Machine Learning with R: Logistic Regression

Shaila Jamal

Data Analysis Support Assistant, DASH, McMaster Library

Ph.D. Candidate, School of Earth, Environment and Society, McMaster University

November 03, 2022

McMaster University | DASH

McMaster University | Library

McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences.

In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Code of Conduct

The DASH program and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:

scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Completion that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

If you would like to be considered for the certificate, verify your participation in this form: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

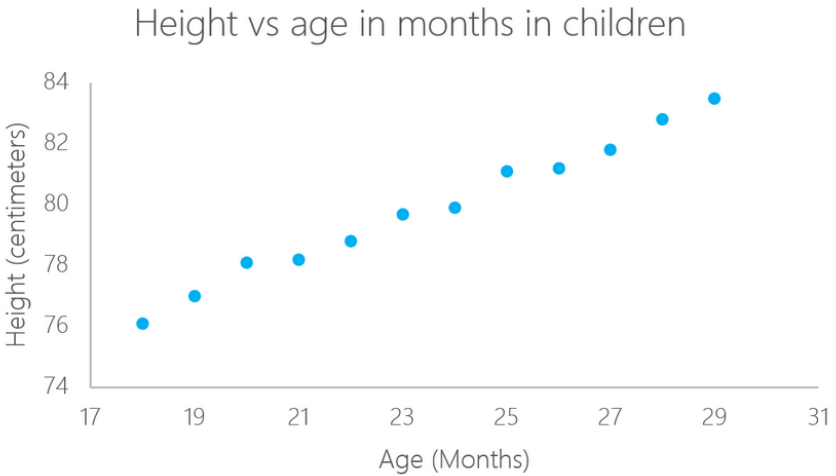
Recordings of the workshops

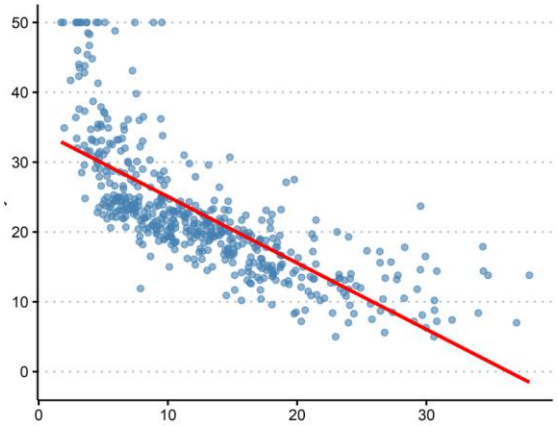
For workshop recordings, check here: Search the Online Learning Catalogue | Sherman Centre for Digital Scholarship (scds.ca)

Linear Regression

Linear relationships mean that “you can fit a line between the two (or more variables)”

- Take the example of the age and height we discussed before.





Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Logistics Regression

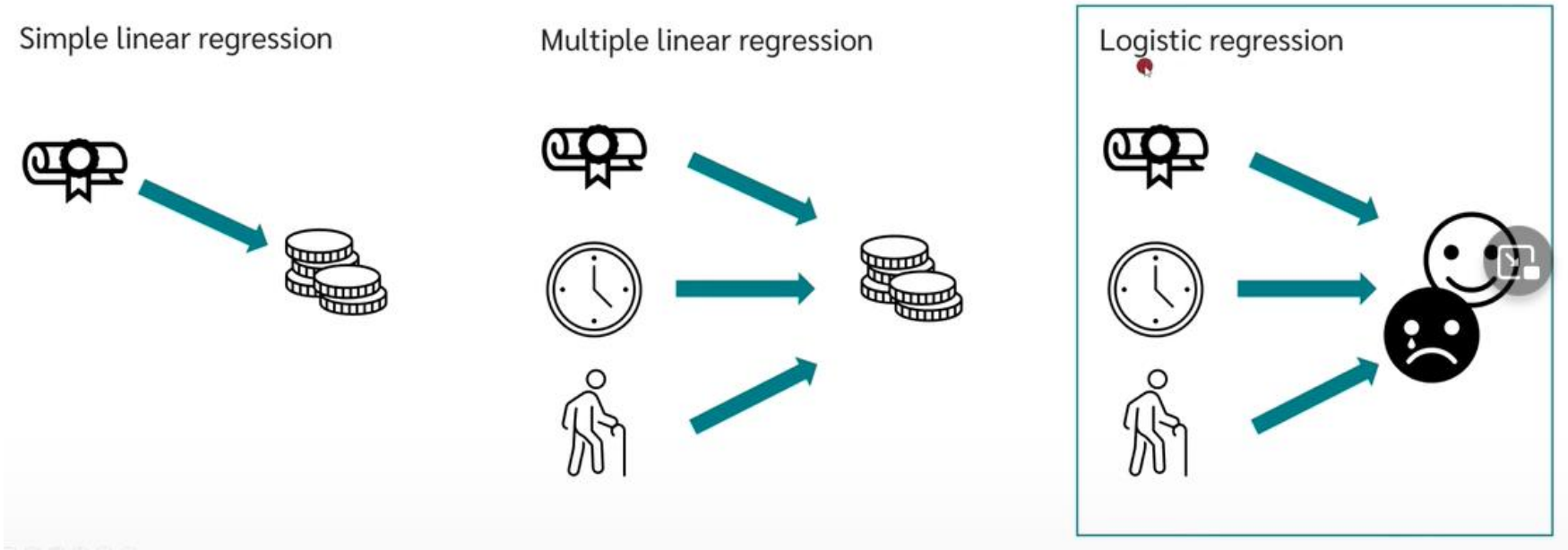
“Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables”.

