26.05.2023

IE 544 DECISION ANALYSIS

ASSIGNMENT 3

GROUP C

ALİ CAN ŞAHİN - 2018402189

SELİN COŞKUN – 2018402213

1. Introduction

In this report some analysis on the data provided by Amazon is made. This data contains some information about the online shopping activities that are done on some products from different sellers. A product can be sold by different number of sellers. Amazon has an algorithm to select a seller which will be on the buy-box. Winning the buy-box certainly give advantage to the sellers to sell their products. This study aims to find an appropriate model to estimate the winner of the buy-box. Before starting with the modeling part, some descriptive analyses are made.

2. Descriptive Data Analysis

2.a) Number of Products/Sellers Overall and Day

Overall, there are 9 different products and 184 different sellers. The total amount of products that are sold each day and the total amount sold for each product type for each day are given in the following tables.

Table 1: Total Amount of Products that are Sold Each Day

	day	TotalProduct
	<chr>	<int>
1	2015-08-11	3276
2	2015-08-12	3407
3	2015-08-13	3220
4	2015-08-14	6296
5	2015-08-15	6402
6	2015-08-16	6240
7	2015-08-17	6744
8	2015-08-18	6382
9	2015-08-19	6481
10	2015-08-20	7016
11	2015-08-21	6724
12	2015-08-22	6645
13	2015-08-23	6629
14	2015-08-24	6097
15	2015-08-25	6529
16	2015-08-26	6232
17	2015-08-27	4476
18	2015-08-28	2568
19	2015-08-29	6435
20	2015-08-30	6320
21	2015-08-31	6265
22	2015-09-01	6408
23	2015-09-02	3486

Table 2: Total Amount Sold for Each Product Type for Each Day

	2015-08-11	2015-08-12	2015-08-13	2015-08-14	2015-08-15	2015-08-16	2015-08-17	2015-08-18	2015-08-19	2015-08-20	2015-08-21	2015-08-22	2015-08-23	2015-08-24	2015-08-25	2015-08-26	2015-08-27	2015-08-28
B002ZV0OJO	506	544	512	1018	1008	961	1012	1010	987	1061	1028	1023	1066	1099	1077	1110	780	440
B0083H1INK	436	462	407	858	876	800	886	861	837	888	880	902	964	1004	1036	1045	746	396
B00AMFLZLG	112	112	112	220	220	204	220	220	220	216	220	222	269	269	309	301	195	110
B00DNSO1OW	362	366	349	667	611	630	691	775	739	701	675	732	772	772	796	897	598	264
B00DNSO41M	168	180	159	186	288	466	404	0	228	674	702	702	391	15	482	0	0	0
B00MVVI1FC	166	175	189	385	399	379	451	456	445	446	484	564	581	568	525	526	369	216
B00VSIT5UE	580	600	560	1100	1100	1020	1120	1120	1080	1120	1100	1080	1120	1120	1120	1100	780	440
B00VSITBJO	540	580	560	1100	1100	1000	1120	1100	1100	1120	1080	1080	1120	1100	1120	1120	780	440
B00YR6BMS2	406	388	372	762	800	780	840	840	845	790	555	340	346	150	64	133	228	262

	2015-08-29	2015-08-30	2015-08-31	2015-09-01	2015-09-02
B002ZV0OJO	1099	1100	1099	1100	553
B0083H1INK	988	969	933	930	474
B00AMFLZLG	275	265	270	275	140
B00DNSO1OW	663	645	636	660	324
B00DNSO41M	0	0	0	0	0
B00MVVI1FC	512	361	279	449	495
B00VSIT5UE	1100	1080	1100	1080	560
B00VSITBJO	1080	1100	1100	1080	520
B00YR6BMS2	718	800	848	834	420

Table 3: Number of Seller Each Day

	Number of Sellers
2015-08-11	96
2015-08-12	97
2015-08-13	95
2015-08-14	99
2015-08-15	111
2015-08-16	107
2015-08-17	113
2015-08-18	98
2015-08-19	104
2015-08-20	103
2015-08-21	106
2015-08-22	96
2015-08-23	105
2015-08-24	103
2015-08-25	94
2015-08-26	96
2015-08-27	96
2015-08-28	88
2015-08-29	100
2015-08-30	102
2015-08-31	90
2015-09-01	105
2015-09-02	111

Amounts of products differ from day to day. On some days, some products were never sold. The total transactions that occurred also differ from day to day but generally it is more than 6000. It can also be seen that the number of sellers also change day by day and all some sellers cannot be suggested because in all days total numbers of sellers is less than 184 which is the number of different sellers.

2.b) Maximum, minimum, and average prices of products, their buy-box and shipping.

Table 4, 5 and 6: Maximum, Minimum and Average Product Prices

	pid	maxPrice
	<chr>	<dbl>
1	B002ZV00J0	1122.41
2	B0083H1INK	1271.56
3	B00AMFLZLG	852.47
4	B00DNSO10W	951.43
5	B00DNSO41M	1171.72
6	B00MVVI1FC	1018.94
7	B00VSIT5UE	999
8	B00VSITBJ0	1259
9	B00YR6BMS2	1633.6

	pid	minPrice
	<chr>	<dbl>
1	B002ZV00J0	163.8
2	B0083H1INK	209.99
3	B00AMFLZLG	259.99
4	B00DNSO10W	88.39
5	B00DNSO41M	197
6	B00MVVI1FC	74.99
7	B00VSIT5UE	975.71
8	B00VSITBJ0	793.01
9	B00YR6BMS2	700

	pid	averagePrice
	<chr>	<dbl>
1	B002ZV00J0	282.3
2	B0083H1INK	331.9
3	B00AMFLZLG	317.9
4	B00DNSO10W	159.9
5	B00DNSO41M	370.3
6	B00MVVI1FC	181.5
7	B00VSIT5UE	997.4
8	B00VSITBJ0	1253.
9	B00YR6BMS2	997.1

