остальных сфер, не включенных в список наиболее значимых.

2.4. Выводы

В этой главе ставилась задача программно реализовать модуль, позволяющий сравнивать алгоритмы классификации, выделять значимые и незначимые факторы, получать результаты моделирования в форме отчета. Модуль реализован в пакете MATLAB (Приложение 3), а его применение дало достаточно точный результат в удобном для восприятия виде. Наличие ошибок объясняется неравным количеством обучающих данных класса «1» и «0», а также наличием среди респондентов «абсолютных оптимистов» и «абсолютных пессимистов», искажающих модель поведения среднестатистического жителя.

Заключение

Работа с данными социологических опросов касательно поставленных задач дала следующие результаты:

1. Изучены три метода «Data mining»: искусственные нейронные сети, логистическая регрессия и деревья решений. Определены параметры соответствующих моделей для лучшей классификации данных и критерии сравнения.
2. Разработан модуль, реализующий рассмотренные методы и выбирающий «лучший» их них. Для модуля разработан формат отчета, содержащий наглядную структуру зависимостей факторов от контрольного вопроса, что способствует интерпретации социологического исследования.
3. Средствами разработанного модуля проведено социологическое исследование с целью выявления наиболее значимых аспектов жизни жителей Санкт-Петербурга, влияющих на удовлетворенность жизнью в целом. Выделено восемь значимых факторов и проведена интерпретация полученных результатов.

Список литературы

1. Калацкая Л.В., Новиков В.А., Садов В.С. Организация и обучение искусственных нейронных сетей: учебное пособие. Минск: БГУ, 2003. 75 с.
2. Калиниченко А.Ю., Тарашнина С.И. Информационные технологии в целях обеспечения социально-экономического прогнозирования развития региона. Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики. 2014. с. 229-234.
3. Мальцева А. В., Шилкина Н. Е., Махныткина О. В. Data mining в социологии: опыт и перспективы проведения исследования // Социологические исследования. 2016. № 3. С. 35-44.
4. Потемкин В.Г., Медведев В.С. Нейронные сети. MATLAB 6. Диалог-МИФИ, 2002. 496 с.
5. Саймон Хайкин. Нейронные сети: полный курс, 2-е издание. Вильямс, 2006. 1104 с.
6. Семенов В. Е. Анализ и интерпретация данных в социологии: учебное пособие. Владим. гос. ун-т, 2009. 132 с.
7. СПб ГУП «СПб ИАЦ». Материалы. http://www.iac.spb.ru
8. Ясницкий Л.Н., Грацилёв В.И., Куляшова Ю.С., Черепанов ФМ. Возможности моделирования предрасположенности к наркозависимости методами искусственного интеллекта // Вестник Пермского университета. Философия. Психология. Социология. 2015. Вып. 1(21). С. 61-73.
9. BaseGroup Labs. Статьи. https://basegroup.ru/community/articles
10. Coursera. Онлайн-курсы. https://www.coursera.org

11. MachineLearning. http://www.machinelearning.ru/
12. MathWorks. Поддержка. http://www.mathworks.com/help

Приложение

Приложение 1. Анкета социологического опроса

1. Ваш пол?
   1.1. Мужчина
   1.2. Женщина
2. Ваш возраст?
   2.1. 18-29 лет
   2.2. 30-39 лет
   2.3. 40-49 лет
   2.4. 50-59 лет
   2.5. 60 лет и старше
3. Ваше образование?
   3.1. Начальное, неполное среднее
   3.2. Среднее полное *(средняя школа)*
   3.3. Начальное профессиональное *(профессиональное училище, лицей)*
   3.4. Среднее профессиональное *(техникум, колледж)*
   3.5. Высшее
4. Как изменится в целом Ваша жизнь в течение ближайших 6 месяцев?
   4.1. Ухудшится
   4.2. Останется прежней
   4.3. Улучшится
5. В каком направлении, на Ваш взгляд, идет развитие России?
   5.1. Дела в стране развиваются в правильном направлении
   5.2. Дела в стране идут по неверному пути
   5.3. Трудно сказать
6. Скажите, пожалуйста, готовы ли Вы на какие-либо формы протеста при защите Ваших интересов?
   6.1. Мои интересы достаточно защищены, протестовать нет необходимости

6.2. Ни в каких формах протеста я участвовать просто не готов (а)

6.3. Готов (а) участвовать в некоторых формах протеста

7. Как Вы думаете, проблема межнациональных отношений – это проблема России в целом или только ее отдельных регионов?