- “A binary outcome is one where there are only two possible scenarios—either the event happens (1) or it does not happen (0)- such as “yes” or “no”, “pass” or “fail”, and so on.”

- However, the independent variables can fall into any of the categories:
  - Continuous – the temperature is degrees or weights in grams
  - Discrete, ordinal - some kind of order on a scale. Example: customer satisfaction level.
  - Discrete, nominal - fits into named groups that do not represent any kind of order or scale or hierarchy. Example, color of eyes: Blue, Brown, Green, Black.

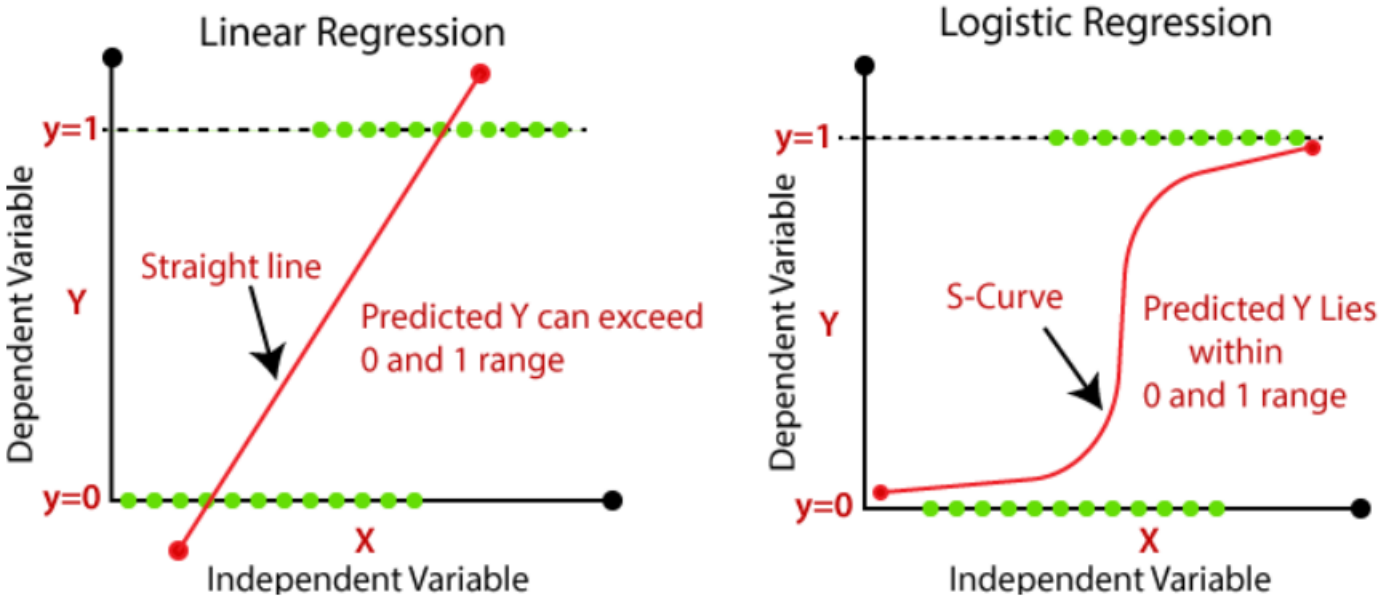
Source: Thanda, A. “What is Logistic Regression? A Beginner's Guide”. October 04, 2022, accessed on October 29, 2022. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

