Prooject4

Tahmid Alam (101008957)

12/14/2021

Background

Keystroke dynamics means the analysis of typing rhythms to discriminate among users. Dr. Roy Maxian and colleagues recruited 51 subjects at CMU who have typed a passcode for a specific system. Subjects completed 8 data-collection sessions (of 50 passwords each), for a total of 400 password-typing samples. They waited at least one day between sessions, to capture some of the day-to-day variation of each subject's typing. In this research they collected a keystroke dynamics data set, developed an evaluation procedure, and measured the performance of a range of anomaly-detection algorithms so that the results can be compared on an equal basis.

Objective

Our objective is to evaluate whether a user is consistent over time in how they type a given passcode. Also, we compare total time for password input between random users to observe if there is any significance difference. We develop a formal and appropriate statistical analysis of the data set.

Data

We have been provided with the typing data from 51 subjects, each typing 400 repetitions of a password in 8 sessions. There are 31 various timing features used by researchers were extracted from the raw data. We can classify the variable into 3 classes, the keydown-keydown times and hold times, keydown-keyup times. During Analysis, we will be refereeing these variables types as- Keydown-Keydown Time -> DD.Time Keydown-Keyup Time -> UD.Time Keyhold Times -> Hold.Time We summed up the timing variables for each repetition during password input, this variable will be referred as , Total Time -> TT.Time

Algorithm

1. Exploratory analysis: We performed different EDA analysis i.e., Total Time vs Repetition, Total Time vs Session Number to examine the trend of consistency over time for a user. We also plotted boxplot, histograms, density plots to find out if there is significant difference among randomly selected different subjects' password input times.

2. Developing Statistical Model: We developed a linear regression model based on individual subject's total time needed for every session to predict the total time needed for future sessions.

3. Model Analysis: To examine the model's accuracy, we split the dataset into an 80:20 sample (training:test), then, build the model on the 80% sample and then used the model thus built to predict the dependent variable on test data. To study the appropriateness of the model, we plotted diagnostics plots for the linear regression.

Exploratory analysis

First, we will be observing whether a user is consistent over time in how they type a given passcode by analyzing total time for per repetition and per session. Also, we will be applying data visualization such as BoxPlots, Histograms, Density Plots to find out if there is any difference between session total times of subject 57 Then to compare the mean of two sessions, we will be doing a unpaired two-samples t-test .

From the plots, we can see that the trend of the Total Time vs Repetition Number is decreasing. This means, with each repetition, the user get used to keystrokes of the password, hence reducing the total time.

From our boxplot, histogram, density, we can see that the boxes of the boxplots does not overlap with each other. So, there is difference between the three sessions of subject57.

We performed two tailed t.test to compare between two session mean total time. From the t test, The p-value of the test is 2.2e-16, which is less than the significance level alpha = 0.05 hence that the average total time for session1 and session8 of subject57 is significantly different from each other with a p-value = 2.2e-16.

Next, we will be observing the total time average session wise of different random subjects and how are they related to each other..

Also, we will be applying data visualization such as BoxPlots, Histograms, Density Plots to find out if there is any difference between the groups. Then to compare the mean of two sessions, we will be doing a unpaired two-samples t-test .

From our boxplot, histogram, density, we can see that the boxes of the boxplots does not overlap with each other. So, there is difference between subjects.

Developing Statistical Model

Linear regression is used to predict the value of an outcome variable Y based on one or more input predictor variables X. The aim is to establish a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, we can use this formula to estimate the value of the response Y, when only the predictors (Xs) values are known. Here, we Developed a linear regression model for subject5. We took the average total time for password input for per session and the sessionIndex as the response variable.

We used Scatter plots to help us visualize any linear relationships between the dependent (response) variable and independent (predictor) variables. The scatter plot along with the smoothing line above suggests a linearly decreasing relationship between the variables.

The function used for building linear models is lm(). The lm() function takes in two main arguments, namely: 1. Formula 2. Data. The data is typically a data.frame and the formula is a object of class formula.

After developing the model we can obtain the summary statistics.

Summary of Linear Regression

STATISTIC	Value	CRITERION
R-Squared	0.87	Higher the better (> 0.70)
Adj R-Squared	0.85	Higher the better
F-Statistic	41.24	Higher the better
Std. Error	0.9424	Closer to zero the better
t-statistic	-6.422	Should be greater 1.96 for p-value to be less than 0.05
AIC	25.46	Lower the better
BIC	25.69	Lower the better

Here, the p-value is 0.000673 for TT.Time, We can consider a linear model to be statistically significant only when both these p-Values are less that the pre-determined statistical significance level, which is ideally 0.05.

R-Squared with 0.8518 tells us is the proportion of variation in the dependent (response) variable that has been explained by this model

F-statistic are measures of goodness of fit. with a F-statistics of 41.24, we can say its a fairly good model.

AIC and BIC is reletively low value which is a indication of good model.

To examine the model's accuracy, we split the dataset into a 80:20 sample (training:test), then, build the model on the 80% sample and then use the model thus built to predict the dependent variable on test data.

Doing it this way, we have the model predicted values for the 20% data (test) as well as the actuals (from the original dataset). By calculating accuracy measures (like min_max accuracy) and error rates (MAPE or MSE), we foundout the prediction accuracy of the model.

STATISTIC	Value	CRITERION
MAPE	8.114901	Lower the better
Min_Max Accuracy	92.09895	Higher the better

Conclusion

From our statistical Analysis, we can see that total time for one subject to put in the keystroke of designated password gradually changes with repetition and session. We have demonstrated the difference with graphical analysis such as plots, boxplot, histogram, density plots. Also, we performed a two tailed t.test to determine if there is a significant difference between the means of two groups. Later, we demonstrated that there are total time differences for every subjects as the vary in mean total time. Lastly, we developed a Linear Regression Model to predict the value of a dependent variable which is Total Time based on an independent variable. The greater the linear relationship between the independent variable and the dependent variable, the more accurate is the prediction. We Also developed a multiple linear regression with total time as a response variable with DD time, UD Time and Hold time as predictor variables.

References

https://www.cs.cmu.edu/~keystroke/KillourhyMaxion09.pdf
https://www.cs.cmu.edu/~keystroke/
https://github.com/RoyMaxion/RoyMaxion.github.io/blob/master/projects/keystroke-benchmark/evaluation-script.R http://r-statistics.co/Linear-Regression.html
http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/ https://stackoverflow.com/questions/65124061/confusion-matrix-for-a-logistic-model
https://github.com/cran/sparklyr/blob/c0effdbed11c95e42ea37193b1cfe2516217516b/R/ml_classification_logistic_regression.R https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 https://www.statmethods.net/graphs/density.html
http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r
https://data.library.virginia.edu/diagnostic-plots/
https://statisticsbyjim.com/regression/choosing-regression-analysis/
https://www.investopedia.com/terms/m/mlr.asp
https://www.northeastern.edu/graduate/blog/statistical-modeling-for-data-analysis/