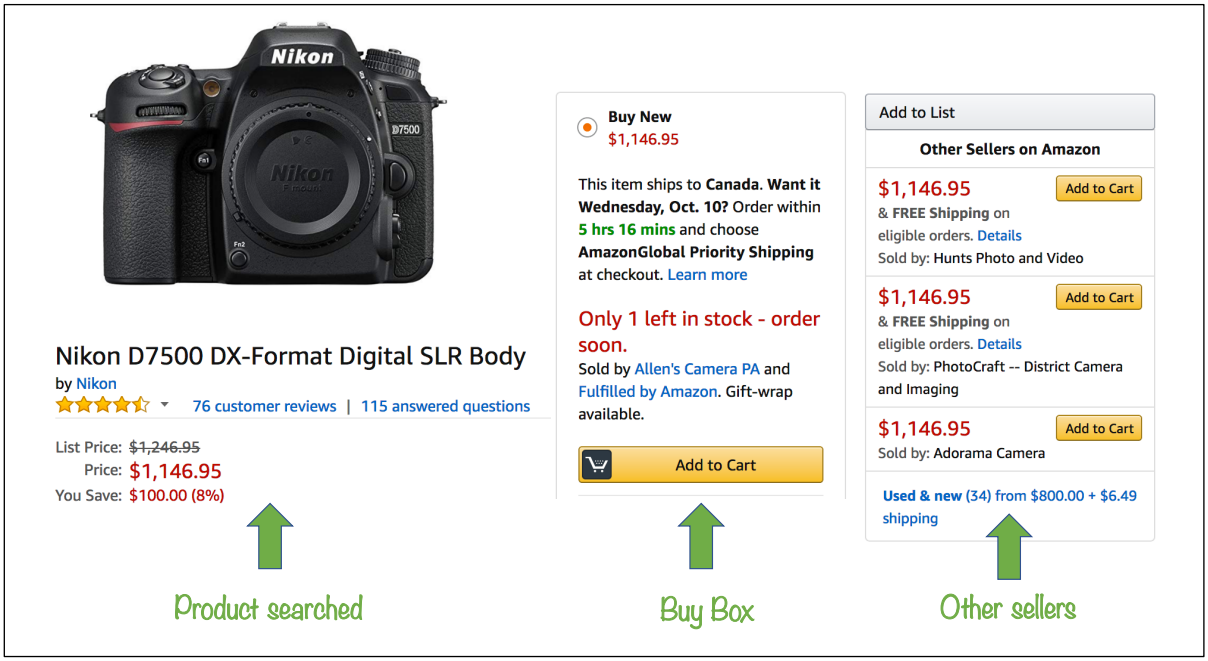
IE 544 Decision Analysis
Assignment 3

How to Win Amazon's Buy Box?

In this assignment, you are going to work with data collected from Amazon.com and try to understand the important factors in winning Amazon's Buy Box. Amazon.com is an online marketplace. Amazon is a (re)seller in this marketplace but there are hundreds or even thousands of other sellers present in the marketplace as well[1].

Amazon.com has a mechanism called the "Buy Box". When a customer searches for a product and there are multiple sellers of this product, she sees a Buy Box with a price and a seller. This means that a seller has "won" the Buy Box for this (possible) transaction. Below the Buy Box, Amazon also lists other sellers and their prices.

The following figure shows an example:



Buy Box clearly outlines the price, the seller (in this example the seller who "won" the Buy Box is Allen's Camera from PA) and whether this shipment will be fulfilled by Amazon or not[2]. Below the Buy Box, Amazon also lists few other sellers and their prices with a link to a larger list (in this example three other sellers are shown with a link to 34 other sellers that sell the same product as new or used). Buyers rarely go through all sellers and usually choose the one suggested by the Buy Box. Therefore, winning the Buy Box is critical for all sellers.

---

[1] Amazon announced that, in 2017, more than 50% of total number of products sold through their platform had been by third party sellers for the first time in their history. In 2018, the number of 3rd party sellers increased further.

[2] Fulfillment by Amazon (FBA) is a program where independent sellers use the logistics infrastructure (warehouses, fulfilment centers, shipment options etc.) of Amazon. Amazon carefully selects independent sellers to include in its FBA program. When buyers see the FBA sign, they know that Amazon is responsible for the delivery of their order.

Amazon does not disclose the details of its algorithm that chooses the winner of the Buy Box but declares that price is not the only factor and claims that seller rating, customer reviews etc. play a role. Possibly there is also a mechanism for sellers to increase their potential to win the Buy Box by paying to Amazon which is not observable to us.

You have access to data collected from amazon.com between 11-08-2015 and 13-09-2015[3]. This data is available as "`amz_train.rds`" in Moodle along with this assignment. Description of the data columns is available in *Table 1*. There is also going to be a test data ("`amz_test.rds`") in the same format for you to test your learning algorithm.