Linear vs logistic Regression



Source: Datatab. “Logistic Regression: An Introduction”. February 05, 2021, accessed on October 29, 2022. https://www.youtube.com/watch?v=3tq4t41MsPc/

Linear vs logistic Regression



Source: “Linear Regression vs Logistic Regression”. accessed on October 29, 2022. https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

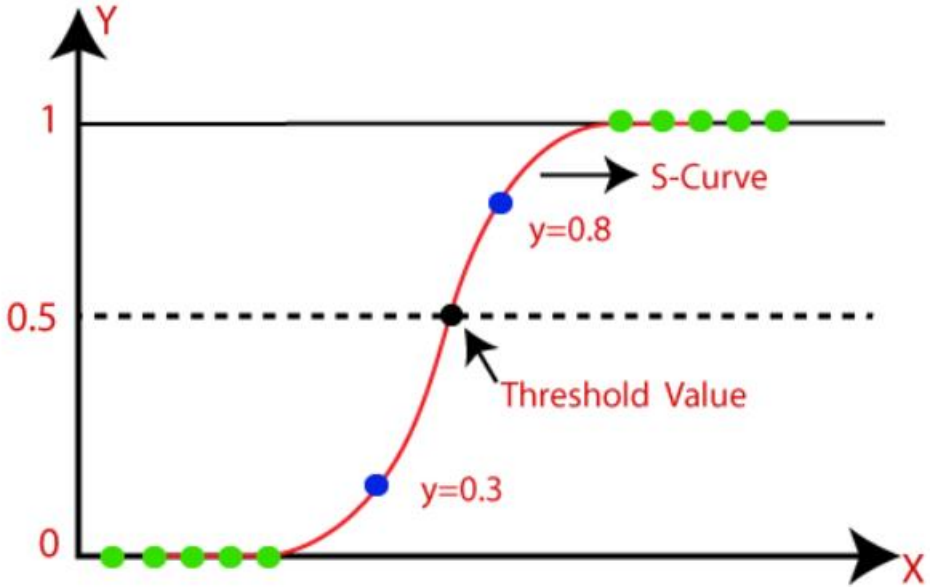
Logistic Regression

- “Logistic regression does not return directly the class of observations.
- It allows us to estimate the probability (p) of class membership. Or in other words, the probability of occurrence of characteristics 1 (= characteristics present) is estimated.

  - For example, in medicine, a common goal is to find which variables have impacts on disease.
  - In this particular case, 0 could stand for ‘not having the disease” and 1 could stand for “having the disease” and how variables like age, gender, and smoking status impact someone’s probability of having that disease.

Source: Kassambara. “Logistic Regression Essentials in RI”. November 03, 2018, accessed on September 29, 2022. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/
Datatab. “Logistic Regression: An Introduction”. February 05, 2021, accessed on October 29, 2022. https://www.youtube.com/watch?v=3tq4t41MsPc/

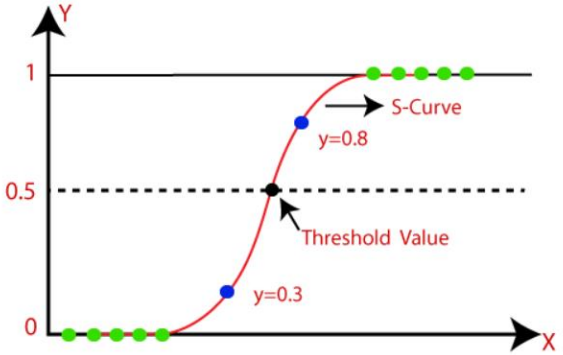
Linear vs logistic Regression

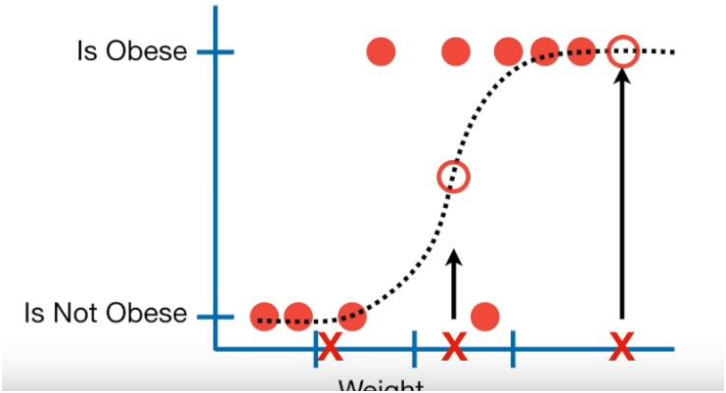


Source: “Linear Regression vs Logistic Regression”. accessed on October 29, 2022. https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Logistic Regression

- The probability will range between 0 and 1.
- You need to decide the threshold probability at which the category flips from one to the other. By default, this is set to p = 0.5, but in reality, it should be settled based on the analysis purpose.”





