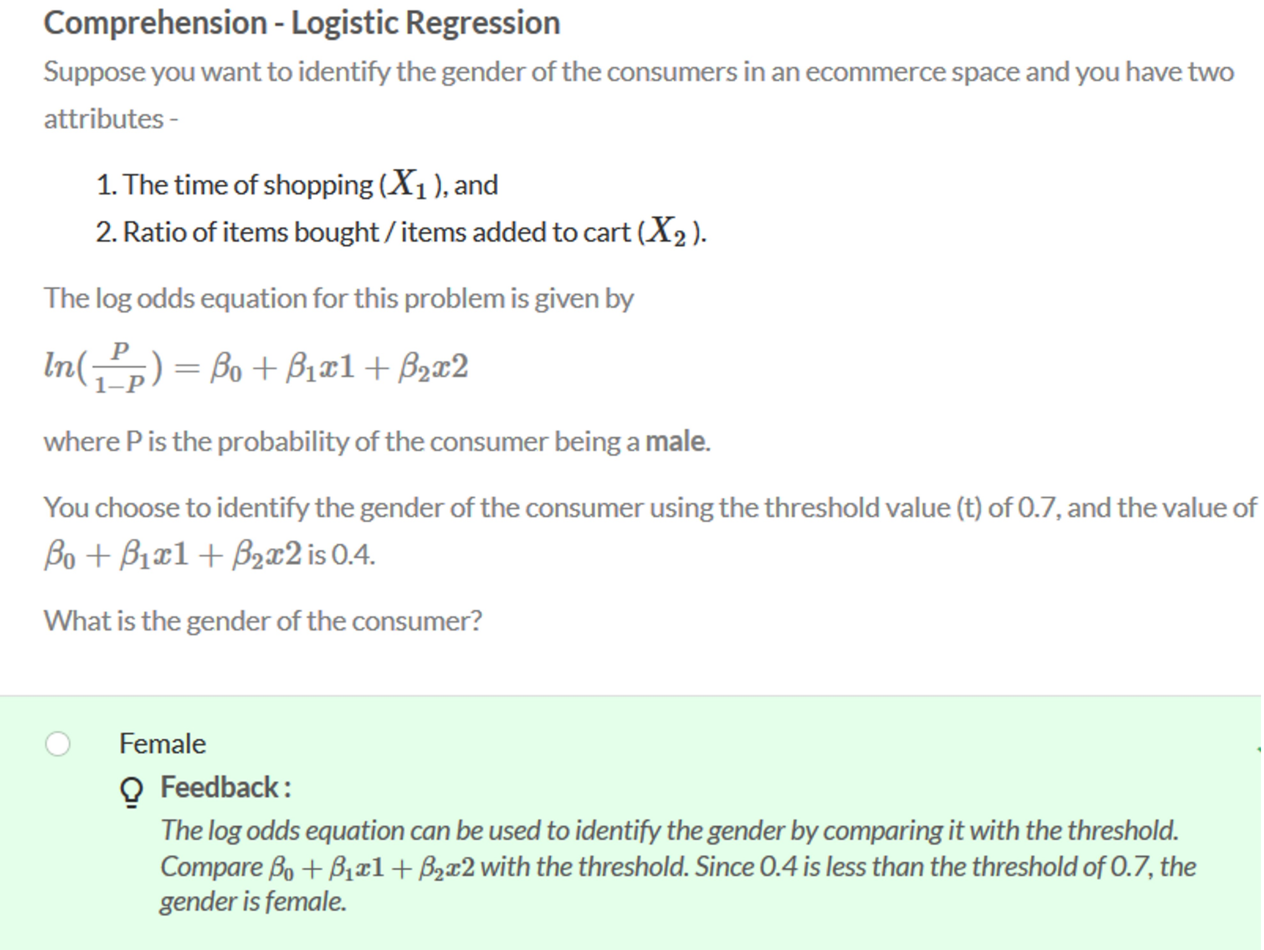
Q&A

08 March 2020 13:40

1.)



Comprehension - Logistic Regression

Suppose you want to identify the gender of the consumers in an ecommerce space and you have two attributes -

1. The time of shopping (X_1), and
2. Ratio of items bought / items added to cart (X_2).

The log odds equation for this problem is given by

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x1 + \beta_2 x2$$

where P is the probability of the consumer being a **male**.

You choose to identify the gender of the consumer using the threshold value (t) of 0.7, and the value of $\beta_0 + \beta_1 x1 + \beta_2 x2$ is 0.4.

