**Assignment-5**

1-a) What metrics do you think are informative for me to understand from the data? How do you determine which metrics are important?

Answer)The metrics which I think are informative to user are as follows:-

Steps(count) on hourly basis and daily basis are the metrics that are important and informative. From the data set metric we can also define some other inferred metrics which are as follows

1)Sum of count of steps in each of the hour during the entire data set

2)Sum of count of steps in each of the weekday during the entire data set

3)Average of the count of steps that are taken in each of the hour during entire data set

4) Average of the count of steps that are taken in each of the weekday during entire data set

For a data to be metric, it needs to satisfy the criteria of:

\*refernce for below discussed topic: http://blog.trak.io/a-beginners-introduction-to-metrics-analytics-for-data-driven-growth/\*

Accessible: If everyone in your team can’t understand what the metric means, then it’s useless. People need to be able to discuss it, remember it, apply it to their daily work. This doesn’t just mean having simple, human-readable names for metrics, but also a clearly understandable explanation of how that metric is calculated and what it means in the context of your product and users.

Actionable: A metric is actionable if you can make a decision based on it’s result, and go and do something about it. The opposite of an actionable metric is a vanity metric. Actionable is important as a metric should motivate us to make a decision based on it. In our case, steps help us conclude the walking done over a period

Comparable: If you can’t compare the metric to another date period, or to a different version of the product, then you have no frame of reference.

Rates or Ratios: Counting absolute numbers can be useful, but monitoring the rate of change of that number, or the ratio of that number to another, is more valuable. Knowing the current speed of a car is OK, but knowing if it's accelerating or braking is far more useful.

1-b) For each of the Daily and Hourly data sets, what form of graph and graphical objects are appropriate? Why? (5 points)

Answer)

For **daily data** analysis, I am choosing a line chart to show count of steps for each day. For plotting the graph I am choosing X axis as Date and Y axis as Number of steps.

Line Chart will show a good comparative visualization of counts of steps across various days. The odds that a day will be go without walking is pretty less. It is expected that a person will definitely walk and take atleast few steps each day. An outlier may be a day when excessive walking is done.

For **hourly data** analysis, I am using Bar Graph. Reason for choosing the bar graph is as follows:

|  |  |  |
| --- | --- | --- |
| Graph Type | Type/Description(Graph Objectives) | Encoding Methods |
| Bar graph | **Frequency Distribution**  It Counts of something per categorical subdivisions (intervals) of a quantitative range | I have used Vertical bars to emphasize individual values |

As we can see that we can easily do Frequency Distribution of the step counts(eg: Task 1c where I have plot the step counts distribution for 24 hour)by taking the subdivisions(intervals) of 24 hours of a day and then we can efficiently plot that to get some meaningful information.

1-c) Generate two visualizations, keeping in mind principles we discussed in class. (12 points)?

Answer)

**First visualization technique**

Here, I’m calculating the most active hour of a person for an entire data set. So for that I’ve sum up the step counts for a particular hour for the entire hourly data set and that procedure I follow to all 24 hours.Then I have taken the average for all the sum values by dividing the number of times that hour duration appear in the data set

Example(Let’s say we are calculating how many steps the person has taken between 1:00 pm and 2:00 pm for the entire dataset ,then I’m sum up the step count values where I find the field1 value 1:00 pm to 2:00 pm in the dataset. After I got the sum values I’m dividing that value to number of times the duration 1:00pm to 2:00pm appear on the dataset)

Code:

dtparts<-NULL

#Read hoursly data file in next step

hrdata<-read.csv(file.choose(),sep=",",header=T)

#Read daily data file in next step

dailydata<-read.csv(file.choose(),sep=",",header=T)

dtparts = t(as.data.frame(strsplit(as.character(hrdata$Start),' ')))

dtparts<-cbind(dtparts,as.data.frame(hrdata$Steps..count.))

dtparts1<-as.data.frame(dtparts)

#dtparts1$V3<-as.numeric(dtparts1$V3)

X <- vector(mode="numeric", length=24)

cnt<-vector(mode="numeric",length=24)

hours<-seq(from=1,to=24,by=1)

for(i in 1:length(dtparts1[,2]))

{

if(dtparts1[i,2]=="0:00")

{

X[1]=X[1]+dtparts1[i,3]

cnt[1]=cnt[1]+1

}

else if(dtparts1[i,2]=="1:00")

{

X[2]=X[2]+dtparts1[i,3]

cnt[2]=cnt[2]+1

}

else if(dtparts1[i,2]=="2:00")

