**Module 13**

**Logistic Regression**

* [Video Transcripts](https://student.emeritus.org/courses/4765/files/3332295?wrap=1)
* [Download Video Transcripts](https://student.emeritus.org/courses/4765/files/3332295/download?download_frd=1)
* [Quick Reference Guide](https://student.emeritus.org/courses/4765/files/3332299?wrap=1)

PCA is not good, use LASSO instead!

Feature limits:

5 Logistic

15 Linear

Logistic regression is somewhat like linear regression, but linear regression is ‘unbounded’, meaning that the value the model returns could be anything based on the input of the data. In logistic regression, the value that the model returns is between 0 and 1. This allows you to set a threshold between two classes more easily. Later in this module, you will apply logistic regression to more than two classes.

**Binary Classification**

Classifying data points into one of two classes

**Decision Boundary**

The region in which the output label of a classifier is ambiguous

**Logistic Regression**

A statistical model that uses a logistic function to model a binary-dependent variable

**Multiclass Classification**

Classifying data points into one of three or more classes

**Multinomial Classification**

Another name for multiclass classification

**One-vs-One**

A technique that breaks up a multiclass dataset into multiple binary classification problems where each problem compares one class against one other class

**One-vs-Rest**

A technique that breaks up a multiclass dataset into multiple binary classification problems where each problem compares one class against every other class; also called one-vs-all

**Notes:**

Logistic regression is a supervised machine learning model for determining the probability of the dependent variable. Its major motivation is predicting the likelihood of categorical outcomes that often present binary data such as yes/no and true/false. Some well-known use cases for logistic regression are email spam classifiers and tools for medical diagnoses (e.g., malignant or benign cancers).

One disadvantage of logistic regression is that it does not handle many features simultaneously. For example, if there are a large number of predictors, it may be difficult to interpret or convey the model as a whole. In some cases, it may be best to sacrifice some details and limit the model to a subset of its most substantial features. To accomplish this, regularization can maintain every feature in the model by reducing the magnitude (i.e., cost) of the variables, thereby maintaining both the accuracy and generalization of the model.

Multiclass logistic regression (also known as multinomial logistic regression) predicts more than two classes. Multiclass regression explains how one nominal-dependent variable is related to one or more independent variables. The independent variables can be either continuous or dichotomous (i.e., binary). This type of analysis is often considered attractive, as it does not assume linearity, normality, or homoscedasticity.

In machine learning, binary classification refers to classifying cases into one of two classes. Comparatively, multinomial (or multiclass) classification refers to classifying instances into one of more than two classes. Although some classification algorithms naturally support multiple classes, others are binary algorithms by nature. Nevertheless, these algorithms can be turned into multinomial classifiers using various strategies, namely one-vs-rest and one-vs-one.

**One-vs-Rest (OvR)**

One-vs-rest (also called one-vs-all) involves breaking up a multiclass dataset into multiple binary classification problems. Each binary classification problem is trained using a binary classifier, and predictions are made using the most accurate model.

Consider the following example featuring a multiclass classification problem with examples for each class: orange, red, and yellow. These could be subdivided into three binary classification datasets as follows:

* Classification Problem 1: red vs. [orange, yellow]
* Classification Problem 2: orange vs. [red, yellow]
* Classification Problem 3: yellow vs. [red, orange]

**One-vs-One (OvO)**

Like one-vs-rest, one-vs-one breaks up a multiclass dataset into multiple binary classification problems. However, instead of dividing the dataset into one binary dataset per class, one-vs-one divides the dataset into one dataset per class versus every other class.

Consider the following example featuring a multiclass classification problem with examples for each class: orange, red, and yellow. These could be subdivided into three binary classification datasets as follows:

* Binary Classification Problem 1: red vs. orange
* Binary Classification Problem 2: red vs. yellow
* Binary Classification Problem 3: red vs. red

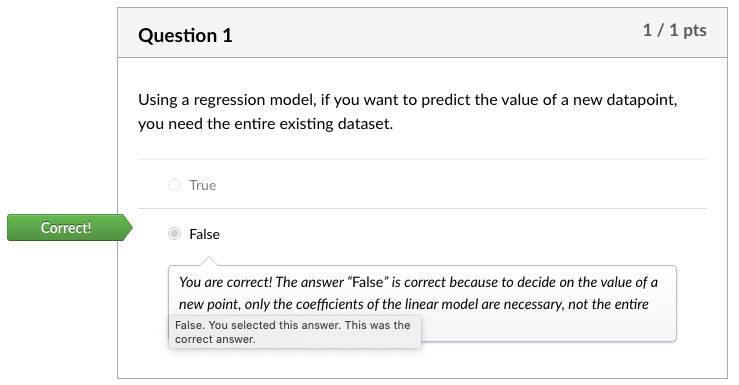
As you can observe, this strategy differs from the previous one-vs-rest strategy in that there is one class vs. another class as opposed to one class vs. multiple classes.