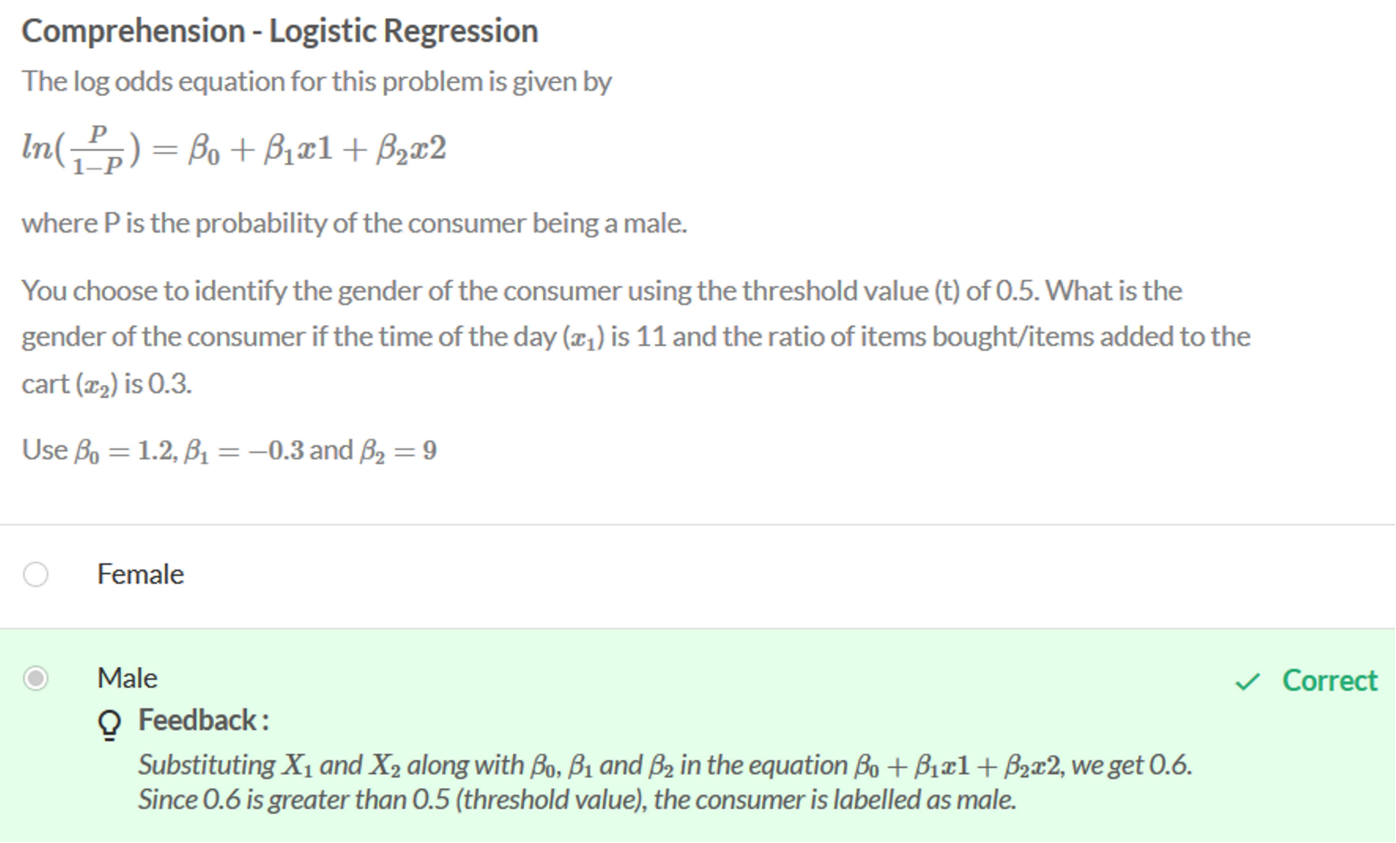
What is the gender of the consumer?

○ Female

Feedback :

The log odds equation can be used to identify the gender by comparing it with the threshold. Compare $\beta_0 + \beta_1 x1 + \beta_2 x2$ with the threshold. Since 0.4 is less than the threshold of 0.7, the gender is female.

2.)



Comprehension - Logistic Regression

The log odds equation for this problem is given by

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x1 + \beta_2 x2$$

where P is the probability of the consumer being a male.