{

X[3]=X[3]+dtparts1[i,3]

cnt[3]=cnt[3]+1

}

else if(dtparts1[i,2]=="3:00")

{

X[4]=X[4]+dtparts1[i,3]

cnt[4]=cnt[4]+1

}

else if(dtparts1[i,2]=="4:00")

{

X[5]=X[5]+dtparts1[i,3]

cnt[5]=cnt[5]+1

}

else if(dtparts1[i,2]=="5:00")

{

X[6]=X[6]+dtparts1[i,3]

cnt[6]=cnt[6]+1

}

else if(dtparts1[i,2]=="6:00")

{

X[7]=X[7]+dtparts1[i,3]

cnt[7]=cnt[7]+1

}

else if(dtparts1[i,2]=="7:00")

{

X[8]=X[8]+dtparts1[i,3]

cnt[8]=cnt[8]+1

}

else if(dtparts1[i,2]=="8:00")

{

X[9]=X[9]+dtparts1[i,3]

cnt[9]=cnt[9]+1

}

else if(dtparts1[i,2]=="9:00")

{

X[10]=X[10]+dtparts1[i,3]

cnt[10]=cnt[10]+1

}

else if(dtparts1[i,2]=="10:00")

{

X[11]=X[11]+dtparts1[i,3]

cnt[11]=cnt[11]+1

}

else if(dtparts1[i,2]=="11:00")

{

X[12]=X[12]+dtparts1[i,3]

cnt[12]=cnt[12]+1

}

else if(dtparts1[i,2]=="12:00")

{

X[13]=X[13]+dtparts1[i,3]

cnt[13]=cnt[13]+1

}

else if(dtparts1[i,2]=="13:00")

{

X[14]=X[14]+dtparts1[i,3]

cnt[14]=cnt[14]+1

}

else if(dtparts1[i,2]=="14:00")

{

X[15]=X[15]+dtparts1[i,3]

cnt[15]=cnt[15]+1

}

else if(dtparts1[i,2]=="15:00")

{

X[16]=X[16]+dtparts1[i,3]

cnt[16]=cnt[16]+1

}

else if(dtparts1[i,2]=="16:00")

{

X[17]=X[17]+dtparts1[i,3]

cnt[17]=cnt[17]+1

}

else if(dtparts1[i,2]=="17:00")

{

X[18]=X[18]+dtparts1[i,3]

cnt[18]=cnt[18]+1

}

else if(dtparts1[i,2]=="18:00")

{

X[19]=X[19]+dtparts1[i,3]

cnt[19]=cnt[19]+1

}

else if(dtparts1[i,2]=="19:00")

{

X[20]=X[20]+dtparts1[i,3]

cnt[20]=cnt[20]+1

}

else if(dtparts1[i,2]=="20:00")

{

X[21]=X[21]+dtparts1[i,3]

cnt[21]=cnt[21]+1

}

else if(dtparts1[i,2]=="21:00")

{

X[22]=X[22]+dtparts1[i,3]

cnt[22]=cnt[22]+1

}

else if(dtparts1[i,2]=="22:00")

{

X[23]=X[23]+dtparts1[i,3]

cnt[23]=cnt[23]+1

}

else if(dtparts1[i,2]=="23:00")

{

X[24]=X[24]+dtparts1[i,3]

cnt[24]=cnt[24]+1

}

}

X

cnt

average<-X/cnt

#write.csv(dtparts1,"dtparts1.csv")

barplot(average, main="stpes in each hour in entire year",names.arg=seq(from=0,to=23,by=1),axes=F)

usr <- par("usr")

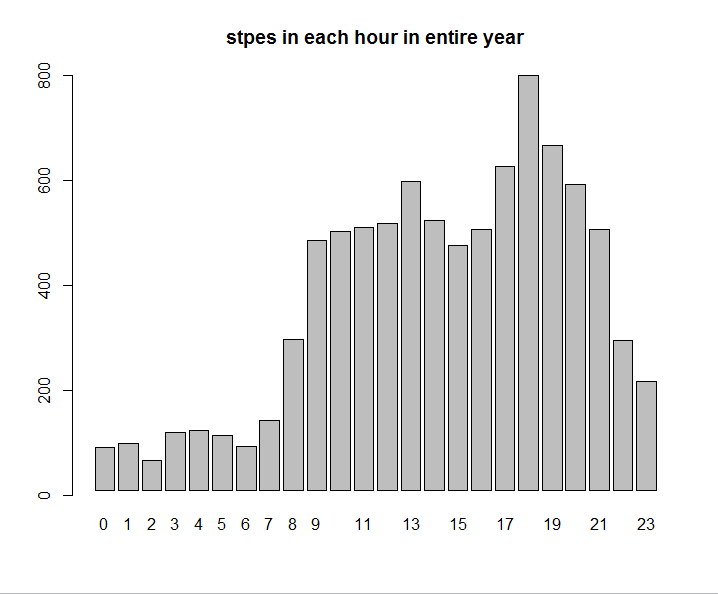
par(usr=c(usr[1:2], 0, 800))