Table 7, 8 and 9: Maximum, Minimum and Average Buy-Box Prices

	pid	maxBboxPrice
	<chr>	<dbl>
1	B002ZV00J0	255
2	B0083H1INK	269.99
3	B00AMFLZLG	259.99
4	B00DNSO10W	88.39
5	B00DNSO41M	320.36
6	B00MVVI1FC	199.99
7	B00VSIT5UE	1029
8	B00VSITBJ0	1259
9	B00YR6BMS2	999.99

	pid	minBboxPrice
	<chr>	<dbl>
1	B002ZV00J0	163.8
2	B0083H1INK	209.99
3	B00AMFLZLG	259.99
4	B00DNSO10W	88.39
5	B00DNSO41M	199.64
6	B00MVVI1FC	74.99
7	B00VSIT5UE	979.17
8	B00VSITBJ0	1245
9	B00YR6BMS2	799

	pid	averageBboxPrice
	<chr>	<dbl>
1	B002ZV00J0	221.6
2	B0083H1INK	255.1
3	B00AMFLZLG	260.0
4	B00DNSO10W	88.39
5	B00DNSO41M	291.4
6	B00MVVI1FC	86.96
7	B00VSIT5UE	999.5
8	B00VSITBJ0	1257.
9	B00YR6BMS2	815.8

Table 10, 11 and 12: Maximum, Minimum and Average Shipping Prices

	pid	maxShippingPrice
	<chr>	<dbl>
1	B002ZV00J0	NA
2	B0083H1INK	NA
3	B00AMFLZLG	42.74
4	B00DNSO10W	25.15
5	B00DNSO41M	46.64
6	B00MVVI1FC	59.17
7	B00VSIT5UE	13.28
8	B00VSITBJ0	13.17
9	B00YR6BMS2	37.7

	pid	minShippingPrice
	<chr>	<dbl>
1	B002ZV00J0	NA
2	B0083H1INK	NA
3	B00AMFLZLG	0
4	B00DNSO10W	0
5	B00DNSO41M	0
6	B00MVVI1FC	0
7	B00VSIT5UE	0
8	B00VSITBJ0	0
9	B00YR6BMS2	0

	pid	averageShippingPrice
	<chr>	<dbl>
1	B002ZV00J0	NA
2	B0083H1INK	NA
3	B00AMFLZLG	1.366
4	B00DNSO10W	2.262
5	B00DNSO41M	5.047
6	B00MVVI1FC	6.131
7	B00VSIT5UE	0.4051
8	B00VSITBJ0	1.190
9	B00YR6BMS2	4.683

Related information is given in the tables. For two products there is no information about shipping. Therefore, we got NA values and for each product there is a free shipping option. These options come from amazon itself as seller. Also, the minimum of the box prices is also the minimum price of corresponding product. It can be seen that providing a product as cheap as possible gives an advantage to the sellers to win the buy-box. The reverse also can be true. When the maximum of prices are examined, the differences between maximum of the buy-box prices and the maximum price of each item are high. Therefore, we can conclude that setting the prices too high, can lead to sellers losing the buy-box. The product "B00VSITBJO" can be exceptional case for this assumption.

2.c) Seller Ratings, Positive Feedbacks and Counts, Product Ratings and Counts

In this section related metrics will be examined by plotting them.