You choose to identify the gender of the consumer using the threshold value (t) of 0.5. What is the gender of the consumer if the time of the day (x_1) is 11 and the ratio of items bought/items added to the cart (x_2) is 0.3.

Use $\beta_0 = 1.2$, $\beta_1 = -0.3$ and $\beta_2 = 9$

○ Female

● Male ✓ Correct

Feedback :

Substituting X_1 and X_2 along with β_0, β_1 and β_2 in the equation $\beta_0 + \beta_1 x1 + \beta_2 x2$, we get 0.6. Since 0.6 is greater than 0.5 (threshold value), the consumer is labelled as male.

3.) Logistic Regression
Which of the following is/are correct about logistic regression?
A logistic regression model calculates the class probabilities of all the classes of the outcome variable, while predicting a test case.
Feedback :
Logistic regression calculates the class probabilities of all the classes present in the outcome variable, using the logistic function. The final class is predicted by providing a cutoff value.
The decision boundary of an LR model is a straight line.
Feedback :
The logistic regression model separates two different classes using a line linearly. The sigmoid curve is only used to calculate class probabilities. The final classes are predicted based on the cutoff chosen after building the model.

4.) SVM Kernels
In which of the following situations would a radial kernel of SVM perform better than a linear kernel, to separate the two groups?

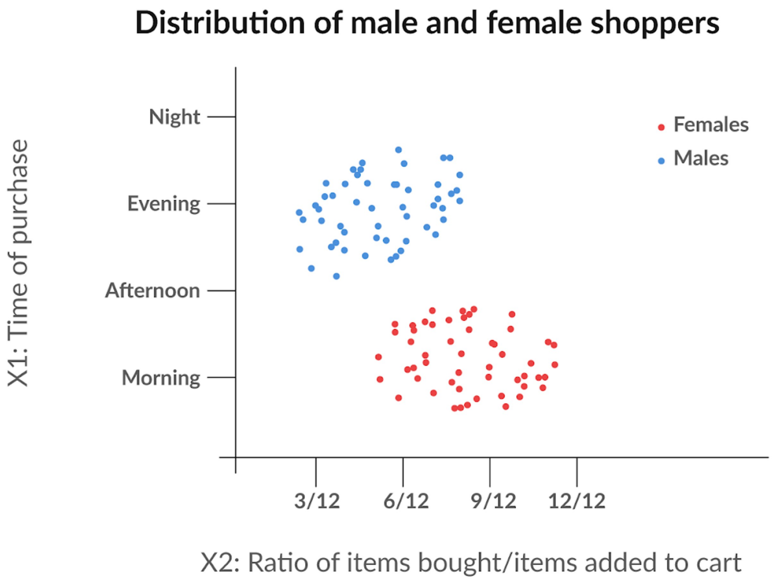


Figure A

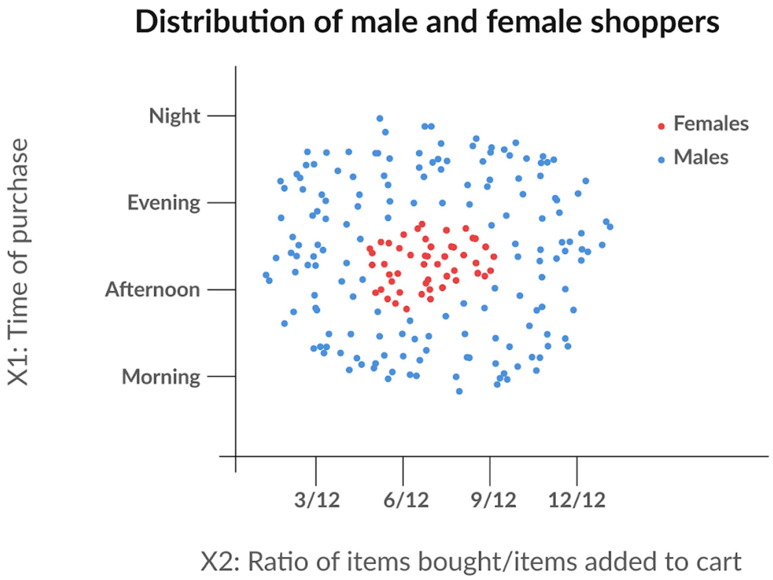


Figure B

Figure B
Feedback :
A radial kernel would be the choice in this case because the data is inseparable using a linear kernel.