axis(2,at=seq(0,800,200))

# [1] 35155.20 38051.20 24575.42 47403.98 48784.40 44432.64 35440.36 57195.31 122364.09 202320.00 209415.84 212482.27 215277.69 249352.57

#[15] 218292.71 197909.51 211138.79 261387.67 334062.35 278380.53 246803.94 210807.80 121333.44 88227.31

**Output:**



Conclusion:

By seeing the graph we can say that between 18:00 Hr and 19:00 Hr,the person is most active(has taken most steps) during the entire data set.

**Second visualization technique**

Here, I’m calculating the most active weekday of a person for an entire data set. So for that I’ve sum up the step counts for a particular weekday for the entire daily data set and that procedure I follow to all 7 weekdays. Then I have taken the average for all the sum values by dividing them to the number of times that weekday appear in the data set

Example(Let’s say we are calculating how many steps the person has taken on Monday for the entire dataset ,then I’m sum up the step count values where I find the field3 value Monday in the dataframe dtparts3. After I got the sum values I’m dividing that value by number of times Monday appear on the dataframe dtparts3)

Code)

#dailydata dataframe has already decalre and initialize in visulization technique 1

dtparts2<-NULL

dtparts2 = t(as.data.frame(strsplit(as.character(dailydata$Start),' ')))

dtparts2<-cbind(dtparts2,as.data.frame(dailydata$Steps..count.))

dtparts3<-as.data.frame(dtparts2)

#dtparts3$V3<-as.numeric(dtparts3$V3)

install.packages("chron")

library(chron)

thetimes1<-chron(dates=as.character(dtparts3[,1]),format=c(dates="m/d/y"))

thetimes1<-as.data.frame(thetimes1)

thetimes1$day <- weekdays(as.Date(thetimes1$thetimes1))

dtparts3<-cbind(dtparts3,as.data.frame(thetimes1$day))

X1 <- vector(mode="numeric", length=7)

cnt1<-vector(mode="numeric",length=7)

for(i in 1:length(dtparts3[,1]))

{

if(dtparts3[i,4]=="Monday")

{

X1[1]=X1[1]+dtparts3[i,3]

cnt1[1]=cnt1[1]+1

}

else if(dtparts3[i,4]=="Tuesday")

{

X1[2]=X1[2]+dtparts3[i,3]

cnt1[2]=cnt1[2]+1

}

else if(dtparts3[i,4]=="Wednesday")

{

X1[3]=X1[3]+dtparts3[i,3]

cnt1[3]=cnt1[3]+1

}

else if(dtparts3[i,4]=="Thursday")

{

X1[4]=X1[4]+dtparts3[i,3]

cnt1[4]=cnt1[4]+1

}

else if(dtparts3[i,4]=="Friday")

{

X1[5]=X1[5]+dtparts3[i,3]

cnt1[5]=cnt1[5]+1

}

else if(dtparts3[i,4]=="Saturday")

{

X1[6]=X1[6]+dtparts3[i,3]

cnt1[6]=cnt1[6]+1

}

else if(dtparts3[i,4]=="Sunday")

{

X1[7]=X1[7]+dtparts3[i,3]

cnt1[7]=cnt1[7]+1

}

}

X1

cnt1

average1<-X1/cnt1

#write.csv(dtparts3,"dtparts3.csv")

