KING KHALID UNIVERSITY
College of Computer Science
Information Sys dept
Date: 16/04/2020

Subject: 632 CIS
Introduction to Data Science
Due Date: 16/04/2020

Project of Introduction to Data Science

Title: Predictions of No-Show Medical Appointments Using Data Mining Techniques

Done by:

Students' name	University id
Hajar Saeed Saad Asiri	441813541
Raneen Abdullah Alghamdi	441813268
Sarah Ali Al Sultan	441813271

Supervised by lecturer:

Mona Moshbab

Predictions of No-Show Medical Appointments Using Data Mining Techniques

1. Introduction

No-show appointments are troublesome for services at all levels of the health care system, described as an appointment in which the patient did not appear for treatment or cancelled on the same day as the appointment. No-shows are a lost profit gain that cannot be recovered for the practice and lead to both diminished patient satisfaction and staff satisfaction [1]. No-show appointments have a negative effect on patients as well as the care teams.

The objective of this project is to study the effect of different factors of patient in predicting whether he/she will attend a medical appointment or not. To achieve this objective, data mining techniques are used to analyse a dataset that contains 110.527 medical appointments. The methodology used in this project consists of four main stages: dataset pre-processing, model building, model training, and model testing and evaluation. In the first stage, some data cleaning and transformation methods are used to prepare the dataset. In the second stage, a logistic regression model is build. In the third stage, the clean dataset is split into two sets: train and test datasets. The train dataset is used to train the logistic model. In the fourth stage, the test dataset is used to test the logistic model. To evaluate the performance of machine learning model, the correct classification accuracy is used.

2. Dataset Description

The dataset consists of 110.527 medical appointments. Each appointment consists of 14 attributes. Table 1 shows the dataset attributes and their descriptions.

3. Exploratory Data Analysis

In this section, some exploratory data analysis techniques are used to explore the distribution of some attributes of the dataset. Also, Statistical tests such as t test and chi-squared test are used to test the significance of some attributes. Also, preparation methods such as attribute transformation and outliers' detection and removing are discussed.

Importing Required Libraries

```
> # import libraries
> library(tidyverse)
> library(lubridate)
> library(caTools)
> # read dataset
```

- **Loading the dataset into R**

```
> data <-  read_csv("dataset.csv")
```

- **Show first five rows of the dataset**

```
> head(data)
```

```
# A tibble: 6 x 14
  PatientId AppointmentID Gender ScheduledDay        AppointmentDay
Age
      <dbl>         <dbl> <chr>  <dttm>              <dttm>
<dbl>
1   2.99e13       5642903 F      2016-04-29 18:38:08 2016-04-29 00:0
0:00    62
2   5.59e14       5642503 M      2016-04-29 16:08:27 2016-04-29 00:0
0:00    56
3   4.26e12       5642549 F      2016-04-29 16:19:04 2016-04-29 00:0
0:00    62
4   8.68e11       5642828 F      2016-04-29 17:29:31 2016-04-29 00:0
0:00     8
5   8.84e12       5642494 F      2016-04-29 16:07:23 2016-04-29 00:0
0:00    56
6   9.60e13       5626772 F      2016-04-27 08:36:51 2016-04-29 00:0
0:00    76
```

Table 1 Attributes and their descriptions

Attribute	Description
PatientId	Identification of patient
AppointmentID	Identification of each appointment
Gender	Male or Female. Female is the greater proportion, woman takes way more care of their health in comparison to man.
DataMarcacaoConsulta	The day of the actual appointment, when they have to visit the doctor.
DataAgendamento	The day someone called or registered the appointment, this is before appointment of course.
Age	How old is the patient.
Neighbourhood	Where the appointment takes place.
Scholarship	True of False .
Hipertension	True or False
Diabetes	True or False
Alcoholism	True or False
Handcap	True or False
SMS_received	1 or more SMS message sent to patient
No-show	True or False

