Министерство образования и науки Российской Федерации

ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина»

Оценка

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

**СИСТЕМА ПРЕДСКАЗАНИЯ НАЛИЧИЯ ИЗМЕН В БРАКЕ ПО ЕГО ПАРАМЕТРАМ**

Отчет по курсовой работе по дисциплине «Методы анализа Big Data»

Дата Подпись

Преподаватель \_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ Мирвода С. Г.

Студенты гр. РИ-450004 \_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ Рубин О. И.

\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ Шахмартова В. С.

ЕКАТЕРИНБУРГ 2019

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc535457751)

[1 ПОДГОТОВКА ДАННЫХ 4](#_Toc535457752)

[2 ВЫЯВЛЕНИЕ ЗАВИСИМОСТЕЙ И ТРЕНИРОВКА МОДЕЛИ 5](#_Toc535457753)

[3 РАЗРАБОТКА ИНТЕРФЕЙСА СИСТЕМЫ 8](#_Toc535457754)

[ЗАКЛЮЧЕНИЕ 10](#_Toc535457755)

[БИБЛИОГРАФИЧЕСКИЙ СПИСОК 11](#_Toc535457756)

[ПРИЛОЖЕНИЕ А. ИСХОДНЫЙ КОД ПРИЛОЖЕНИЯ НА R 12](#_Toc535457757)

# ВВЕДЕНИЕ

Целью настоящей курсовой работы работы является рассмотрение и применение методов статистики и машинного обучения для анализа набора данных: определение переменной для предсказания, выявление признаков от которых она зависит, подготовка набора данных, избавление от выбросов и неполных данных, подбор модели по которой будет происходить предсказание, тренировка модели на части исходного набора данных данных, проверка качества модели на валидационных данных. В результате выполнения настоящей курсовой работы необходимо получить модель, которая будет выполнять предсказания, а также разработать пользовательский интерфейс, который бы позволил использовать разработанную модель.

# 1 ПОДГОТОВКА ДАННЫХ

В качестве набора данных для исследования был выбран датасет, содержащий информацию о различных параметрах брака (продолжительность, удовлетворенность и т.д.), а также о количестве измен в браке.

Рисунок 1 - Наименование переменных набора данных

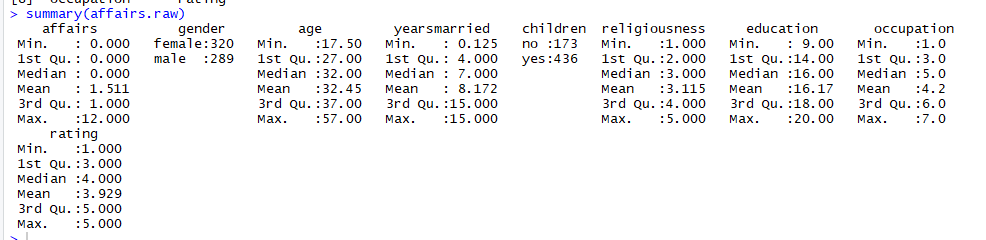


Рисунок 2 — Основные статистические характеристики каждой

В ходе первичного анализа были построенны диаграммы рассеяния, из которых было выявлено, что явных зависимостей между параметрами нет. Далее была выполнена попытка перевести задачу к разряду классификации, разделив все элементы выборки (браки) на те, в которых были измены («unfair») и те, в которых не было измен («fair»). В качестве метода для анализа данных и построения предсказательной модели необходимо использовать классификацию. Для этого необходимо очистить данные от выбросов, выбрать метрику, алгоритм классификации и натренировать модель.

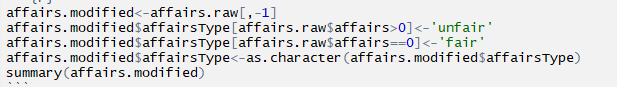


Рисунок 3 – Разделение на классы «fair» и «unfair»

# 2 ВЫЯВЛЕНИЕ ЗАВИСИМОСТЕЙ И ТРЕНИРОВКА МОДЕЛИ

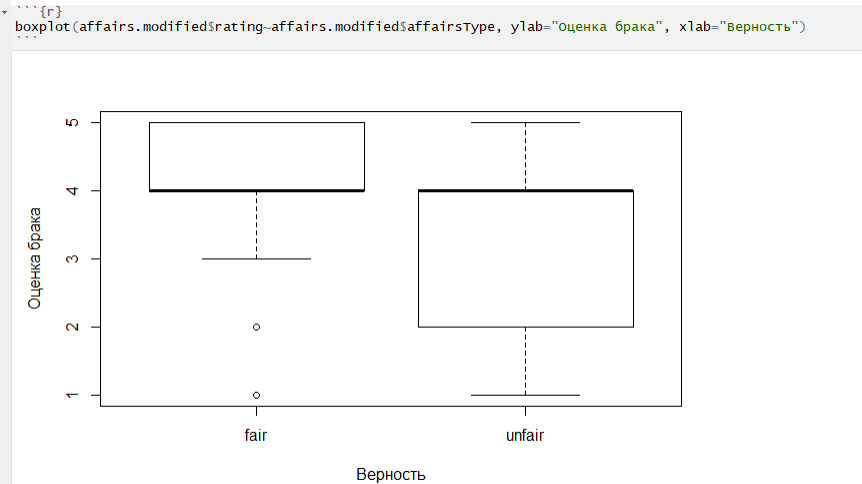
1. В ходе проверки зависимостей посредством функции «boxplot» было выявлено, что существует зависимость между наличием измен в браке и уровнем удовлетворенности им.
2. 
3. Рисунок 4 – Зависимость между оценкой брака и наличием измен
4. На рисунке 4 явно видно, что присутствуют выбросы. Необходимо от них избавиться.

Рисунок 5 — Избавление от выбросов

Далее данные были разделены на два множества – обучающее (для тренировки модели) и валидационное (для валидации точности предсказаний).

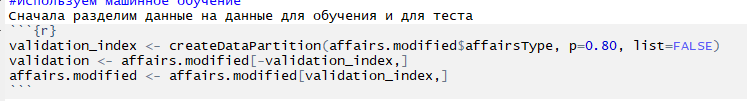


Рисунок 6 - Выбор данных для тренировки модели

Далее был выбран наиболее подходящий алгоритм для тренировки модели, результат тренировки которой будем получать при помощи кроссвалидации и показателя «Accuracy» (см. рис 7). Метрика выбрана так как имеем одинаковое количество экземпляров каждого из трех классов и таким образом оценка, полученная по этой метрике, будет корректна. В качестве алгоритмов для тренировки модели были выбраны LDA, CART, kNN, SVM, randomForest.

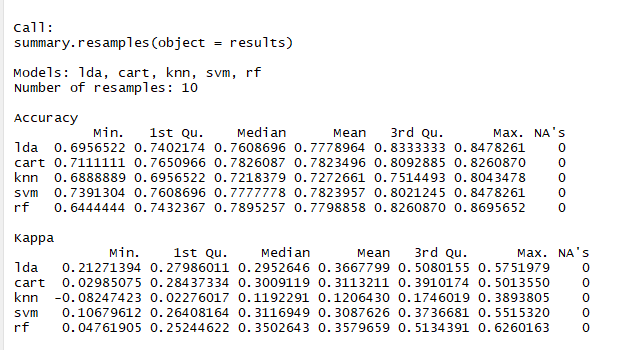


Рисунок 7 – Результаты тренировки модели

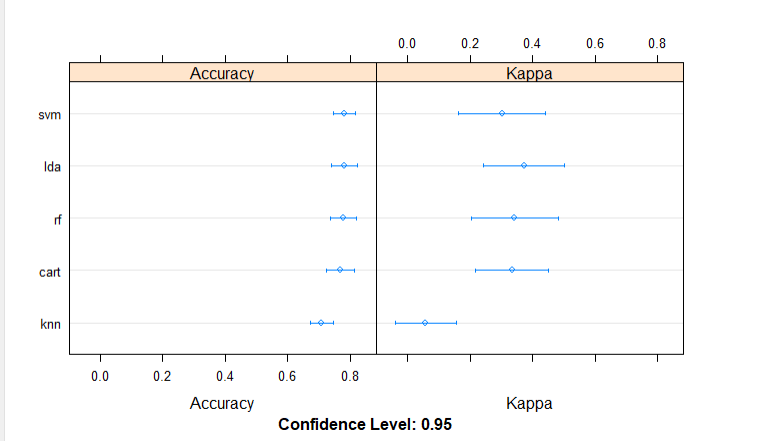


Рисунок 8 — Визуализация метрик «Accuracy» и «Kappa» для алгоритмов

Протестируем работу модели на валидационных данных, попробуем предсказать по ним наличие измен в браке.

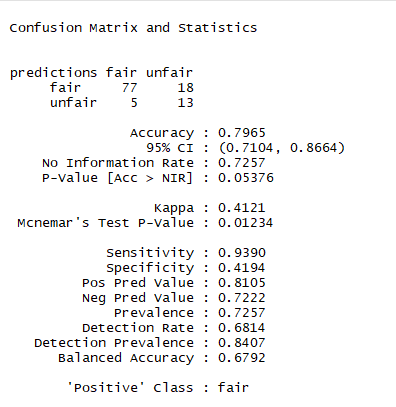


Рисунок 9 — Confusion matrix натренированной модели

# 3 РАЗРАБОТКА ИНТЕРФЕЙСА СИСТЕМЫ

Для разработки приложения использовалась библиотека для создания графических интерфейсов на языке R «gWidgets».

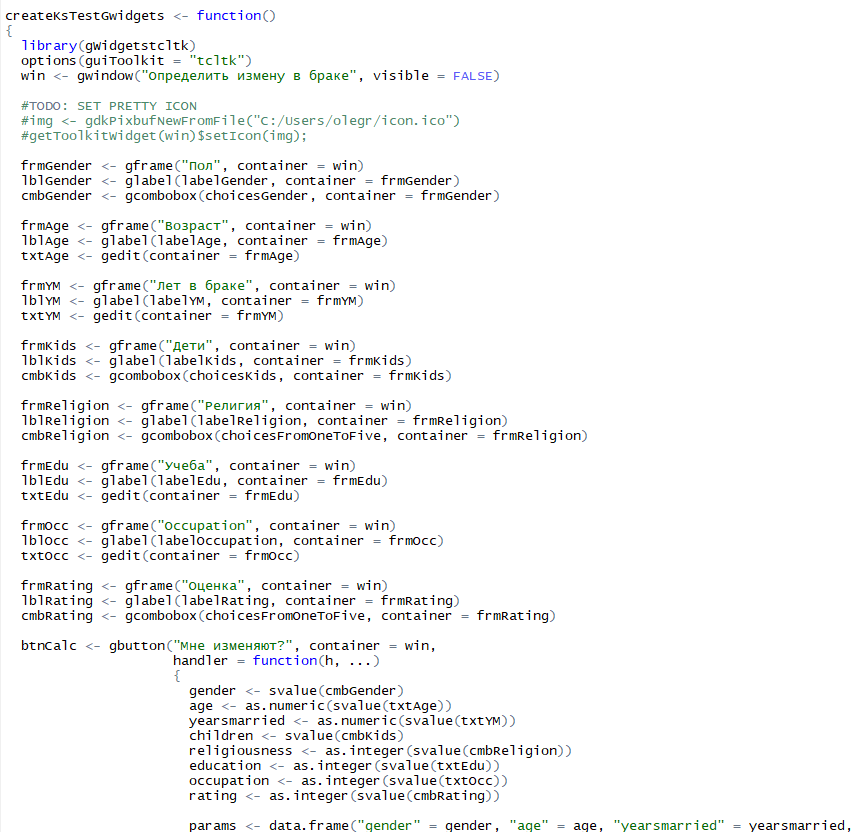
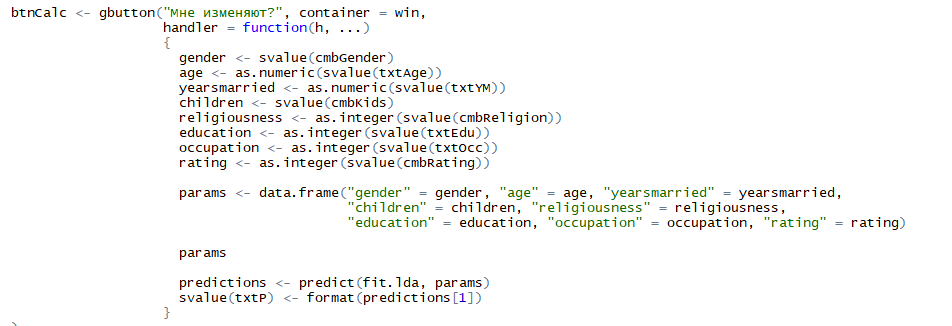
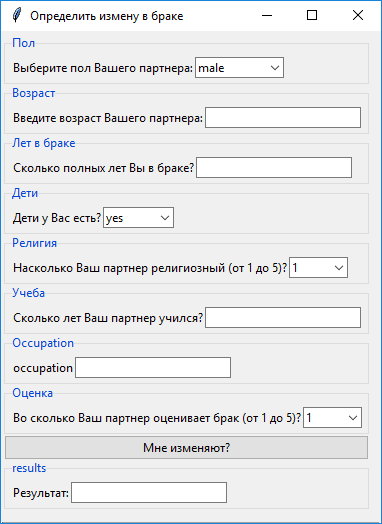


Рисунок 10 — Код создания окна на R

Обработка параметров для дальнейшего предсказания была реализована в обработчике события нажатия на кнопку, являющуюся элементом разработанного интерфейса.

Рисунок 11 – Код обработчика входных параметров

Фактически в итоге было разработано десктопное приложение, способное по введенным на форме в интерфейсе параметрам при нажатии на кнопку угадать, есть ли в браке измена.

Рисунок 12 — Интерфейс приложения

После ввода всех необходимых параметров и нажатия на кнопку для начала расчетов приложение выдаст либо значение «fair», что будет означать что в браке нет измен, либо «unfair», что будет означать обратно, в поле «Результат».

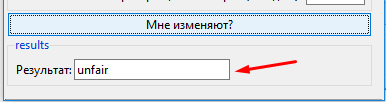


Рисунок 13 — Результат угадывания

# ЗАКЛЮЧЕНИЕ

В результате выполнения настоящей курсовой работы на основании исходного набора данных была получена модель, способная по параметрам брака определять, были в нем измены или нет.

Помимо модели, было разработано простое десктопное приложение с пользовательским интерфейсом для ввода параметров брака и дальнейшего вывода результата предсказания честности брака. Данное приложение может использоваться как супругами для проверки преданности своих партнеров по браку, так и любыми другими людьми, кто, зная чьи-либо данные, может проверить его или ее на верность браку.

# БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Тьюки, Дж. Анализ результатов наблюдений / Дж. Тьюки. – М.: Мир, 1981 – 346 с.

2. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.

3. Сара Бослаф Статистика для всех. / Пер. с англ. П. А. Волкова, И. М. Флямер, М. В. Либерман, А. А. Галицына. – М.: ДМК Пресс, 2015. – 586 с.: ил.

4. Use R in C# [Электронный ресурс]. – Режим доступа: https://coders-corner.net/2015/11/01/use-r-in-c/.

# ПРИЛОЖЕНИЕ А. ИСХОДНЫЙ КОД ПРИЛОЖЕНИЯ НА R

install.packages("caret");

install.packages("e1071");

install.packages("gWidgets");

install.packages("gWidgetstcltk");

#install.packages("RGtk2");

library(caret)

library(gWidgets)

library(gWidgetstcltk)