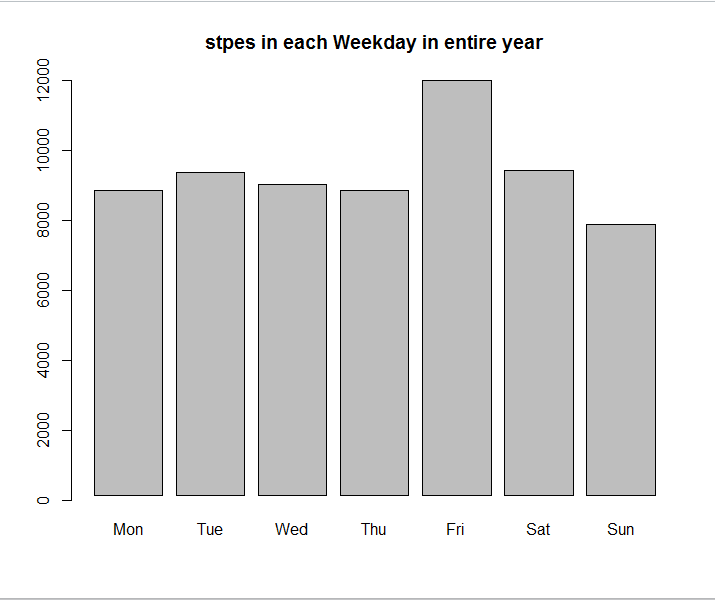
barplot(average1, main="stpes in each Weekday in entire year",names.arg=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"),axes=F)

usr <- par("usr")

par(usr=c(usr[1:2], 0, 12000))

axis(2,at=seq(0,12000,2000))

Output



Conclusion

By seeing the graph we can say that on Friday, the person is most active(has taken most steps) during the entire data set.

Q-2)

Q-a)If you have a classification model that outputs predicted probabilities, how could you convert those probabilities to class predictions?

Answer

As we have model the computes predicted probabilities, we can make a graph by taking x axes as

that predicted probabilities. After generating the graph we can set a Threshold and can classify

everything above that threshold as correct and everything below that prediction as wrong

Q-b)Why are predicted probabilities (rather than just class predictions) required to

generate an ROC curve?

Answer

Our motivation to use ROC curve is to study the performance of classifier at various threshold levels. Hence, if we have class predictions, we can just have a certain values and can't have thresholds. Predictive probabilities facilitate the thresholds. So, to fix the threshold values(if we don't have predictive probabilities then we don’t have any accurate base to fix the threshold values),we require predicted probabilities as to measure the performance of classifier at all possible threshold values.

Q-c)Could you use an ROC curve for a regression problem? Why or why not?

Answer

No

This is because a regression problem is not a classification problem. In other words, ROC curve is a plot between True Positives & False Positives. Such fields can not be obtained from a regression problem.

Q-d)What's another term for True Positive Rate?

Answer

Recall or Sensitivity

Q-e) If I wanted to increase specificity, how would I change the classification threshold?

Answer

We can increase specificity by increasing the value of threshold till we get specificity(true negative rate) equal to 1(ideal case)

Q-f) Is it possible to adjust your classification threshold such that both sensitivity and specificity increase simultaneously? Why or why not?

Answer

No,we can not increase both sensitivity and specificity simultaneously. As they are inversely related. What I mean by this is

1)if we increase the threshold value,we are increasing the proportion of negative and decreasing

the proportion of positives. And by going through the formula of

(specificity= measures the proportion of negatives that are correctly identified as such)

(sensitivity-measures the proportion of positives that are correctly identified as such)

if we increase the threshold value, we are increasing specificity and decreasing sensitivity

2)if we decrease the threshold value, we are decreasing the proportion of negative and increasing

the proportion of positives. And by going through the formula of

(specificity= measures the proportion of negatives that are correctly identified as such)

(sensitivity-measures the proportion of positives that are correctly identified as such)

if we decrease the threshold value , we are decreasing specificity and increasing sensitivity

Q-g)What are the primary benefits of ROC curves over classification accuracy?

Answer

It allows us to visualize the performance of our classifier and it works well for unbalanced classes. And also it doesn't require to set the classification threshold

Q-h)What should you do if you have a low AUC value like 0.15?

Answer

If AUC value is low then our classifier model is not that good. We can get the higher AUC values if we interchange our prediction(yes to no and no to yes).By doing that we can get AUC value (1-0.15=0.85) for this specific scenario. Alternatively , having more predictors or less predictors than what is required in analysis may lead to a low AUC.Hence, consider changing the set of predictors.

Q-i)What's a real-world scenario in which you would prefer high specificity (rather than high sensitivity) for your classifier?

Answer

Consider a doctor is testing a patient for the presence of a disease. This particular disease is treatable, but the treatment has very serious side(life threatening) effects. In this case doctor wants a test that has high specificity, because there are major drawbacks to a false positive

(Reference-http://www.stomponstep1.com/sensitivity-specificity-screening-tests/)