L1 (LASSO) - Mean Absolute Error

L2 (Ridge). - Mean Squared Error

**Module Issues:**

In Quiz 13.1, Question 1, the question is not clear if it is for *logistic* or *linear* regression model.



Codio Activity 1 the entirety is not very clear

Codio Activity 5 Problem 2: There are 2 ‘b’s in the legend, I believe ‘green’ line is supposed to be ‘c’

Codio Activity 5 Problem 4: **x = np.linspace(0, 2600, 100)** should be defined!

Codio Activity 5 Problem 4: LogisticRegression estimator should be created as **clf** and fit.

Codio Activity 6 Problem 2: LogisticRegression(penalty = 'l1', solver = 'liblinear', random\_state=42, max\_iter = 1000)

Codio Activity 8 Problem 10: *selected\_features* must be defined as selected\_features = feature\_names[ [int(i[1:]) **for** i **in** lgr\_pipe\_.named\_steps['selector'].get\_feature\_names\_out()]]

**Quizes:**

Using a regression model, if you want to predict the value of a new datapoint, you need the entire existing dataset. : False

*You are correct! The answer “*False*” is correct because to decide on the value of a new point, only the coefficients of the linear model are necessary, not the entire dataset.*

Which type of regression model outputs a continuous-valued probability that a datapoint belongs to a specific class? : Logistic regression

*You are correct! The answer “*Logistic regression*” is correct because this is the type of regression in which the output is the probability of the input belonging to a specific class.*

Logistic regression is very sensitive toward outliers. : False

*You are correct! The answer “*False*” is correct because logistic regression is not sensitive toward outliers.*

Assume that you have a single feature X and two classes Y1 and Y2. What do you call the ratio of the area of class Y1 and Y2? : Odds Ratio

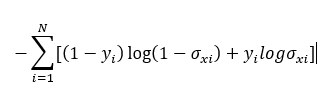
*You are correct! The answer “*Odds ratio*” is correct because the ratio of the two areas is known as odds ratio.*

Given that the odds ratio is equal to “ e-Z ” where “Z” is a linear function of x, what is the formula for Z?  : Z = β0 + β1x

*You are correct! The answer “*Z = β0 + β1x*” is correct because this is the correct representation for Z.*

What is the formula for the logistic function, otherwise known as the sigmoid function? *Choose all that apply*. : *“1/1 + e-Z” and “1/1 + odds ratio”.*

*You are correct! The answers “*1/1 + e-Z*” and “*1/1 + odds ratio*” are correct because these formulas are the correct representations for logistic function.*



The above formula is used to calculate the sigmoid function. : False

*You are correct! The answer “*False*” is correct because the given formula is for “cross-entropy”.*

In Python, to find the “β0” value using the LogisticRegression() object, what code would you write? : lr.intercept\_[0]

*You are correct! The answer “*lr.intercept\_[0]*” is correct because this is the function called on the LogisticRegression() object to get the value of the intercept*“*β0*”.

The symbol “N” represents the complete samples in a dataset. : True

*You are correct! The answer “*True*” is correct because “N” represents the complete samples in a dataset.*

You are given a model with two features, X1 and X2, as well as a categorical class Y. In the logistic function for this model, what would “Z” be equal to? : Z = β0 + β1x1 + β2x2

*You are correct! The answer “*Z = β0 + β1x1 + β2x2*” is correct because it is a linear combination of the inputs with an intercept term “*β0*” and coefficients “*β1*” and “*β2*” for the two inputs x1 and x2.*

What does it mean when there is ‘no solution’ to the optimization of a logistic regression? : The data is cleanly separable

*You are correct! The answer “*The data is cleanly separable*” is correct because the coefficients “*β1*” and “*β2*” can be increased to the point where the solution will converge toward an infinitely steep step function where the data is completely separable.*