Source: Kassambara. “Logistic Regression Essentials in RI”. November 03, 2018, accessed on September 29, 2022. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/
StatQuest. “Logistic Regression”. March 05, 2018. accessed on November 02. https://www.youtube.com/watch?v=yIYKR4sgzI8

Logistic Regression: log odds

- “Log odds are an alternate way of expressing probabilities.”
- But they don’t represent the same thing.
- “Odds are the ratio of something happening to something not happening, while probability is the ratio of something happening to everything that could possibly happen.”
- “Odds can be seen as the ratio of “successes” to “non-successes”.
- “Technically, odds are the probability of an event divided by the probability that the event will not take place. For example, if the probability of being diabetes-positive is 0.5, the probability of “won’t be” is 1-0.5 = 0.5, and the odds are 1.0.”

Source: Thanda, A. “What is Logistic Regression? A Beginner's Guide”. October 04, 2022, accessed on October 29, 2022. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/
Kassambara. “Logistic Regression Essentials in Rl”. November 03, 2018, accessed on September 29, 2022. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Logistic Regression: Assumptions

- “The dependent variable is binary —i.e. It fits into one of two clear-cut categories.”
- “There should be no, or very little, multicollinearity between the predictor variables. This means that there should not be a high correlation between the independent variables.
  - Note: In statistics, certain tests can be used to calculate the correlation between the predictor variables. For example, “Spearman’s rank correlation coefficient” or “the Pearson correlation coefficient.”
  - “The Wald test/ Wald Chi-Squared Test is a parametric statistical measure to confirm whether a set of independent variables are collectively 'significant' for a model or not.
- “The independent variables should be linearly related to the log odds.”

Source: Thanda, A. “What is Logistic Regression? A Beginner's Guide”. October 04, 2022, accessed on October 29, 2022. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

Logistic Regression: Data Preparation

“Performing the following steps might improve the accuracy of your model

- Remove potential outliers.
- Make sure that the predictor variables are normally distributed. If not, you can use log, root, or other transformations.
- Remove highly correlated predictors to minimize overfitting. The presence of highly correlated predictors might lead to an unstable model solution.”

Source: Kassambara. “Logistic Regression Essentials in RI”. November 03, 2018, accessed on September 29, 2022. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Logistic Regression in R: Coefficients

```
model <- glm( diabetes ~., data = train.data, family = binomial)
summary(model)$coef
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.50372    1.31719  -7.215 5.39e-13
## pregnant     0.04571    0.06218   0.735 4.62e-01
## glucose      0.04230    0.00657   6.439 1.20e-10
## pressure    -0.00700    0.01291  -0.542 5.87e-01
## triceps      0.01858    0.01861   0.998 3.18e-01
## insulin     -0.00159    0.00139  -1.144 2.52e-01
## mass         0.04502    0.02887   1.559 1.19e-01
## pedigree     0.96845    0.46020   2.104 3.53e-02
## age          0.04256    0.02158   1.972 4.86e-02
```

- “Estimate: the intercept (b0) and the beta coefficient estimates associated to each predictor variable.
- Std.Error: the standard error of the coefficient estimates. This represents the accuracy of the coefficients. The larger the standard error, the less confident we are about the estimate.
- z value: the z-statistic, which is the coefficient estimate (column 2) divided by the standard error of the estimate (column 3).
- Pr(>|z|): The p-value corresponding to the z-statistic. The smaller the p-value, the more significant the estimate is.”