- **Show the structure of the data**

```
> str(data)
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame':     110527 o
bs. of  14 variables:
 $ PatientId     : num  2.99e+13 5.59e+14 4.26e+12 8.68e+11 8.84e+12
...
 $ AppointmentID : num  5642903 5642503 5642549 5642828 5642494 ...
 $ Gender        : chr  "F" "M" "F" "F" ...
 $ ScheduledDay  : POSIXct, format: "2016-04-29 18:38:08" "2016-04-2
9 16:08:27" ...
 $ AppointmentDay: POSIXct, format: "2016-04-29" "2016-04-29" ...
 $ Age           : num  62 56 62 8 56 76 23 39 21 19 ...
 $ Neighbourhood : chr  "JARDIM DA PENHA" "JARDIM DA PENHA" "MATA DA
PRAIA" "PONTAL DE CAMBURI" ...
 $ Scholarship   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Hipertension  : num  1 0 0 0 1 1 0 0 0 0 ...
 $ Diabetes      : num  0 0 0 0 1 0 0 0 0 0 ...
 $ Alcoholism    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Handcap       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SMS_received  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ No-show       : chr  "No" "No" "No" "No" ...
```

```
 - attr(*, "spec")=
  .. cols(
  ..   PatientId = col_double(),
  ..   AppointmentID = col_double(),
  ..   Gender = col_character(),
  ..   ScheduledDay = col_datetime(format = ""),
  ..   AppointmentDay = col_datetime(format = ""),
  ..   Age = col_double(),
  ..   Neighbourhood = col_character(),
  ..   Scholarship = col_double(),
  ..   Hipertension = col_double(),
  ..   Diabetes = col_double(),
  ..   Alcoholism = col_double(),
  ..   Handcap = col_double(),
  ..   SMS_received = col_double(),
  ..   `No-show` = col_character()
  .. )
```

- **Check Missing values in the dataset**

Medical appointments that contain missing values are checked using is.na() R function as follows:

```
> sapply(data,function(x)sum(is.na(x)))
     patient_id  appointment_id          gender    schedule_day appo
intment_day
              0               0               0               0
0
            age    neighborhood     scholarship    hypertension
diabetes
              0               0               0               0
0
     alcoholism         handicap    sms_received         no_show
              0               0               0               0
```

Results show that there are no missing values in the dataset.

- **Show descriptive summary statistics**

The R function summary is used to show some important statistical measures about the attributes of the dataset. These statistical measures include maximum, minimum, mean, and median.

```
> summary(data)
   patient_id         appointment_id    gender     schedule_day
 Min.   :3.922e+04   Min.   :5030230   F:71840   Min.   :2015-11-10
07:13:56
 1st Qu.:4.173e+12   1st Qu.:5640286   M:38687   1st Qu.:2016-04-29
10:27:01
 Median :3.173e+13   Median :5680573             Median :2016-05-10
12:13:17
 Mean   :1.475e+14   Mean   :5675305             Mean   :2016-05-09
07:49:15
 3rd Qu.:9.439e+13   3rd Qu.:5725524             3rd Qu.:2016-05-20
11:18:37
 Max.   :1.000e+15   Max.   :5790484             Max.   :2016-06-08
20:07:23
```

```
 appointment_day                    age                neighborhoo
d   scholarship
 Min.   :2016-04-29 00:00:00   Min.   : -1.00   JARDIM CAMBURI : 771
7   0:99666
 1st Qu.:2016-05-09 00:00:00   1st Qu.: 18.00   MARIA ORTIZ    : 580
5   1:10861
 Median :2016-05-18 00:00:00   Median : 37.00   RESISTÊNCIA    : 443
1
 Mean   :2016-05-19 00:57:50   Mean   : 37.09   JARDIM DA PENHA: 387
7
 3rd Qu.:2016-05-31 00:00:00   3rd Qu.: 55.00   ITARARÉ        : 351
4
 Max.   :2016-06-08 00:00:00   Max.   :115.00   CENTRO         : 333
4
                                                (Other)        :8184
9
 hypertension diabetes   alcoholism handicap   sms_received   no_sho
w
 0:88726      0:102584   0:107167   0:108286   0:75045      Length:1
10527
 1:21801      1:  7943   1:  3360   1:  2042   1:35482      Class :c
haracter
                                    2:   183                Mode  :c
haracter
                                    3:    13
                                    4:     3
```

The statistical summary of the attributes show that the minimum age of in the dataset is -1, maximum age is 115, and the average age of patients is about 37 years. The patient with age -1 is inconsistent and is remove from the dataset. The statistical summary also, shows that female patients in the dataset is about twice male patients.