Which regularization terms can be added to the logistic regression calculation to avoid a ‘no solution’ issue? *Choose all that apply*. : L1, L2

*You are correct! The answers “*L1*” and “*L2*” are correct because these regularizations could be added to the cost function to penalize large coefficient values.*

The regularization parameter in scikit-learn is called “c”. To increase the strength of the regularization, the value of “c” decreases. : True

*You are correct! The answer “*True*” is correct because the regularization parameter in scikit-learn is called “c”, not lambda, and it corresponds to the inverse of lambda. Thus, to increase the strength of the regularization, you actually need to decrease “c”.*

When initializing a “LogisticRegression()” object, which constructor is set to what value to make it a Lasso regularization? : penalty= ’l1’

*You are correct! The answer “*penalty= ’l1’*” is correct because this is the constructor with the value used to make the logistic regression a LASSO regularization.*

Which of the following is **not** a method for adapting binary classification models to multiclass problems? : One-vs-zero

*You are correct! The answer “*one-vs-zero*” is correct because it is not a method for adapting binary classification models to multiclass problems.*

For a one-vs-one approach, if there are five classes, how many binary models should be built for binary classifiers? : 10 (5 x (5-1) / 2)

*You are correct! The answer “*10*” is correct because the total number of executions of the binary classifier is given using K × (K*− *1) / 2. Thus, the result is 5 × 4 / 2 = 10.*

One-vs-rest has the advantage over one-vs-one because in one-vs-rest, the number of models that need to be trained grows quadratically with the number of classes : False

*You are correct! The answer “*False*” is correct because the advantage one-vs-rest has is that the number of models that need to be trained grows linearly with the number of classes.*

Is the size of the training problems smaller in one-vs-one or one-vs-rest?  : One-vs-one

*You are correct! The answer “*one-vs-one*” is correct because the size of the training problems is smaller in one-vs-one by a factor of K*− *1.*

What is the value in the denominator of the multinomial logistic regression formula? : 1 +

∑k-1k=1 exp(−βk.x)

*You are correct! The answer “*1 + ∑k-1k=1 exp(−βk.x)*” is correct because this is the value in the denominator of the multinomial logistic regression formula.*

When creating a “LogisticRegression()” object, what constructor do you use to declare the logistic regression as “ovr” or “multinomial”? : multi\_class

*You are correct! The answer “*multi\_class*” is correct because this is the constructor used in creating a “LogisticRegression()” object to declare the type of logistic regression.*

**Try-It Activity 13.1:** Using Logistic Regression to Make Business Decisions

**Las Vegas Strip Data Set**

<https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip#>

* The dataset and its features
* The classification problem
* A business decision that can be supported using the results of the classification model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Abstract**: This dataset includes quantitative and categorical features from online reviews from 21 hotels located in Las Vegas Strip, extracted from TripAdvisor ([[Web Link]](http://www.tripadvisor.com/)). | | |  | | |
| **Data Set Characteristics:** | N/A | **Number of Instances:** | 504 | **Area:** | Business |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 20 | **Date Donated** | 2017-07-23 |
| **Associated Tasks:** | Classification, Regression | **Missing Values?** | N/A | **Number of Web Hits:** | 120448 |

 ————— o —————

**Clickstream of Online Shopping Dataset**

<https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping>

**Dataset Information:**

The dataset contains information on clickstream from online store offering clothing for pregnant women. Data are from five months of 2008 and include, among others, product category, location of the photo on the page, country of origin of the IP address and product price in US dollars.

**Attribute Information:**

The dataset contains 14 variables described in a separate file (See 'Data set description')

Data description ìe-shop clothing 2008î

Variables:

1. YEAR (2008)

========================================================

2. MONTH -> from April (4) to August (8)

========================================================

3. DAY -> day number of the month

========================================================

4. ORDER -> sequence of clicks during one session

========================================================

5. COUNTRY -> variable indicating the country of origin of the IP address with the

following categories:

1-Australia

2-Austria

3-Belgium

4-British Virgin Islands

5-Cayman Islands

6-Christmas Island

7-Croatia

8-Cyprus

9-Czech Republic

10-Denmark

11-Estonia

12-unidentified

13-Faroe Islands

14-Finland

15-France

16-Germany

17-Greece

18-Hungary

19-Iceland

20-India

21-Ireland

22-Italy

23-Latvia

24-Lithuania

25-Luxembourg

26-Mexico

27-Netherlands

28-Norway

29-Poland

30-Portugal

31-Romania

32-Russia

33-San Marino

34-Slovakia

35-Slovenia

36-Spain

37-Sweden

38-Switzerland

39-Ukraine

40-United Arab Emirates

41-United Kingdom

42-USA

43-biz (\*.biz)