Table 13: sid_rating versus bbox

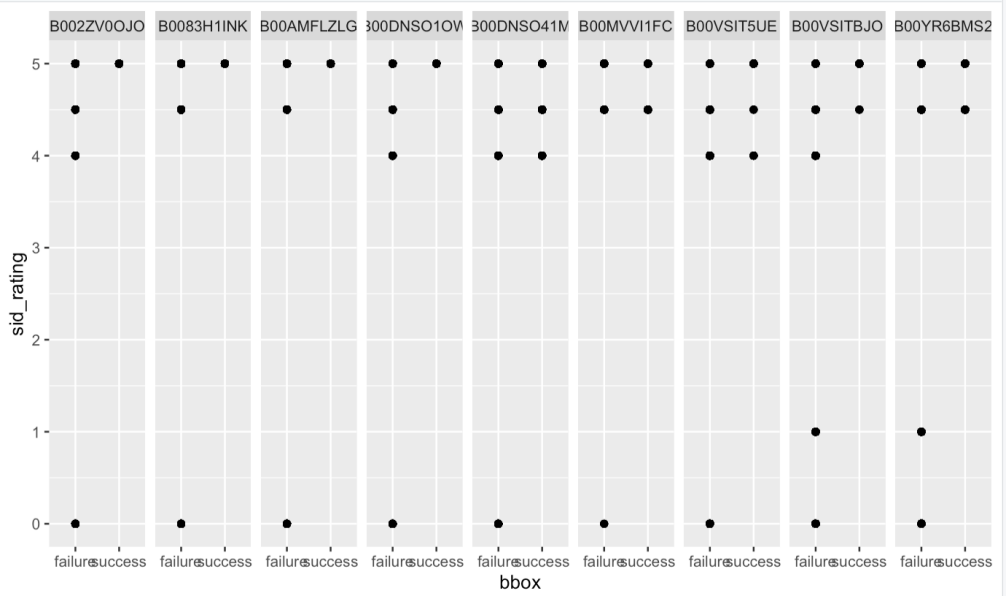


Table 14: sid_pos_fb versus bbox

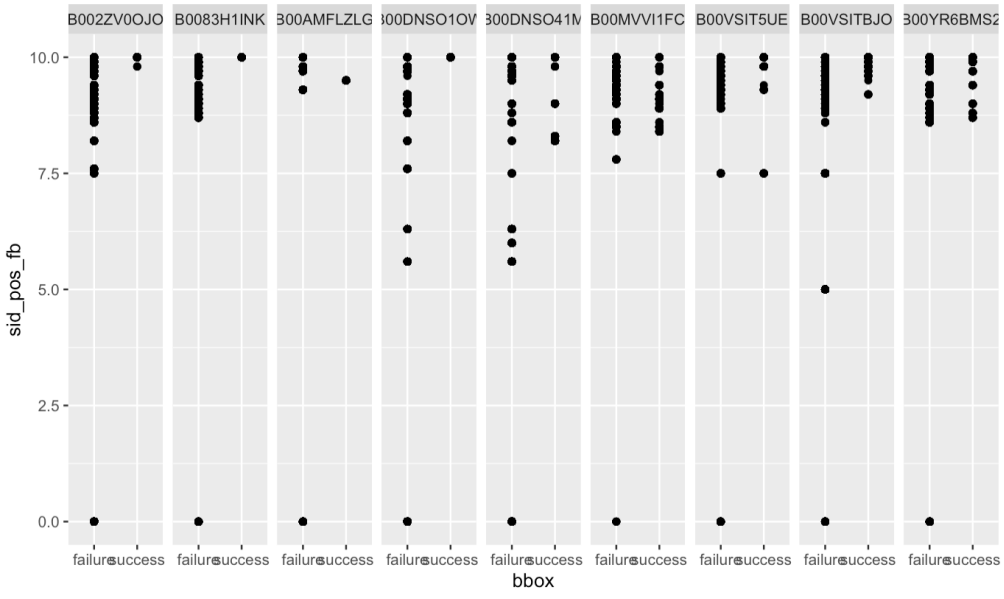


Table 15: sid_rating_cnt versus bbox

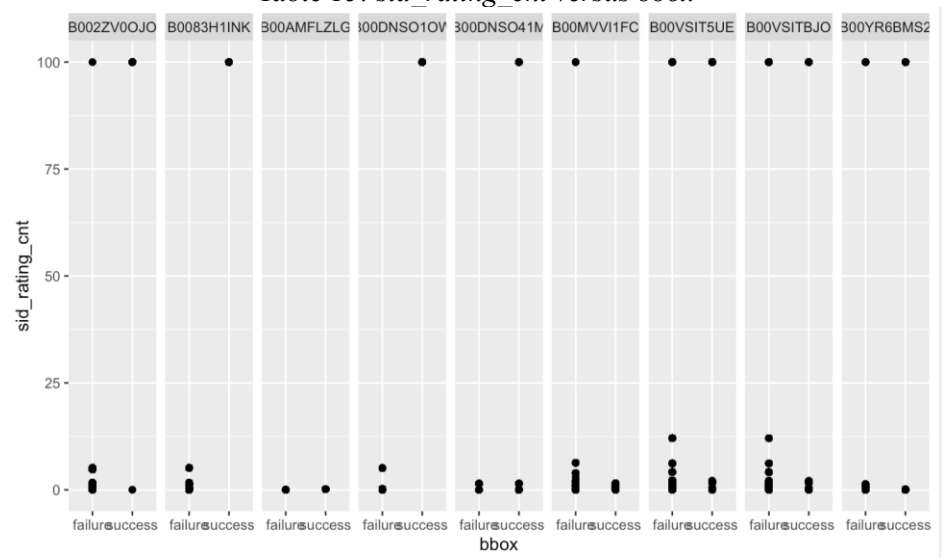


Table 16: pid_rating versus bbox

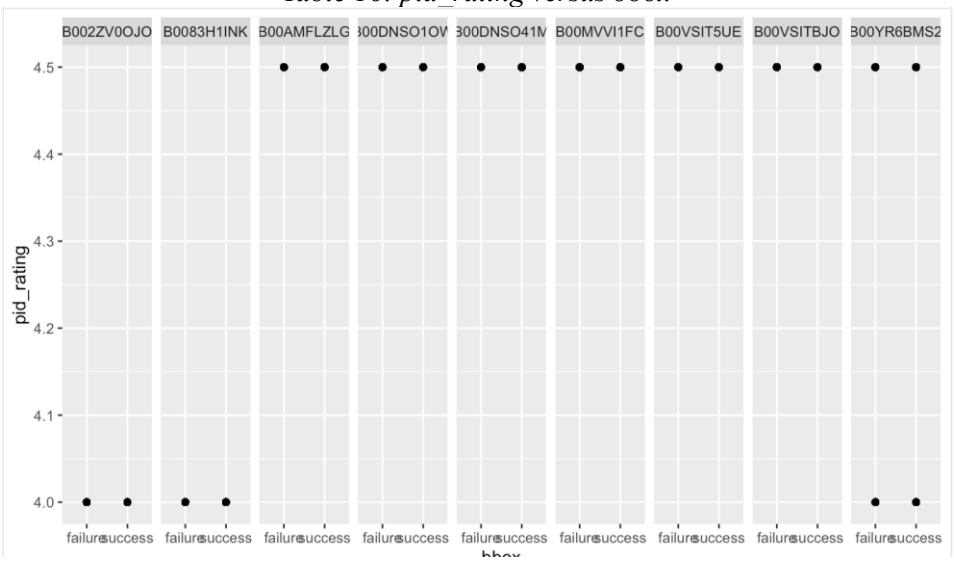
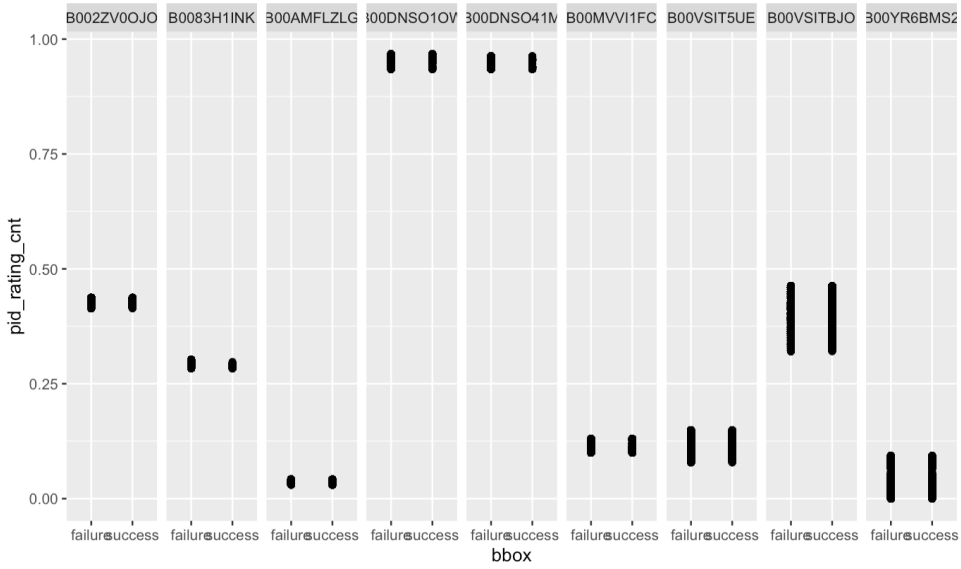