   7.1. Это проблема России в целом

   7.2. Это проблема отдельных регионов

   7.3. Трудно сказать

8. Как бы Вы оценили характер межнациональных отношений в России в целом?

   8.1. Как очень напряженные, взрывоопасные

   8.2. Как напряженные

   8.3. Как в целом спокойные

   8.4. Как полностью спокойные, мирные

   8.5. Трудно сказать

9. Существует ли, на Ваш взгляд, в настоящее время угроза распада России?

   9.1. Да, существует значительная угроза

   9.2. Да, угроза существует, но не слишком значительная

   9.3. Нет, такой угрозы не существует

   9.4. Трудно сказать

10. Как, на Ваш взгляд, влияют на ситуацию в Петербурге поселившиеся в нашем городе приезжие из других регионов России?

    10.1. В основном положительно

    10.2. В основном отрицательно

    10.3. Никак не влияют

11. Как, на Ваш взгляд, влияют на ситуацию в Петербурге поселившиеся в нашем городе приезжие из стран ближнего зарубежья *(СНГ и Прибалтика)*?

    11.1. В основном положительно

    11.2. В основном отрицательно

    11.3. Никак не влияют

12. Как, на Ваш взгляд, влияют на ситуацию в Петербурге, поселившиеся в нашем городе приезжие из стран дальнего зарубежья *(страны, не входившие в состав СССР)*?

12.1. В основном положительно

12.2. В основном отрицательно

12.3. Никак не влияют

13. Как, на Ваш взгляд, в целом, влияют на ситуацию в Петербурге трудовые мигранты?

13.1. В основном положительно

13.2. В основном отрицательно

3.Никак не влияют

14. Вы лично испытываете раздражение или неприязнь по отношению к представителям той или иной национальности?

14.1. Безусловно, не испытываю

14.2. Скорее не испытываю

14.3. Скорее испытываю

14.4. Безусловно, испытываю

14.5. Трудно сказать

15. С какими из следующих проблем Вам приходилось встречаться прошедшей зимой? *(При вводе данных указанный вариант ответа фиксируется как «1» в соответствующей переменной, неуказанный – «0»)*

15.1. Отключение теплоснабжения в жилых домах

15.2. Недостаточное теплоснабжение в жилых домах

15.3. Отключение горячего водоснабжения в жилых домах

15.4. Снижение качества горячей воды в жилых домах

15.5. Отключение электроэнергии в жилых домах

15.6. Неудовлетворительная уборка от снега дворов, гололед во дворах

15.7. Неудовлетворительная уборка от снега тротуаров улиц, гололед на тротуарах, улицах

15.8. Неудовлетворительная уборка от снега проезжей части дорог

15.9. Перебои в работе общественного транспорта

15.10. Протечки и другие повреждения кровли

15.11. Угрожающие Вашей безопасности сосульки на кровле

16. Опасаетесь ли Вы того, что на Вас или Ваших близких могут напасть преступники на улице, в подъезде или в других местах?

16.1. Да, опасаюсь

16.2. Нет, не опасаюсь

16.3. Трудно сказать

17. Доверяете ли Вы банкам и подобным им финансовым учреждениям?

17.1. Доверяю полностью

17.2. Доверяю, но лишь отчасти

17.3. Совершенно не доверяю

17.4. Трудно сказать

18. Представьте себе, если бы Вы, например, получили большое наследство, то как бы Вы отнеслись к работе, обеспечивающей Вам средства к существованию?

18.1. Лучше было бы все равно работать

18.2. Лучше было бы не работать

18.3. Трудно сказать

19. Справедливо ли, на Ваш взгляд, утверждение, что для жизнь мужчины неполноценна, если у него нет детей?

19.1. Да, справедливо

19.2. Нет, не справедливо

19.3. Трудно сказать

20. Справедливо ли, на Ваш взгляд, утверждение, что жизнь женщины неполноценна, если у нее нет детей?