44-com (\*.com)

45-int (\*.int)

46-net (\*.net)

47-org (\*.org)

========================================================

6. SESSION ID -> variable indicating session id (short record)

========================================================

7. PAGE 1 (MAIN CATEGORY) -> concerns the main product category:

1-trousers

2-skirts

3-blouses

4-sale

========================================================

8. PAGE 2 (CLOTHING MODEL) -> contains information about the code for each product

(217 products)

========================================================

9. COLOUR -> colour of product

1-beige

2-black

3-blue

4-brown

5-burgundy

6-gray

7-green

8-navy blue

9-of many colors

10-olive

11-pink

12-red

13-violet

14-white

========================================================

10. LOCATION -> photo location on the page, the screen has been divided into six parts:

1-top left

2-top in the middle

3-top right

4-bottom left

5-bottom in the middle

6-bottom right

========================================================

11. MODEL PHOTOGRAPHY -> variable with two categories:

1-en face

2-profile

========================================================

12. PRICE -> price in US dollars

========================================================

13. PRICE 2 -> variable informing whether the price of a particular product is higher than

the average price for the entire product category

1-yes

2-no

========================================================

14. PAGE -> page number within the e-store website (from 1 to 5)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++

If you use this dataset, please cite:

£apczyÒski M., Bia≥owπs S. (2013) Discovering Patterns of Users' Behaviour in an E-shop -

Comparison of Consumer Buying Behaviours in Poland and Other European Countries,

ìStudia Ekonomiczneî, nr 151, ìLa sociÈtÈ de l'information : perspective europÈenne et

globale : les usages et les risques d'Internet pour les citoyens et les consommateursî, p. 144-

153.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++

**Try-It Activity 13.1:** Using Logistic Regression to Make Business Decisions

**Overview**

* The dataset and its features
* The classification problem
* A business decision that can be supported using the results of the classification model

**Clickstream of Online Shopping Dataset Information:**

The dataset contains information about clickstream from online store offering clothing for pregnant women. Data are from five months of 2008 and include, among others, product category, location of the photo on the page, country of origin of the IP address and product price in US dollars.

<https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping>

**Dataset Features:**

The dataset contains 14 variables described:

Variables:

1. YEAR (2008)

2. MONTH -> from April (4) to August (8)

3. DAY -> day number of the month

4. ORDER -> sequence of clicks during one session

5. COUNTRY -> country name indicating the originating IP address.

6. SESSION ID -> variable indicating session id (short record)

7. PAGE 1 (MAIN CATEGORY) -> concerns the main product category (4 categories)

8. PAGE 2 (CLOTHING MODEL) -> contains information about the code for each product

(217 products)

9. COLOUR -> colour of product (14 colors)

10. LOCATION -> photo location on the page, the screen has been divided into six parts.

11. MODEL PHOTOGRAPHY -> variable with two categories

12. PRICE -> price in US dollars

13. PRICE 2 -> variable informing whether the price of a particular product is higher than

the average price for the entire product category

14. PAGE -> page number within the e-store website (from 1 to 5)

**Classification**

As explained above the dataset contains information on an online store offering clothing for pregnant women. The data has high priced merchandise where it is marked in the attribute name *price 2* where can be used to categorize merchant as high or regular priced.