Table 17: pid_rating_cnt



As it can be seen from Table 13, for all products, the seller must have sid_rating more than 4 to win the buy-box. Also, the minimum sid_pos_fb is 7.5 (Table 14). The sellers below 7.5 could not win the buy-box for all products. This finding will later be used in the modelling part.

For all product types sid_rating_cnt do not follow.a pattern (Table 15). Generally, they are low, and the high amounts do not provide sellers to win the buy-box. Here, amazon itself is an exception. Its sid_rating_cnt is 100 where nobody else has this value.

All pid_ratings are above 4 (Table 16). The pid_rating amounts do not change for product except the last one. This type of product has two different values. Additionally, there seems no relation between bbox and pid_rating. It seems logical because probably sellers are ranked not by the product but other metrices. Similar comments can be made also for pid_rating_cnt (Table 17). There is no relation between bbox and generally they are less than 0.5.

2.d) Percentage of buy-box successes when the amazon seller

When the amazon is the seller buy-box successes percentage is **92.63%.** For each product this amount can be changed.

Figure 1

	pid	totalSale	totalSucces	Perc
1	B002ZV0OJO	1105	1102	99.72851
2	B0083H1INK	956	956	100.00000
3	B00DNSO1OW	1108	1108	100.00000
4	B00DNSO41M	123	123	100.00000
5	B00MVVI1FC	8	NA	NA
6	B00VSIT5UE	680	490	72.05882
7	B00VSITBJO	1072	859	80.13060
8	B00YR6BMS2	802	785	97.88030

Generally, Amazon has big advantage over other sellers. However, there are surprising outcomes. Amazon do not sell the product “B00AMFLZLG” and it has no successes in the product “B00MVVI1FC” (it gives NA value because of the count() function in R). Therefore, selling these items can give advantage to the sellers.

Figure 2

	week	totalSale	totalSucces	Perc
1	32	395	360	91.13924
2	33	1997	1887	94.49174
3	34	2012	1877	93.29026
4	35	1450	1299	89.58621

As it can be seen that each week, amazon has huge success percentage. These weeks represent the i^{th} week of the year for the date in data.

2.e) Prime, FBA, Page and Rank information by seller and product

Here, because of the size of the dataset some plots will be provided related to given metrices.

Figure 4

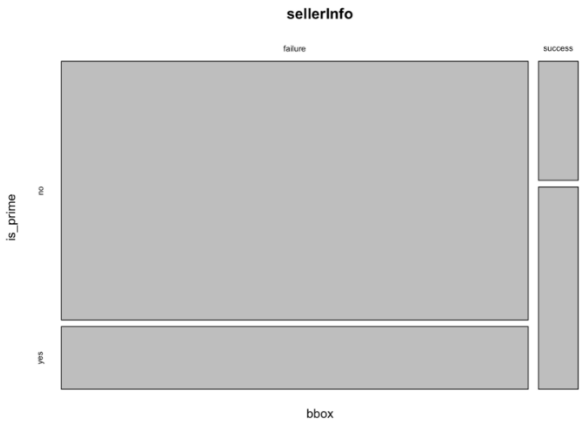
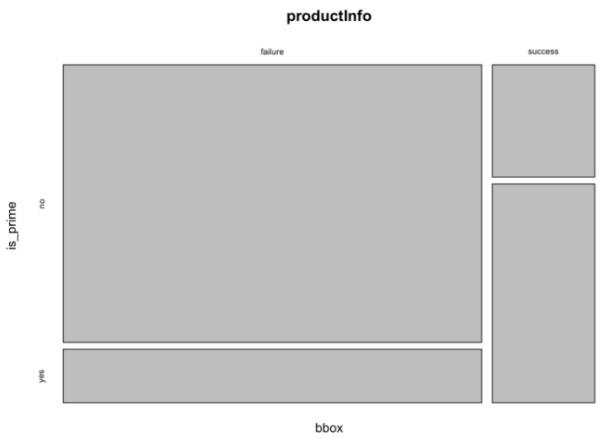


Figure 5



For sellers and products, when the Amazon prime option is available, success numbers increase. Same comment can also be made for is_fba:

Figure 6

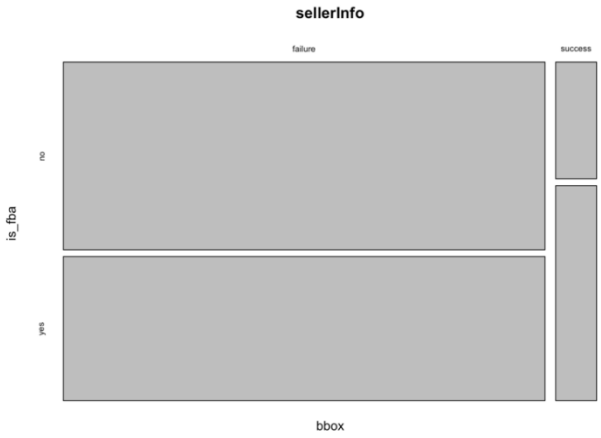


Figure 7

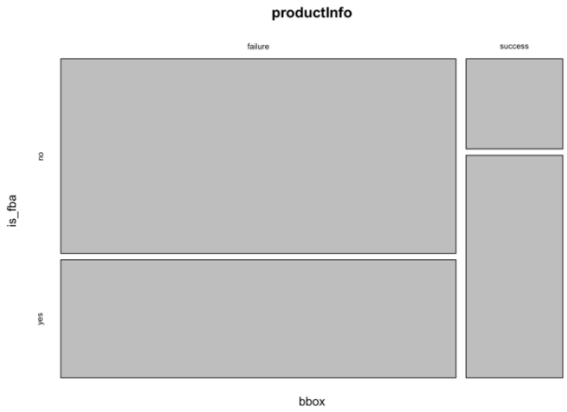


Figure 8

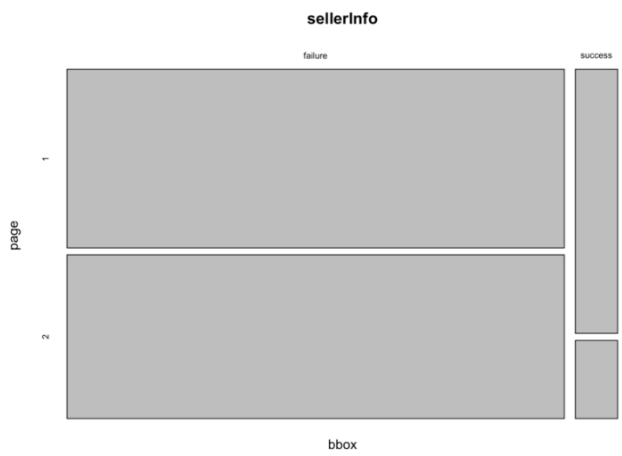


Figure 9

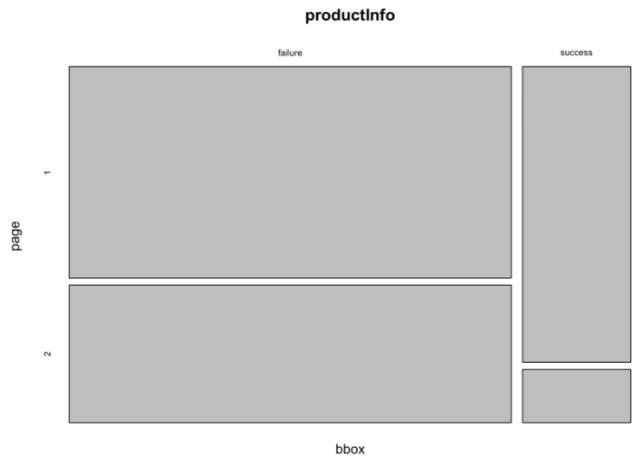


Figure 10

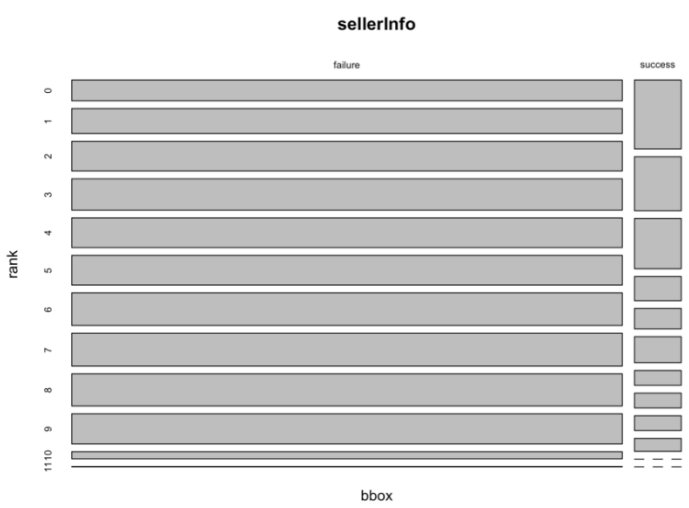
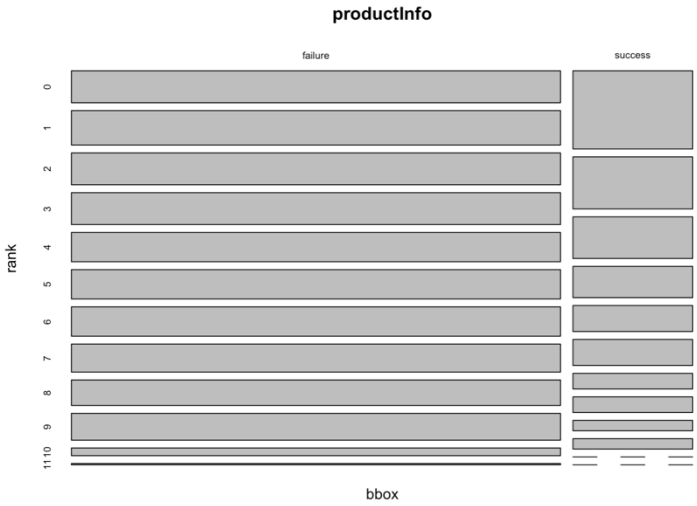