20.1. Да, справедливо

20.2. Нет, не справедливо

20.3. Трудно сказать

21. Как Вы считаете, обязательно ли для женщины в семье, независимо от мужа, иметь собственный источник дохода?
    21.1. Да
    21.2. Нет
    21.3. Трудно сказать
22. Согласны ли Вы с утверждением, что семья – это главное предназначение женщины?
    22.1. Да
    22.2. Нет
    22.3. Трудно сказать
23. Ваше семейное положение?
    23.1. Женат/замужем (брак официально зарегистрирован)
    23.2. Состою в гражданском браке
    23.3. Не женат/не замужем
    23.4. Разведен/разведена
    23.5. Вдовец/вдова
24. У Вас есть дети в возрасте до 18 лет?
    24.1. Нет детей
    24.2. Один ребенок
    24.3. Двое детей
    24.4. Трое и более детей
25. Как бы Вы оценили уровень доходов в Вашей семье?
    25.1. Денег не хватает даже на продукты питания
    25.2. На продукты питания денег хватает, но покупка одежды уже вызывает затруднения
    25.3. Денег хватает на продукты и одежду, однако покупка вещей длительного пользования для нас является проблемой
    25.4. Мы можем без труда приобретать вещи длительного пользования, но нам сложно приобретать дорогие вещи

25.5. можем позволить себе приобретать такие дорогие вещи, как квартира, дача

26. Оцените, пожалуйста, по пятибалльной системе эффективность работы: *(Оценка осуществляется по пятибалльной системе: 0 - совершенно неэффективна - 4 - очень эффективна.)*

26.1. Президента Российской Федерации В.В. Путина

26.2. Правительства Российской Федерации

26.3. Государственной Думы

26.4. Судебной власти в Российской Федерации

26.5. Губернатора Санкт-Петербурга Г.С. Полтавченко

26.6. Городской администрации *(Правительства города)*

26.7. Законодательного Собрания *(ЗакСа)*

26.8. Правоохранительных органов *(полиции, ОМОН, ГИБДД, прокуратуры)*

26.9. Муниципальных органов власти *(Органов местного самоуправления)*

27. Скажите, пожалуйста, в последнее время Вам приходилось часто сталкиваться с какими-либо городскими проблемами?

27.1. Да

27.2. Нет

28. В целом Вас устраивает жизнь в настоящее время?

28.1. Не устраивает

28.2. Устраивает

Приложение 2. Отчет на основе результатов модуля

Reports

The result of the module operation

```
Model with the best quality: Logistic regression
```

Significant factors

```
ans =

    factors1    significance1                          nameFactors
    ________    _____________    ____________________________________________________________

     2          0.087912         'age'
     4          0.055632         'evaluation of the quality of life in the future'
     5          0.074176         'evaluation of the Russian Federation'
    14          0.057005         'attitude to the representatives of different nationalities'
    16          0.054945         'assessment of the level of crime'
    24          0.054258         'the number of children under the age of 18 years'
    25           0.11195         'evaluation of the level of income in the family'
    27          0.059753         'existence of urban problems, often related to the respondent'
```

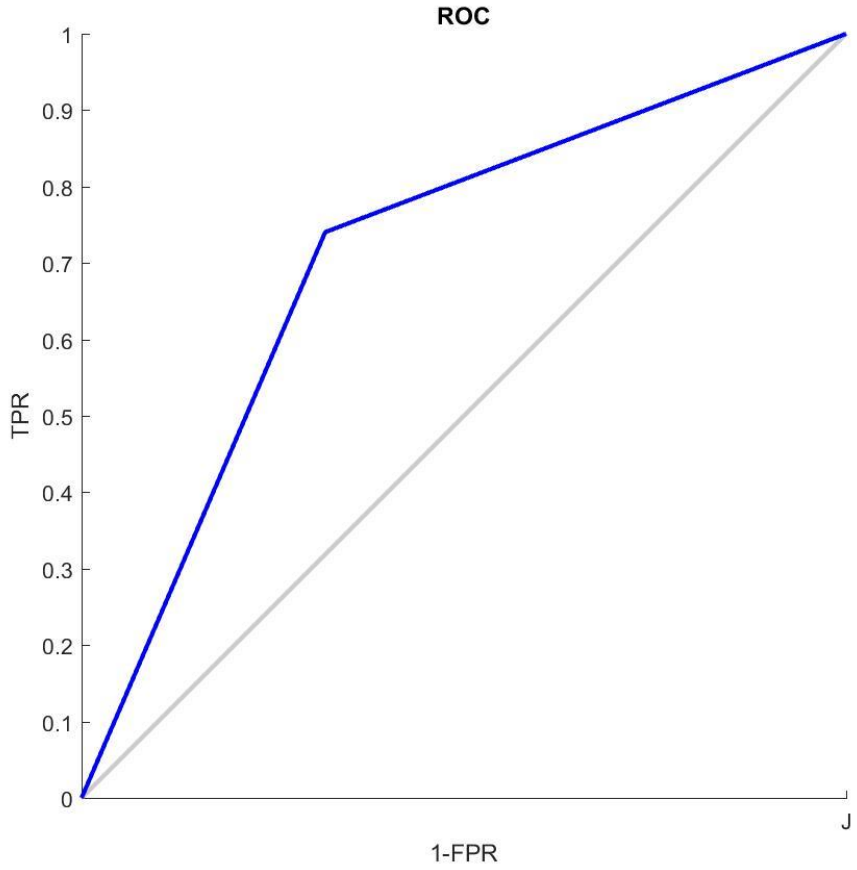
Insignificant factors

```
ans =

    factors2                              nameFactors2
    ________    ____________________________________________________________

     1          'sex of the person'
    19          'male attitudes towards children in the opinion of the respondent'
    20          'female ratio of children in the opinion of the respondent'
    23          'family status'
```