5.) Comparing Different Machine Learning Models - II
Which model gives the lowest accuracy?
Logistic Regression
Feedback :*Run the logistic regression model and check the accuracy on the test dataset. The accuracy on the test set is around 56%.*

6.) Comparing Different Machine Learning Models - II
Which kernel function in SVM works best in this case?
Radial
Feedback :*SVM with the radial kernel gives the highest accuracy and outperforms all the other*

kernels.

7.) Pros and Cons of Different Machine Learning Models
Which of the following algorithms will not perform well when the relationship between the dependent and independent variable in a data set is nonlinear?
Logistic regression
Feedback :*Since logistic regression separates the different classes of the dependent variable using a line, there should be linear dependence between the dependent and independent variable for it to work well.*
Correct

8.) Pros and Cons of Different Machine Learning Models
Which of these algorithms' result is easiest to explain to someone who does not possess any knowledge of machine learning?
Decision trees
Feedback :*Because of the intuitive nature of decision trees, their results can be easily interpreted and explained.*

9.) Pros and Cons of Different Machine Learning Models
Among logistic regression, decision trees and support vector machines, which one is best suited for a dataset having lots of categorical variables?
Decision trees
Feedback :*Decision trees are best suited for a dataset with a lot of categorical data because of the way in which node splitting is performed. Decision trees do not need the categorical features to be converted into numeric features.*

10.) Pros and Cons of Different Machine Learning Models
Among logistic regression, decision trees and support vector machines, which one is best suited if you want to explain the model to, say, a doctor?
Decision trees
Feedback :*Out of the listed machine learning models, decision trees are the easiest to explain because of the similarity of the decision-making process between trees and humans.*

11.)Pros and Cons of Different Machine Learning Models
Say you work for a large e-commerce company such as Amazon and need to build a classification model to classify a user as likely to buy / unlikely to buy. You have a large number of features and observations, and have to deploy the model in real time.
Compare the pros and cons of logistic regression, decision trees and support vector machines in such a case. Write your arguments and the final choice/approach in the box below.
Cons: 1. Logistic regression might not perform as well as other algorithms in terms of accuracy and other such performance metrics because of the potential nonlinearity in the dataset. 2. Decision trees are prone to overfit the data by creating complex rules which mug up the whole data. 3. Support vector machines might not be appropriate for this task since it requires the model to be deployed in real time, and as discussed earlier, SVMs are resource hungry and slow as compared to other machine learning models. Pros: 1. Since the project is to be deployed in real time, logistic regression and decision trees will be the right choice since they are faster to build than support vector machines. 2. In general, support vector machines give a really good performance as compared to logistic regression or decision trees when the number of features is large. In the end, you have to test and compare all the models in terms of the following - 1. Predictive power (accuracy, sensitivity and specificity, AUC etc.), and 2. Computational cost After analysing the above, you have to choose the model that gives a right balance of both the goals.

12.) Pros and Cons of Different Machine Learning Models
Among logistic regression, decision trees and support vector machines, which one is the least suited for a nonlinear decision boundary?
Logistic regression
Feedback :*Logistic regression is a linear model and it can not create a nonlinear boundary.*

13.) CART vs CHAID
What is the main difference between CART and CHAID trees?
CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node).
Feedback :*As explained in the video, CART can only build binary trees, whereas CHAID can build multiway trees.*

14.) CART vs CHAID
Suppose you are asked to build a decision tree model for a classification problem, by your manager. The aim of the organisation is to understand the driver KPIs (Key Performance Indicators), i.e. features that play an important role in the problem at hand. Which type of tree would be more appropriate for this task?
CHAID (Chi-square Automatic Interaction Detection)
Feedback :*CHAID trees are suitable when you need to understand the driver KPIs, instead of predicting the class.*

15.) Random Forests vs Decision Trees
Which of the following are the advantages of random forests over decision trees?
Random forests solve the problem of overfitting, a problem commonly faced by decision trees.
Feedback : Random forests use bagging along with sampling the features randomly at each node split. This prevents them from overfitting the data, unlike decision trees.
Correct
Random forests do not require tree pruning.
Feedback : There is no need to prune trees in a random forest because even if some trees overfit the training set, it will not matter when the results of all the trees are aggregated.
Correct

16.) End-to-End Modelling - II
Which is the more appropriate approach towards model building?
Start from a simple model; and only build complex models if the simple ones do not meet the required standards.
Feedback :*Starting from a basic model helps in two ways: 1) If the model performs as per requirement, there is no need to go to complex models. This saves time and resources. 2) If it does not perform well, it can be used to benchmark the performance of other models.*