Figure 11

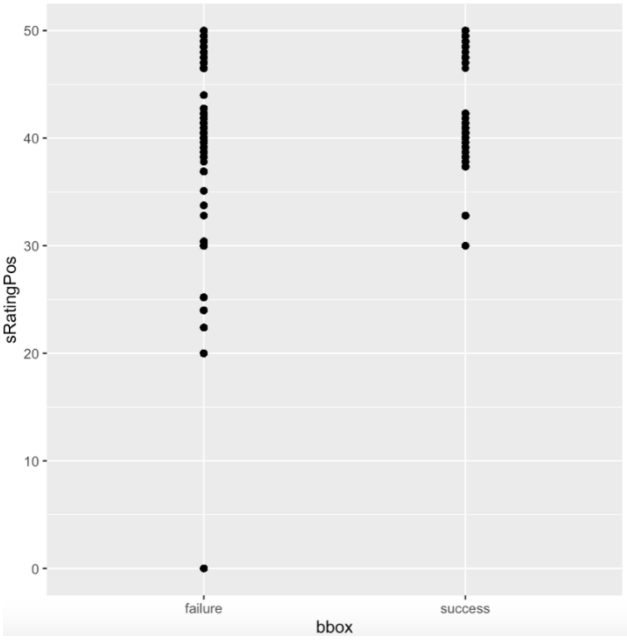


Page and rank also have similar behavior. As the seller/product becomes in front, winning the buy box increases.

2.f) Other Descriptive Statistics

It has been tried to examine the effects of new 5 variables. These variables are PoS which is equal to sid_rating/price; sRatingPos which is equal to sid_rating*sid_pos_fb PidOP which is equal to pid_rating/price. By defining these variables, it is aimed to reveal the effects of ratings clearly. It seemed that pid_values do not significantly affect the winning of buy-box. From graphs it can be seen that there is not much relationship. It can only be said that for the extreme high values for PoS and PidOP values there is only success for buy-box. A seller who has as high as possible values can win the buy-box. Also, for sRatingPos there is a threshold for sellers. Sellers win the buy-box if they have high sRatingPos values. Sellers who have sRatingPos values below 30 do not win the buy-box.

Figrue 12



Two additional columns are added to the data frame. One is the sid_pos_fbIndicator. It is a threshold on sid_pos_fb. As it was told before, there is no seller with sid_pos_fb below 7.5 who wins the buy-box. Therefore, this outcome added as indicator variable to the data frame. Last but not least, an additional column related to price is added. This new column is IsBestPrice. This column shows the transaction with the lowest price. It compares the same products on the same day and hour and gives the transaction with the lowest price, the number 1. This is more like a competitor analysis column. The percentage of winning the buy-box with best price is estimated as 50%. This column can have too much information behind it.

2.g) Scales of data

price	ratio
sid_rating	interval
sid_rating_cnt	interval
shipping	ratio
page	ordinal
rank	ordinal
pid_rating	interval
pid_rating_cnt	interval
is_fba	nominal
Is_prime	nominal
bbox_price	ratio

2.h) Outliers in the data

It seems that some data can be considered as outliers concerning some variables. For example, for sRatingPos we mentioned about the threshold is 30. Below 30 there is no success but there are few data points closing to the value 30. Similar conclusions can also be made for PoS. Generally, there seems to be no exact relationship between buy-box and PoS but for larger values of PoS it seems buy-box is success. However, there are few data points here and can be thought of as outliers.

Figure 13

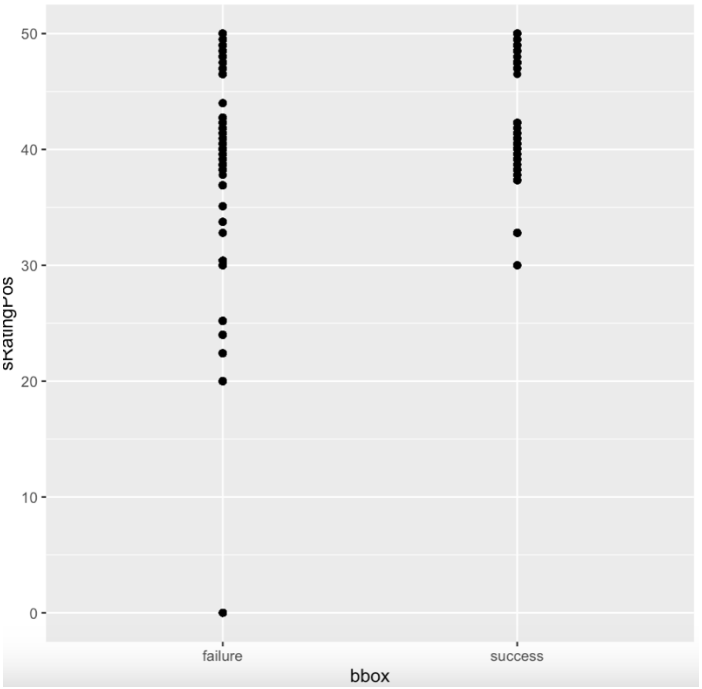
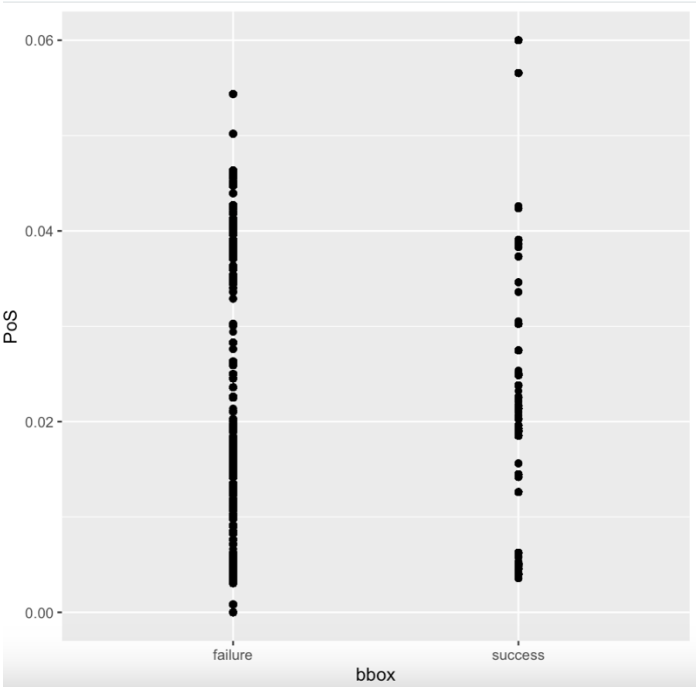
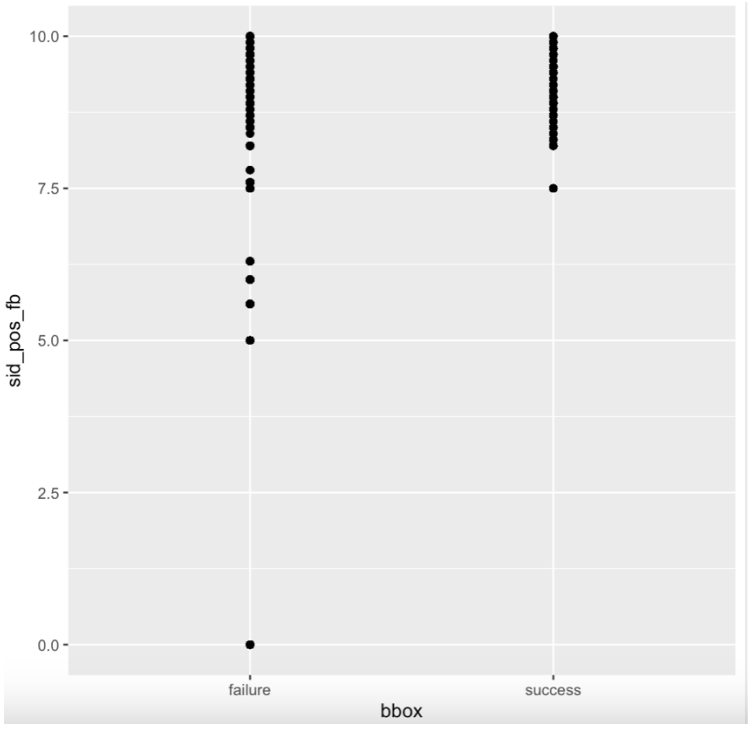


Figure 14



Similarly, sid_pos_fb shows similar behavior. Below 7.5 there is no success but there are a few data points towards 7.5.

Figure 15



Despite these findings, these data are not removed from the dataset. It is thought that, there may be some other relations between other variables and these variables together with these extreme points. However, in the dataset there are some NA values. In these experiments, rows with NA values are discarded from the data and model is built on that.

3. How can a Seller win the buy-box?

It seems that when the amazon itself is the seller, it has a huge advantage. The probability that it will win the buy-box is very large compared to other sellers. Therefore, it can be a good idea for sellers to sell products for which the amazon itself is not the seller.

Additionally, the price has a huge impact on the buy-box algorithm. The column IsBestPrice shows that 50% of the successes come from the best price. Therefore, a seller should investigate its competitors carefully and try to give a better price compared to others. In some cases, seller beats Amazon just a tiny margin on prices.

The ratings related to sellers such as sid_rating, sid_pos_fb, and sid_rating_cnt increase the probability of winning the buy-box algorithm. For example, there is no seller which has sid_pos_fb below 7.5 winning the buy-box. Also, the winning sellers has sid_ratings higher than 4.5. To indicate these effects new columns are added to the data frame and outcomes of these columns are examined above. Generally, it can be said that sellers should try to increase their metrics related to customer satisfaction.

Lastly, fba and prime can have importance. In some observations, a seller wins the buy-box even if its metrices are close to others win the buy-box because it has fba and prime. Therefore, a seller can try to obtain these features.