Roc-analysis



Приложение 3. Реализация модуля в пакете MATLAB

```
AUCmax=0;
for checkConst1 = 1:3
    [m,n] = size(A);
    inputSet = A(1:m-1,1:n);
    outputSet = A(m,1:n);
    [m,n] = size(B);
    inputTestSet = B(1:m-1,1:n);
    outputTestSet = B(m,1:n);
    excludedFctr = -1;
    factors = [1:m-1]';
    AUCmax1=0;
    while excludedFctr ~= 0
        excludedFctr = 0;
        [m,n] = size(inputSet);
        AUC = zeros(m,1);
        TPR = zeros(m,1);
        FPR = zeros(m,1);5
            for j = 1:m
            if j == 1
            inputSet1 = inputSet(2:m, 1:898);
                inputTestSet1 = inputTestSet(2:m, 1:299);
            end
            if j == m
            inputSet1 = inputSet(1:m-1, 1:898);
                inputTestSet1 = inputTestSet(1:m-1, 1:299);
            end
            if j > 1 && j < m
            inputSet1 = inputSet([1:j-1, j+1:m], 1:n);
```

```
            inputTestSet1 = inputTestSet([1:j-1, j+1:m], 1:299);
                end
switch checkConst1
    case 1
    PR = minmax(inputSet);
    net = newff(PR, [3 3 1]);
    net.trainFcn = 'traingda';
    net.trainParam.epochs = 1000;
    for averagvle = 1:15
            net = train(net, inputSet, outputSet);
            outputCode = (sim(net, inputTestSet1))';
            outputCode= round ((outputCode+1)/2);
            [MTPR,MFPR,~,MAUC] =
    perfcurve(outputTestSet,outputCode',1);
            TPR(j) = TPR(j) + MTPR(2)/15;
            FPR(j) = FPR(j) + MFPR(2)/15;
            AUC(j) = AUC(j) + MAUC/15;
    end
    clear averagvle;
    case 2
    bHat = glmfit(inputSet1',outputSet','binomial','logit');
    outputCode = glmval(bHat, inputTestSet1', 'logit');
    for i = 1:299
            if outputCode(i,1) < 0.674
            outputCode(i,1) = 0;
            end
            if outputCode(i,1) >= 0.674
            outputCode(i,1) = 1;
            end
```

```
        end
        clear i;
        [MTPR,MFPR,~,AUC(j)] = perfcurve(outputTestSet,outputCode,1);
        TPR(j) = MTPR(2);
        FPR(j) = MFPR(2);
        case 3
        T = fitctree(inputSet1',outputSet);
        outputCode = ((predict( T, inputTestSet1'))+1)/2;
        [MTPR,MFPR,~,AUC(j)] = perfcurve(outputTestSet,outputCode,1);
        TPR(j) = MTPR(2);
        FPR(j) = MFPR(2);
end
clear MTPR;
clear MFPR;
clear MAUC;

        if AUC(j) >= AUCmax1
                AUCmax1 = AUC(j);
                excludedFctr = j;
                maxoutputCode1 = outputCode;
        end
                        end
                result1 = [AUC(:),TPR(:),FPR(:),factors(:)];
                if excludedFctr ~= 0
                inputSet(excludedFctr,:) = [];
                inputTestSet (excludedFctr,:) = [];
                factors(excludedFctr) = [];
                end
        end
```

```
        if AUCmax1 > AUCmax
                result = result1;
                AUCmax = AUCmax1;
                maxoutputCode = maxoutputCode1;
                checkConst = checkConst1;
        end
end
opts.outputDir = 'report';
opts.format = 'pdf';
opts.showCode = false;
publish ('report',opts);
```