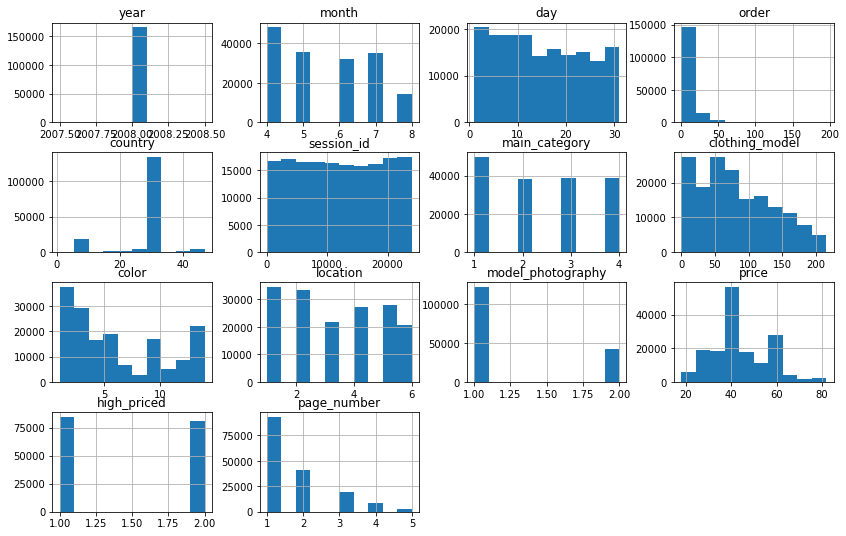
**Business Decision**

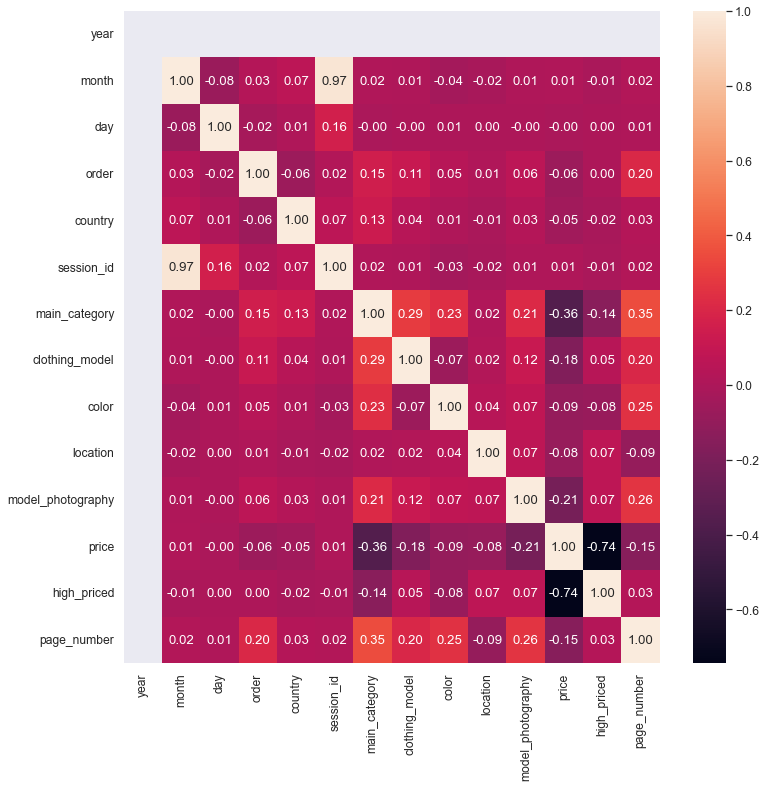
Once merchants are classified with this information, it can be further used to customize recommendations and offerings to serve high profile customers from high priced categories as top recommendations besides items from regular category, similarly for normal customers can be recommended items from regular category as top choices. The presentation order of those can be set by this classification.

**Classification Model**

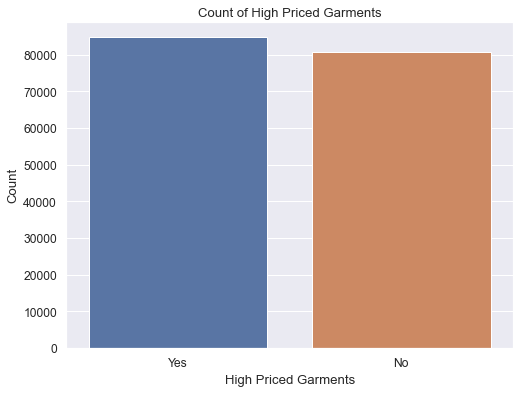
The dataset is almost ready to use, columns are renamed for easiness, price column smoothened out by np.log() function.

I removed ['year', 'month', 'day', 'order', 'session\_id’] attributes from the dataset for modeling.