#require(RGtk2)

affairs.raw<-read.csv('./Affairs.csv')[, -1]

affairs.modified<-affairs.raw[,-1]

affairs.modified$affairsType[affairs.raw$affairs>0]<-'unfair'

affairs.modified$affairsType[affairs.raw$affairs==0]<-'fair'

affairs.modified$affairsType<-as.character(affairs.modified$affairsType)

affairs.modified<-affairs.modified[!(affairs.modified$rating<3&affairs.modified$affairsType=='fair'),]

validation\_index <- createDataPartition(affairs.modified$affairsType, p=0.50, list=FALSE)

validation <- affairs.modified[-validation\_index,]

affairs.modified <- affairs.modified[validation\_index,]

percentage <- prop.table(table(affairs.modified$affairsType)) \* 100

cbind(freq=table(affairs.modified$affairsType), percentage=percentage)

control <- trainControl(method="cv", number=10)

metric <- "Accuracy"

set.seed(13)

fit.lda <- train(affairsType~., data=affairs.modified, method="lda", metric=metric, trControl=control)

validation$affairsType <- NULL

predictions <- predict(fit.lda, validation)

#Some sample data to test against

x1 <- rnorm(100)

x2 <- runif(100)

labelGender <- "Выберите пол Вашего партнера:"

labelAge <- "Введите возраст Вашего партнера:"

labelYM <- "Сколько полных лет Вы в браке?"

labelKids <- "Дети у Вас есть?"

labelReligion <- "Насколько Ваш партнер религиозный (от 1 до 5)?"

labelEdu <- "Сколько лет Ваш партнер учился?"

labelOccupation <- "occupation"

labelRating <- "Во сколько Ваш партнер оценивает брак (от 1 до 5)?"

labelP <- "Результат:"

#Choices for comboboxes

choicesGender <- c("male", "female")

choicesKids <- c("yes", "no")

choicesFromOneToFive <- c("1", "2", "3", "4", "5")

createKsTestGwidgets <- function()

{

library(gWidgetstcltk)

options(guiToolkit = "tcltk")

win <- gwindow("Определить измену в браке", visible = FALSE)

#TODO: SET PRETTY ICON

#img <- gdkPixbufNewFromFile("C:/Users/olegr/icon.ico")

#getToolkitWidget(win)$setIcon(img);

frmGender <- gframe("Пол", container = win)

lblGender <- glabel(labelGender, container = frmGender)

cmbGender <- gcombobox(choicesGender, container = frmGender)

frmAge <- gframe("Возраст", container = win)

lblAge <- glabel(labelAge, container = frmAge)

txtAge <- gedit(container = frmAge)

frmYM <- gframe("Лет в браке", container = win)

lblYM <- glabel(labelYM, container = frmYM)

txtYM <- gedit(container = frmYM)

frmKids <- gframe("Дети", container = win)

lblKids <- glabel(labelKids, container = frmKids)

cmbKids <- gcombobox(choicesKids, container = frmKids)

frmReligion <- gframe("Религия", container = win)

lblReligion <- glabel(labelReligion, container = frmReligion)

cmbReligion <- gcombobox(choicesFromOneToFive, container = frmReligion)

frmEdu <- gframe("Учеба", container = win)

lblEdu <- glabel(labelEdu, container = frmEdu)

txtEdu <- gedit(container = frmEdu)

frmOcc <- gframe("Occupation", container = win)

lblOcc <- glabel(labelOccupation, container = frmOcc)

txtOcc <- gedit(container = frmOcc)

frmRating <- gframe("Оценка", container = win)

lblRating <- glabel(labelRating, container = frmRating)

cmbRating <- gcombobox(choicesFromOneToFive, container = frmRating)

btnCalc <- gbutton("Мне изменяют?", container = win,

handler = function(h, ...)

{

gender <- svalue(cmbGender)

age <- as.numeric(svalue(txtAge))

yearsmarried <- as.numeric(svalue(txtYM))

children <- svalue(cmbKids)

religiousness <- as.integer(svalue(cmbReligion))

education <- as.integer(svalue(txtEdu))

occupation <- as.integer(svalue(txtOcc))

rating <- as.integer(svalue(cmbRating))

params <- data.frame("gender" = gender, "age" = age, "yearsmarried" = yearsmarried,

"children" = children, "religiousness" = religiousness,

"education" = education, "occupation" = occupation, "rating" = rating)

params

predictions <- predict(fit.lda, params)

svalue(txtP) <- format(predictions[1])

}

)

frmResults <- gframe("results", container = win)

lblP <- glabel(labelP, container = frmResults)

txtP <- gedit(container = frmResults)

visible(win) <- TRUE

}

createKsTestGwidgets()