4. Modelling Section

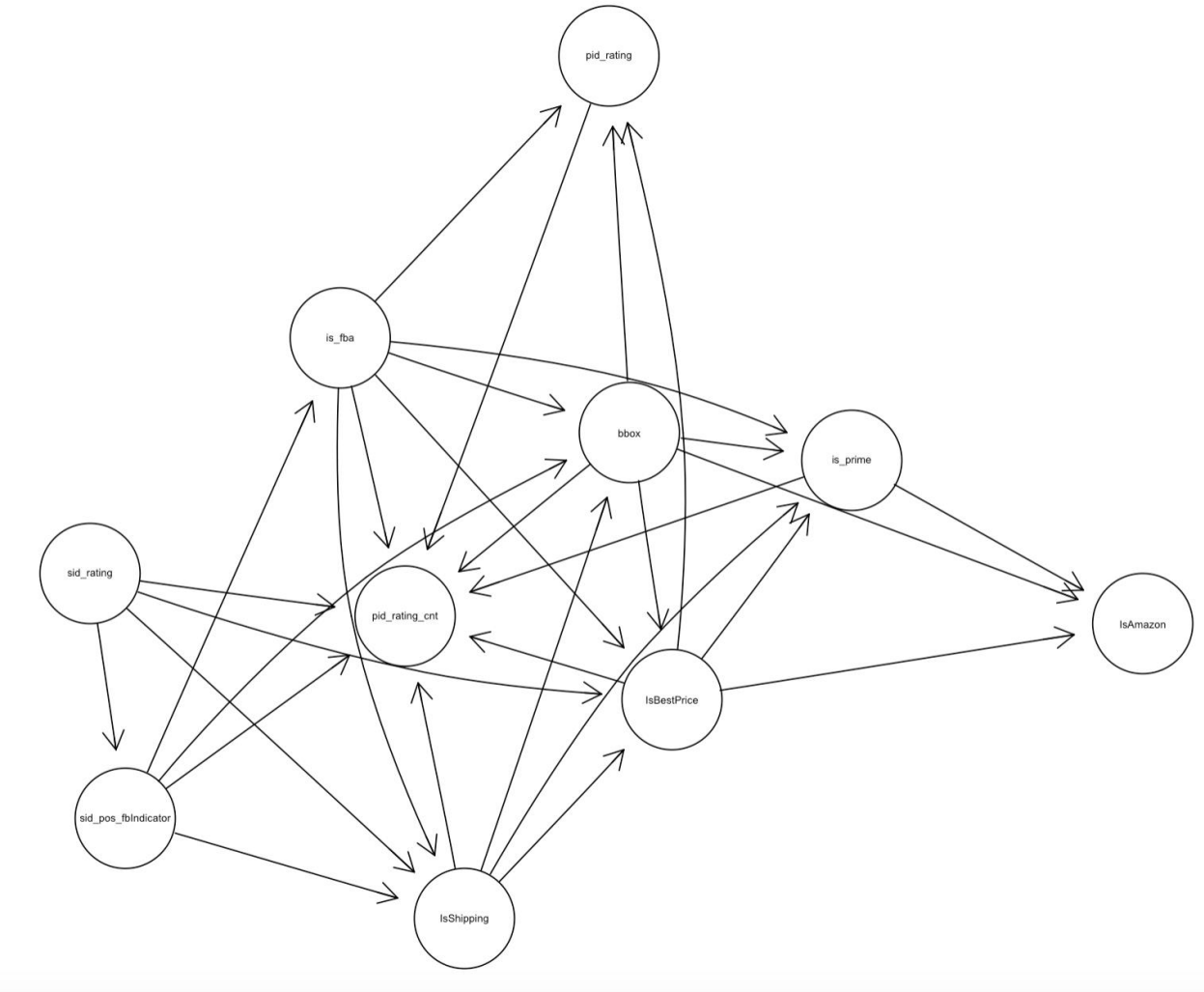
In this section different models from score-based, constraint-based and hybrid algorithms are used. In this section generally, trial and error procedure is used. Firstly, some columns like pid, sid, bboxprice, bbox sid etc. are removed from the training data because it is thought that they are irrelevant to the buy-box. Then, models are started to be constructed. Firstly, score-based algorithms are tried. Here, the algorithm hc from bnlearn package is mainly used. After every model construction the total score of the model is examined and some columns are removed, or some columns are adjusted to increase the score of the algorithm in an iterative way. In these iterations some columns are discretized, meaning that values are grouped as 1 or 0 respecting some values. A total of 8 different models are tried in score-based algorithm and the one with the highest score, -169.399, is selected.

After score-based algorithms, 4 constraint-based algorithms and 2 hybrid algorithms is tried. Two of the constraint-based algorithms whitelist is used because one node was not corresponded to nobody. To compare these different types of models, cross validation is used, and it seems that the hc algorithm gives the least expected loss. Therefore, for the prediction the hc algorithm is used. The final dataset and the DAG corresponding to this model is given as follows:

Figure 16

	sid_rating	pid_rating	pid_rating_cnt	is_fba	is_prime	bbox	sid_pos_fbIndicator	IsBestPrice	IsShipping	IsAmazon
1	1	4	0.4139535	no	no	failure	1	0	0	0
3	1	4	0.4139535	no	no	failure	1	0	0	0
4	1	4	0.4139535	no	no	failure	1	0	1	0
5	0	4	0.4139535	no	no	failure	0	1	0	0
6	1	4	0.4139535	no	no	failure	1	0	0	0

Figure 17



After the model is selected, bootstrapping is applied to the model. After the bootstrapping the score of the algorithm is measured but its score was below the model selected. Therefore, for the prediction stage the model with no bootstrapped is used. Next step is the learning the parameters of the selected model. For this purpose bn.fit() function from bnlearn package is used and parameters are learned.

The probability that amazon wins the buy-box is calculated as 92.56%. Here, to compute that probability cqquery() function from bnlearn is used. As event bbox == “success” and as evidence IsAmazon == 1 is given. To compute that probability logic sampling procedure is used. As can be noticed, this value is close to the value that is found earlier in this study.

After learning the parameters, predictions on test data are made. Before making predictions, some manipulations on test date are made and it is made have the same structure as the model data. Here, the predict() function from bnlearn is used to predict the values of bbox. However, because of the parameters of the pid_rating and pid_rating_cnt this function always gives “failure” outcome. Therefore, the predictions are made by hand.

Each row of the test data is examined via cpquery to find the probability of the bbox == “success”. If the probability is larger than 0.5 then the result is assigned as success.

To estimate the accuracy of the prediction, some metrices are used. 93.31% of the predictions are correct. They are the same as the test data. Also, our predictions give good predictions to find the probability of Amazon winning the buy-box algorithm. 99.34% of the predictions are made correctly to find the winner of the buy-box is Amazon.

However, there are also some disadvantages of the model. When the sid is Amazon, 60% of the predictions are correct. Possibly the model gives more weight to the Amazon sellers to win the buy-box. Additionally, when the seller is not amazon, 11.53% of the predictions are correct. This probability is a little bit low. Models are bad at predicting the status of the non-Amazon sellers.

These steps and manipulations on the datasets are shown in the code file. Considering the length of this report they are not fully included here. Every step and the results are shown in the code file by comment lines.

5. Conclusion

Winning the buy-box clearly depends on multiple criteria. Obviously, the price plays an important role. When the seller provides a price as low as possible, its chance of winning the buy-box increases. Besides that, Amazon gives big importance to customer satisfaction. Therefore, ratings related to customer satisfaction should be tried to increase by sellers. Lastly, zero shipping and fba and prime options carry big importance. In some observations some sellers win the buy-box algorithm even if they do not provide the cheapest price.

For this kind of larger datasets containing many observations and many variables, using DAGs and BNs gives many advantages. Firstly, computational advantages play an important role. Using bnlearn package many actions can be taken from the datasets and predictions can be made fast. Also, this kind of dataset has variables that are dependent on each other’s. Belief networks are good and appropriate tools to reveal these kinds of dependency structure and dependent probabilities. Therefore, using BNs and especially bnlearn gives huge advantage to solve the problem.