- **Removing outliers from the dataset**

```
> data <-data[!(data$age<= 0),]
```

- **Converting class label attribute to categorical**

To be able to use the statistical dunctions, visualizations, and the classification model, the attribute NO_SHOW is transformed into categorical attribute.

```
> data$no_show <- as.factor(data$no_show)
```

- **Plotting Patients age versus class label**

To find out if the age of a patient affects his/her attendance in the appointment, the age attribute versus no show attribute is visualized using box plot R function.

Figure 1 show a boxplot of Age vs. No Show. Table 2 shows that the average age of patients who showed up is about 39.1% and the average age of patients who showed up is about 35.3%.

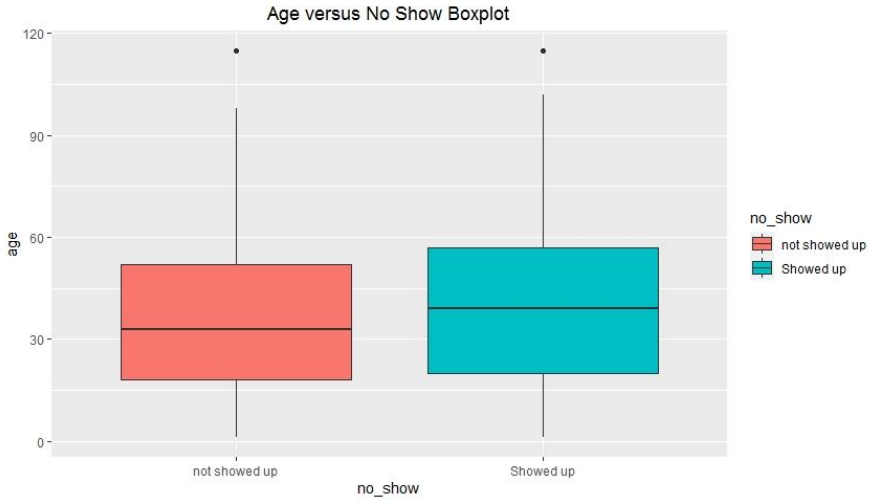

Figure1 Age vs. No Show boxplot

- **Show statistical summary of Age attribute**

Table 2 Patients distribution per age

no_show attribute	age_mean
not_showed_up	35.3%
Showed_up	39.1%

- **Testing the Age attribute**

The t test measure is used to test the relationship between patient's age and the no show attribute as follows:

```
t.test(data$age ~ data$no_show)
Welch Two Sample t-test

data:  data$age by data$no_show
t = -22.682, df = 34965, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.069750 -3.422322
sample estimates:
```

Table 2 T test result of Age Attribute

mean in group not showed up	35.32915%
mean in group Showed up	39.07519%

The small value of P-value in the t test indicates that there is a significant relation between patient age and patient's attendance.

- **Explore the relationship between Gender attribute and No Show**

Figure 2 Gender distribution in the dataset

Figure 2 shows that number of female patients is greater than number of male patients in the dataset.

```
> table(data$gender, data$no_show)
                not showed up Showed up
Female                14275      55843
Male                   7405      29464
```

Table 3 Patients distribution per gender

Gender	Not showed up	Showed up
Female	14275	55843
Male	7405	29464

- **Testing the Gender attribute**

To find out if patient's gender affects his/her attendance at medical appointments or not, the chi-squared test is used.

The result of chi-squared test shows that p value is greater than 0.05. Thus the patient' gender has no significant influence on patient's attendance.

```
chisq.test(table(data$gender,data$no_show))
```

```
        Pearson's Chi-squared test with Yates'
        continuity correction

data:  table(data$gender, data$no_show)
X-squared = 1.1052, df = 1, p-value = 0.2931
```

The result of chi-squared test shows that p value is more than 0.05, so gender difference is not significant.

4. Model Building

The problem in the dataset is to predict if a patient will attend a medical appointment or not. Therefore, the problem is a binary classification problem. The target variable "no show" has two binary values: true, or false.

Therefore, the logistic regression model was selected to predict patients' attendances.

Since our problem is a binary classification problem, binomial logistic model is used in this project.