Source: Kassambara. “Logistic Regression Essentials in RI”. November 03, 2018, accessed on September 29, 2022. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Logistic Regression in R: Coefficients

Interpretations of the coefficients:

- "The coefficient estimate of the variable glucose is b = 0.045, which is positive. This means that an increase in glucose is associated with increase in the likelihood/probability of being diabetes-positive."
- "The coefficient for the variable pressure is b = -0.007, which is negative. This means that an increase in blood pressure will be associated with a decreased likelihood/probability of being diabetes-positive."
- Being pregnant (a binary variable) is also positive, indicating that being pregnant increases the likelihood/ probability of being diabetes-positive compared to not being pregnant.

```
model <- glm( diabetes ~., data = train.data, family = binomial)
summary(model)$coef
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.50372    1.31719  -7.215 5.39e-13
## pregnant     0.04571    0.06218   0.735 4.62e-01
## glucose      0.04230    0.00657   6.439 1.20e-10
## pressure    -0.00700    0.01291  -0.542 5.87e-01
## triceps      0.01858    0.01861   0.998 3.18e-01
## insulin     -0.00159    0.00139  -1.144 2.52e-01
## mass         0.04502    0.02887   1.559 1.19e-01
## pedigree     0.96845    0.46020   2.104 3.53e-02
## age          0.04256    0.02158   1.972 4.86e-02
```

Source: Kassambara. "Logistic Regression Essentials in RI". November 03, 2018, accessed on September 29, 2022. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Logistic Regression in R: Odds ratio

Interpretations of the odds ratio:

- "The interpretation of the odds ratio depends on whether the predictor is categorical or continuous/ numeric."
- Odds ratios for continuous predictors:
  - Odds ratios that are greater than 1 indicate that the event is more likely to occur as the predictor increases. Odds ratios that are less than 1 indicate that the event is less likely to occur as the predictor increases.
- Odds ratios for categorical predictors
  - For categorical predictors, the odds ratio compares the odds of the event occurring at 2 different levels of the predictor (assume Level A and B) one being the reference (Level B). Odds ratios that are greater than 1 indicate that the event is more likely at level A. Odds ratios that are less than 1 indicate that the event is less likely at level A.

Source: "Odds Ratios for Fit Binary Logistic Model". accessed on November 03, 2022. https://support.minitab.com/en-us/minitab/20/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/interpret-the-results/all-statistics-and-graphs/odds-ratios/

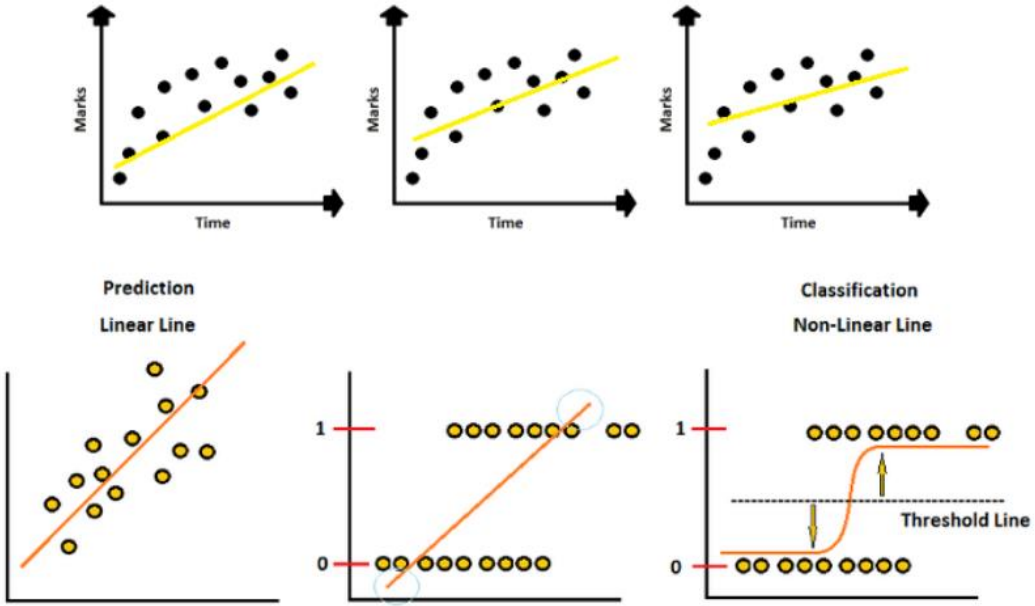
Materials are available at: shorturl.at/aeMZ5

https://drive.google.com/drive/folders/1nyHoAuBAJY6QFjNs5LEOxOOlOpqTudJ3?usp=sharing

Thank you!

Questions: jamals16@mcmaster.ca

Linear vs logistic Regression



Source: Chauhan, A. “Linear Regression vs. Logistic Regression”. July 1, 2021. accessed on October 29, 2022. https://pub.towardsai.net/machine-learning-fcf74f121167