The tasks of the assignment are as follows:

1. (20 points) Using the training data and prepare a descriptive analytics report that contains at least the following: (Do not produce tens of tables and graphs. Use aggregation, summarization and appropriate graphs. The aim is to describe interesting aspects of your data).
    a. Number of Products, number of sellers (overall and by day)
    b. Average, max, min price, buy-box price, shipping cost by product
    c. Seller ratings, positive feedbacks and counts, product ratings and counts by product
    d. Percentage of buy-box successes when Amazon is the seller against every other seller (total and by product and week)
    e. Prime, FBA, Page and Rank information by seller and by product
    f. Any other descriptive statistics you deem appropriate (consider defining new features, particularly with respect to prices)
    g. What are the scales on which the data in each column is measured on?
    h. Are there outliers in the data? If so, should they be discarded?

2. (10 points) On the basis of the observations you made in part (1), interpret what your data seems to describe with respect to how a seller wins the buy-box.

3. (30 points) Using the training data and using at least one score-based, one constraint-based, and one hybrid algorithm from **bnlearn**, learn the structure of a BN (a DAG). Leave the date (epoc) out of consideration in learning. Use a whitelist and/or a blacklist in accordance with your answers to part (2). Experiment with leaving some columns out (simpler the better) and consider discretizing continuous variables if necessary. Choose a model (or a model for each product) by explaining why you make that model choice (consider using cross-validation and/or bootstrapping).

4. (25 points) For your chosen model(s),
    a. learn parameters from the training data using **bnlearn**,
    b. compute the probability that Amazon wins the buy-box
    c. make a prediction (in making a prediction, your model **cannot contain** any variable related to buy-box: `bbox-price`, `bbox-sid`, bbox or other features based on them as they are not observed yet)
    d. report the accuracy of your prediction using the test data. I will make test data with bbox related columns available 2 days before the due date for you to finish this step.
5. (10 points) Give a conclusion about the buy-box winning factors and the appropriateness of using BNs in this framework.

[3] The source of the data is: Chen, Le, Alan Mislove, and Christo Wilson. 2016. "An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace." In *Proceedings of the 25th International Conference on World Wide Web*, 1339–1349. WWW '16. Switzerland. https://doi.org/10.1145/2872427.2883089.

Grading: There is no unique answer to these questions. You will be graded on the basis of your creativity in handling the problem, how you support your answers with sound arguments based on theory and data. Remaining 5 points is for the format and clarity of your report.

Use only **bnlearn** in R and probabilistic networks as your models. Other classification methods are not interesting to us for the purposes of this assignment.

Table 1 Description of the Data

COLUMN	DESCRIPTION
1	pid - (string) The unique product id (or ASIN) that Amazon assigns to each product.
2	epoc -(POSIXct) The timestamp at which this sample was collected.
3	sid - (string) The unique seller id that Amazon assigns to each seller. Note that if the seller is Amazon, the seller id has been saved as "amazon".
4	price - (num) The item price listed by a seller. Note that the item price cannot be 0 for any item. For samples where the item price was unavailable (due to an error in downloading the page, scraping the data or because the price will only be displayed after adding the item to the cart), the item price has been saved as "0". Users are advised to ignore these (price==0) data points when running experiments involving item price.
5	sid_rating - (num) The "star" rating of a seller ranging from 0 to 5 (both inclusive).
6	sid_pos_fb - (num) The positive feedback score of a seller ranging from 0 to 100 (both inclusive).
7	sid_rating_cnt - (num) The number of ratings received by a seller. Scaled between 0 and 100
8	shipping - (num) The shipping price listed by a seller. Note that the shipping price can be 0. For samples where shipping price was unavailable (due to an error in downloading the page, scraping the data or because the price will only be displayed after adding the item to the cart), the shipping price has been saved as "NA". Users are advised to ignore these data points when running experiments involving shipping price.
9	page - (factor) The page number at which a seller was listed. Minimum page number is "1".
10	rank - (factor) The rank of a seller on the page it was listed. Minimum rank is "0".
11	pid_rating - (num) The "star" rating of a product ranging from 0 to 5 (both inclusive). For samples where the rating was unavailable, the rating has been saved as "nan".
12	pid_rating_cnt - (num) The number of ratings received by a product. Scaled between 0 and 1.
13	is_fba - (factor) Is this offer listing Fulfilled By Amazon? <yes/no>.
14	is_prime - (factor) Does the seller offer Prime shipping for this listing? <yes/no>.
15	bbox_sid - (string) The seller selected by Amazon as the default (Buy Box) seller for the product.
16	bbox_price - (num) The price by the Buy Box seller on the product page for a product.
17	bbox – (factor) Whether the seller “won” the Buy Box or not (failure/success). i.e, if bbox_sid==sid then success else failure