Other reasons for choosing the logistic regression model are the logistic regression model is a widely used model because it is very efficient, does not require too many computational resources, it's highly interpretable, it doesn't require input features to be scaled, it doesn't require any tuning, it's easy to regularize, and it outputs well-calibrated predicted probabilities.

4.1. Logistic Regression Model

Before building the logistic model, we split the dataset randomly into 70% for training the model and 30% for testing and evaluating the performance of the model as follows:

```
> # Divide the dataset into two  70% train and 30% test sets
> set.seed(100)
> split = sample.split(df$no_show, SplitRatio = 0.70)
> train = subset(df, split == TRUE)
> test = subset(df, split == FALSE)


> # Logistic Regression
> model1 <- glm(formula = no_show ~ . ,  data = train, family =binom
ial(link = 'logit') )
> summary(model1)

Call:
glm(formula = no_show ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1476   0.5322   0.6084   0.6872   1.0323

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.3373609  0.0220912  60.538  < 2e-16 ***
age            0.0075069  0.0004903  15.311  < 2e-16 ***
genderMale     0.0202746  0.0198013   1.024   0.3059
scholarship1  -0.1799873  0.0293662  -6.129 8.84e-10 ***
hypertension1  0.0320405  0.0293627   1.091   0.2752
diabetes1     -0.0520383  0.0409388  -1.271   0.2037
alcoholism1   -0.1296159  0.0536699  -2.415   0.0157 *
handicap1      0.0004409  0.0712021   0.006   0.9951
handicap2      0.0285616  0.2326642   0.123   0.9023
handicap3     -0.3580161  0.8124158  -0.441   0.6594
handicap4     -1.1189538  1.4361582  -0.779   0.4359
sms_received1 -0.6372546  0.0187534 -33.981  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75496  on 74890  degrees of freedom
Residual deviance: 73973  on 74879  degrees of freedom
AIC: 73997

Number of Fisher Scoring iterations: 4
```

The summary of the logistic model shows the attributes with important parameters such as p-value. The P-value of an attribute can be used to determine whether this attribute is significant or not. Since the attributes gender, hypertension, diabetes, handicap1, Handicap2, handicap3, and handicap4 have p-values less than 0.05, therefore these attribute are not significant in the prediction of patients' attendance at medical appointments. Results also, show that the age and sms_reveived attributes are significant in the prediction of patients' attendance at medical appointments since they have p-values less than 0.05.

5. Model Evaluation

To evaluate the performance of the model, the model is tested using the test dataset. The estimated results are compared with the actual target values, and correct classification accuracy is calculated as follows:

```
> predicted <- predict(model1,newdata=test,type='response')
> pred_test <-  ifelse(predicted>0.5,1,0)
> tab <- table(predicted = pred_test, actual = test$no_show)
> tab
         actual
predicted not showed up Showed up
        1          6504     25592
```

Results show that the logistic model achieved a correct classification accuracy of about 80% on the test dataset.

```
                actual
predicted not showed up Showed up
        1       20.26421  79.73579
```

Figure 1 Confusion Matrix of Logistic Regression Model

6. Conclusions

Deciding the chances that a patient will 'no-show' an appointment will bring major financial and organizational benefits to health care providers. Practices that consistently recognize and work with patients to eliminate no-shows help patients resolve treatment challenges and provide positive clinical outcomes for patients.

In this project, data mining techniques are applied to predict the probability of a patient showing a medical appointment. First, the dataset was loaded into R. Then Data

cleaning tasks such as outliers removal and transformations of attributes' data type. Then, exploratory and visualization data analysis techniques as well as statistical tests such as t test and chi-squared test were applied on the age and gender attributes. Finally, the clean dataset was split into train and test sets. The train set was used to train a logistic regression model. The test set was used to evaluate the prediction performance of the mode. Experimental results show that the model achieved a prediction accuracy of about 80%.

References

[1] Anderson, R. T., Camacho, F. T., & Balkrishnan, R. (2007). Willing to Wait?: The influence of patient wait time on satisfaction with primary care. BMC Health Services Research

[2] Medical Appointment No Shows, Why do 30% of patients miss their scheduled appointments?, https://www.kaggle.com/joniarroba/noshowappointments

[3] Nina Zumel and John Mount, "Practical Data Science with R SECOND EDITION", Manning Publications, November 2019.

[4] https://rstudio.com/

[5] https://www.r-project.org/about.html