I checked data distribution, 51% and 49% as below:



Then, I created a model:

# since there is no aplhanumeric columns, we can just scale and classify data, no transformation needed

# Note: n\_neighbors=5 and weights='uniform' by default!

estore\_pipeline = Pipeline([('scale', StandardScaler()),

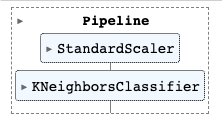
('knn', KNeighborsClassifier(n\_neighbors=5, weights='distance'))])

estore\_pipeline.fit(X\_train, y\_train)

estore\_preds = estore\_pipeline.predict(X\_test)

estore\_proba = estore\_pipeline.predict\_proba(X\_test)

estore\_pipeline



I just tweaked the hyper-parameters as *n\_neighbors=5* and *weights='distance’* which yielded good result:

fig, ax = plt.subplots(1, 2, figsize = (14, 7))

ConfusionMatrixDisplay.from\_predictions(y\_test, estore\_preds, values\_format='d',

display\_labels=['Yes', 'No'], ax=ax[0])

ax[0].set\_title('Confusion Matrix')

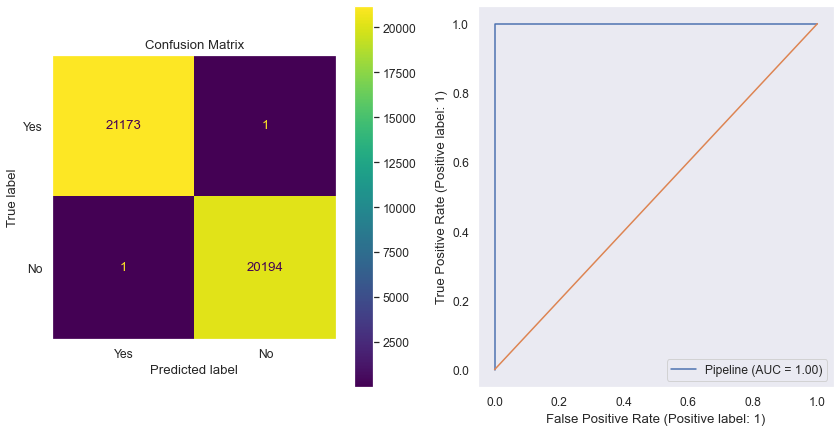
ax[0].grid(False)

RocCurveDisplay.from\_estimator(estore\_pipeline, X\_test, y\_test, pos\_label=1, ax = ax[1])

ax[1].plot(np.array([0, 1]), np.array([0, 1]))

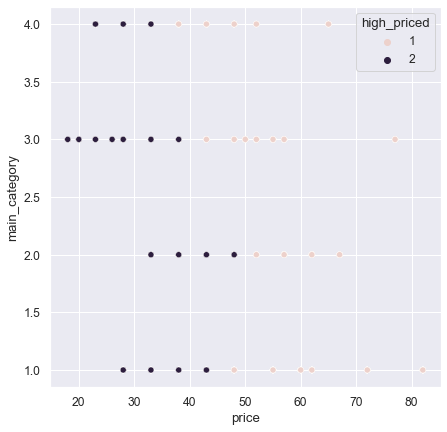
ax[1].grid()

plt.show()



**Conclusion**

The Business Decision objective is achievable with this dataset by a classification model to classify garments as high or regular priced items so it can be used in customizing offerings per customer by the online store.



 ————— o —————

**Discussion 13.1: Business Application of Logistic Regression**

How logistic regression could be used in the real world

* Describe a context where logistic regression would be appropriate
* Describe the necessary dataset to support logistic regression in your chosen context
* If you can, suggest ways to gather this data

**Using Logistic Regression For Forecasting Rain In Real-World:**

I searched around for an interesting topic but I decided to go with a classical logical regression example although we discussed it in one of the sessions.

I addressed the above 3 points in this conceptual discussion.

**Logistic Regression Application**

Logistic Regression is a **white-box** model can be derived for explaining how it arrived to that conclusion therefore results are *back traceable* which can be used for medical, banking, financial applications and any field under a regulation. For that reason, the model application can be wide-spread as it addresses the regulatory concern already.

Logistic Regression can be used to predict rain with probability, it is ideal because the expected outcome is categorical as well as probability of it also provided, it gives end users an understanding how rain is probable on a certain day.

**Necessary Dataset**

Related data attributes for this exercise would be *temperature, humidity, dew point, precipitation*, and *atmospheric pressure* either those measurements can be done individually or looked up from other online sources.

**Gathering Dataset**

Historical weather data collection can be looked upon <https://openweathermap.org/api> which is free open source for educators and students. For other individuals, some sites offer past weather data like <https://www.timeanddate.com/weather/@5341628/historic> where such information can be gathered to build above dataset.

Odds ratio calculation from coefficients:

