BOĞAZIÇI UNIVERSITY

DECISION ANALYSIS
IE 544

ASSIGNMENT 3

How to Win Amazon's Buy Box?

Authors:
Mert Sarıkaya
Y. Harun Kıvrıl
13 June 2021

Introduction

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.

In this assignment, we are going to work with data collected from Amazon.com and try to understand the important factors in winning Amazon's Buy Box.

1 Descriptive Analysis

Before modelling anything it is important to describe the data in hand and make observations that helps to construct and judge models later on.
In this assignment the Amazon BuyBox dataset is used. This dataset is stored in two parts: train and test. Train data has 17 columns and 130278 observations and the test has 41854 observations with the same columns. Train data has observations from 11-08-2015 to 02-09-2015 and the test data includes observations from the period between 13-09-2015 and 21-09-2015.

1.1 Number of Sellers and Products

There are **9** different products in the data and there are **184** unique sellers for them. Also the number of sellers for each product and its daily change is desired and following plots are obtained.

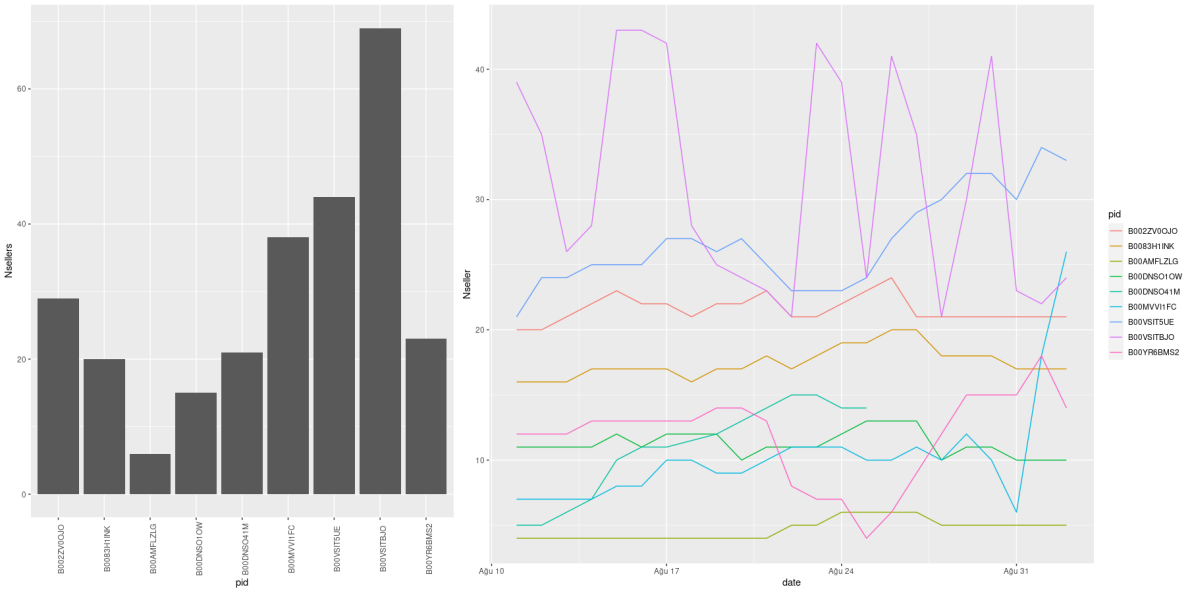

Figure 1: Number of sellers by product and by date and product

This plots show that the number of sellers can be very different for each product. This may introduce different competition characteristics for each product. Also, the number of competitors are not constant at each day in addition the deviations are not the same for each product. For example B00VSITBJO can have about 20 more sellers than previous day while B002ZV0OJO gains/losses maximum 3 sellers in consequent days. In addition, it is observable from the daily plot that B00DNSO41M has no records after 25th August.

1.2 Product Prices and Shipping

The price of a product is one of the most important metric in an e-commerce environment. It is expected from price to have an impact on winning the buy box.

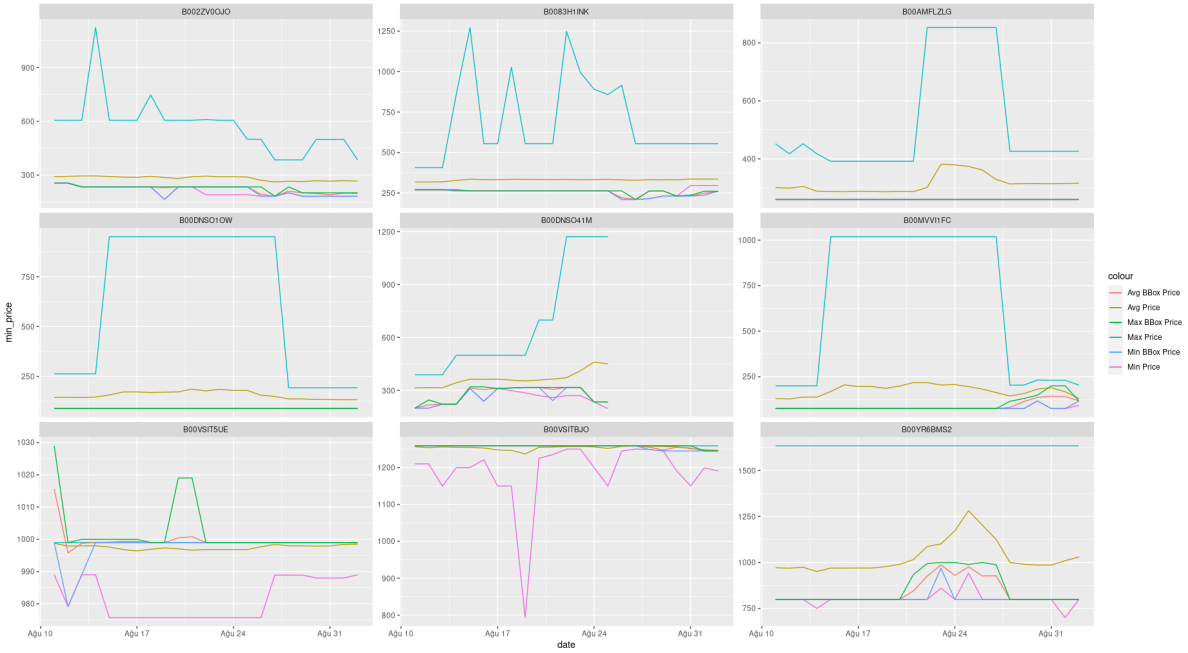

Figure 2: Min, max, avg price and buy box price by product

This plot shows that for every product the buy box price is close to the minimum price but offering the lowest price does not guarantee wining the buy box. It is also possible to see there are some outlier prices that puts the maximum price to an unrealistic level. In addition it is possible to see that buy box price can fluctuate through days and even inside the same day. It is also strange to see that the minimum price is not always at the bottom sometimes minimum buy box price appears less than the minumum price which should be impossible. This means that some entries which wins bbox is missing. Maybe there are other missing entries however it is not possible to be sure.

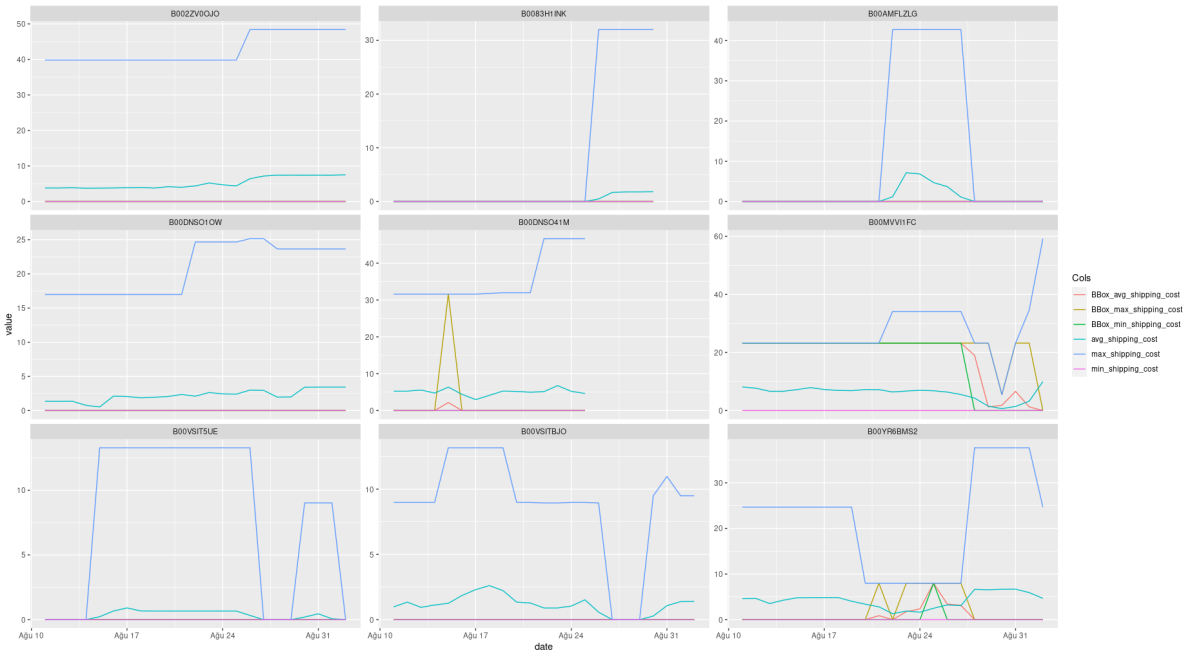

Figure 3: Min, max, avg price and buy box shipping by product

For the majority of the products the average buy box shipping is free. For two products this fact changes for a couple of days and for B00MVVI1FC surprisingly it follows the max shipping cost in the majority of the dates. This may indicate the different characteristics of the products.

1.3 Ratings, Positive Feed backs and Rating Counts

Table below shows us the average, minimum, maximum, bbox_min and bbox_max ratings of sellers for each products. We can see that the sellers' ratings are distributed between 0 and 5, and they are most of the time skewed to 5 side (except for B00AMFLZLG). From this table we can observe that stores with sid_rating ¡ 4 cannot win BBox, and there is probably a positive correlation between sid_rating and BBox winning probability. 5 point corresponds for Amazon itself.

Table 1: Seller Ratings

	pid	avg_sid_rating	max_sid_rating	min_sid_rating	BBox_avg_sid_rating	BBox_max_sid_rating	BBox_min_sid_rating
1	B002ZV0OJO	4.46	5.00	0.00	5.00	5.00	5.00
2	B0083H1INK	4.52	5.00	0.00	5.00	5.00	5.00
3	B00AMFLZLG	2.68	5.00	0.00	5.00	5.00	5.00
4	B00DNSO1OW	4.49	5.00	0.00	5.00	5.00	5.00
5	B00DNSO41M	4.20	5.00	0.00	4.51	5.00	4.00
6	B00MVVI1FC	4.77	5.00	0.00	4.53	5.00	4.50
7	B00VSIT5UE	4.87	5.00	0.00	4.98	5.00	4.00
8	B00VSITBJO	4.46	5.00	0.00	5.00	5.00	4.50
9	B00YR6BMS2	4.63	5.00	0.00	4.94	5.00	4.50

For the positive feedback case, situation is also similar to the above sid rating case. There was

no entity of winning bbox case where positive feedback of seller is less than 8.20. But we cannot conclude that the seller with maximum positive feedback rate wins the bbox, because there are sellers whose positive feedback rates are less than the maximum positive feedback rate in the race. 10 point corresponds for Amazon itself.

Table 2: Positive Feedbacks

	pid	avg_sid_pfb	max_sid_pfb	min_sid_pfb	BBox_avg_sid_pfb	BBox_max_sid_pfb	BBox_min_sid_pfb
1	B002ZV0OJO	8.80	10.00	0.00	10.00	10.00	9.80
2	B0083H1INK	8.95	10.00	0.00	10.00	10.00	10.00
3	B00AMFLZLG	5.29	10.00	0.00	9.50	9.50	9.50
4	B00DNSO1OW	8.87	10.00	0.00	10.00	10.00	10.00
5	B00DNSO41M	8.04	10.00	0.00	8.95	10.00	8.20
6	B00MVVI1FC	9.57	10.00	0.00	8.95	10.00	8.40
7	B00VSIT5UE	9.60	10.00	0.00	9.85	10.00	7.50
8	B00VSITBJO	8.73	10.00	0.00	9.96	10.00	9.20
9	B00YR6BMS2	9.16	10.00	0.00	9.86	10.00	8.70

Below rating counts are scaled between 0 and 100. 100 corresponds for Amazon itself. This table is relatively harder to comment on. But for two products, B0083H1INK and B00DNSO1OW, min and max rating counts are 100, so for these products Amazon always wins the competition.

Table 3: Rating Counts

	pid	avg_sid_cnt	max_sid_cnt	min_sid_cnt	BBox_avg_sid_cnt	BBox_max_sid_cnt	BBox_min_sid_cnt
1	B002ZV0OJO	5.98	100.00	0.00	99.73	100.00	0.02
2	B0083H1INK	5.70	100.00	0.00	100.00	100.00	100.00
3	B00AMFLZLG	0.03	0.14	0.00	0.14	0.14	0.14
4	B00DNSO1OW	8.16	100.00	0.00	100.00	100.00	100.00
5	B00DNSO41M	2.50	100.00	0.00	26.39	100.00	0.00
6	B00MVVI1FC	0.25	100.00	0.00	0.62	1.57	0.00
7	B00VSIT5UE	3.70	100.00	0.00	48.53	100.00	0.00
8	B00VSITBJO	5.44	100.00	0.00	79.11	100.00	0.00
9	B00YR6BMS2	6.53	100.00	0.00	77.35	100.00	0.00

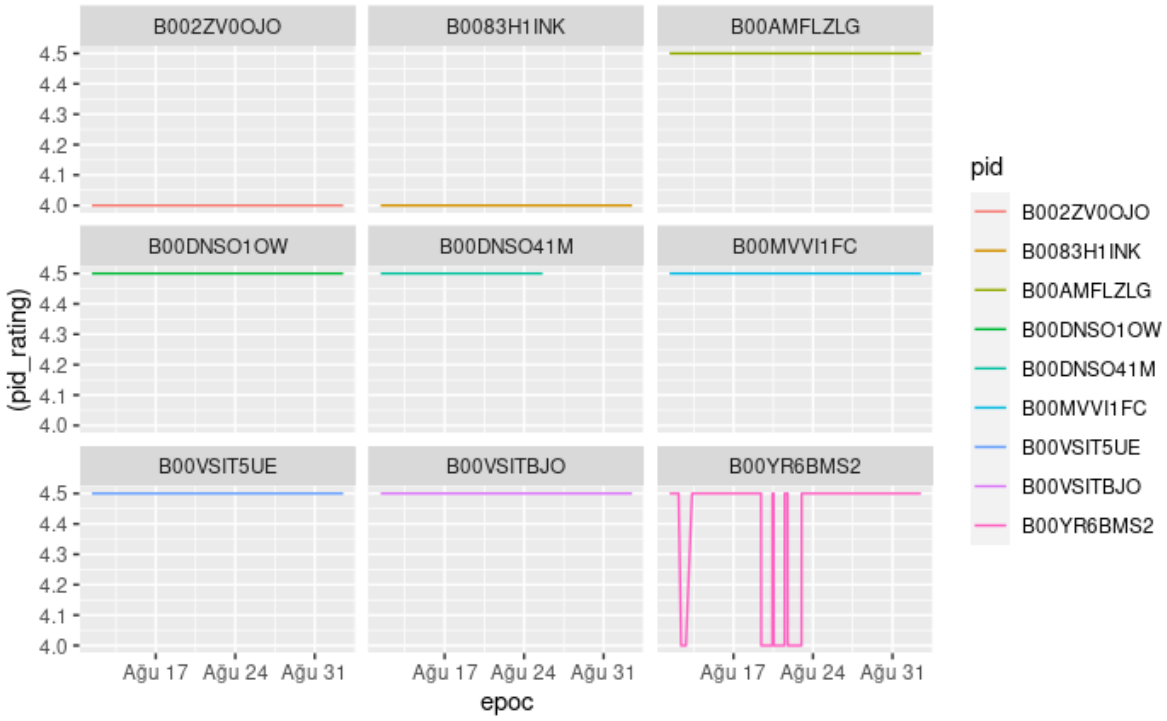

Figure 4: Rating by product and date

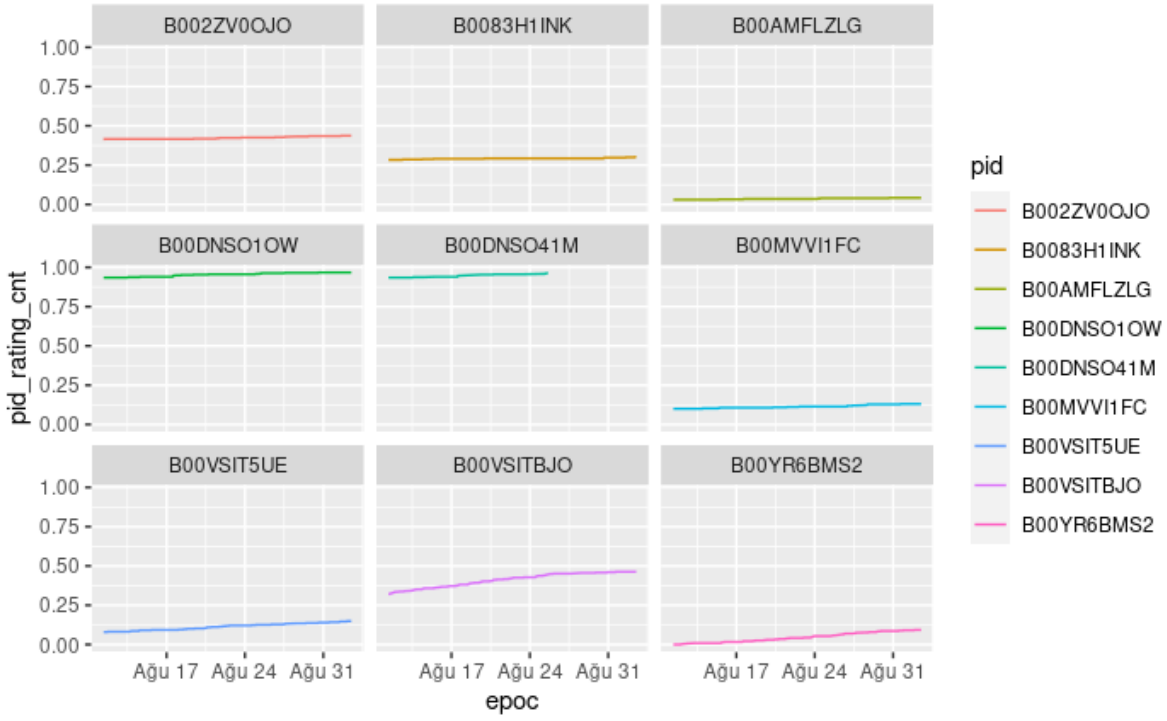

Figure 5: Rating count by product and date

1.4 Amazon on Buy Box

The owner of the platform is also a seller, therefore there is a high chance that it favor itself in the buy box competition. It appears that amazon wins **92.64%** of the buy boxes that it competes. This is a huge ratio and it means that the statistics that is given before for bbox winners are mainly amazon's statistics. In this competition being amazon seems like an incredible boost. When the ratios of each product and weekly ratios are investigated, following plots appear.

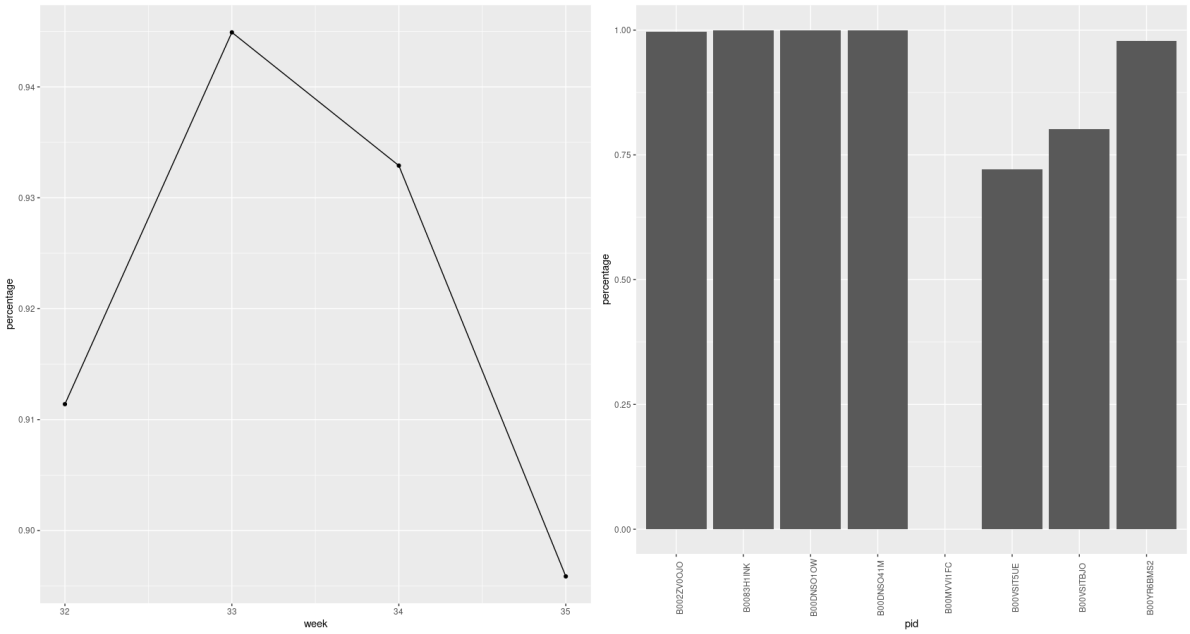

Figure 6: Weekly and by product buy box win ratio of amazon

It appears that amazon winning the buy box of the four products all the time. For one product amazon is not a seller and for one product amazon losses all(8) the competitions which it participates.

1.5 Prime, FBA, Page and Rank

These four information also may be helpful for wining the buybox. Prime means the seller is a part of amazon prime program. FBA means fullfilled by amazon and means that amazon is handling the shipping process. It is important to mention that all sellers have prime also has fba property. In order to understand its effect in buy box wins, prime and fba of all sellers and buy box winner sellers are compared.

Table 4: Prime And FBA Ratios

	pid	prime_ratio	fba_ratio	BBox_prime_ratio	BBox_fba_ratio
1	B002ZV0OJO	0.05	0.05	1.00	1.00
2	B0083H1INK	0.05	0.05	1.00	1.00
3	B00AMFLZLG	0.00	0.00	0.00	0.00
4	B00DNSO1OW	0.08	0.08	1.00	1.00
5	B00DNSO41M	0.02	0.02	0.26	0.26
6	B00MVVI1FC	0.00	0.00	0.00	0.00
7	B00VSIT5UE	0.18	0.48	1.00	1.00
8	B00VSITBJO	0.53	1.00	1.00	1.00
9	B00YR6BMS2	0.12	0.52	0.79	0.80

In this table it seems like having prime and fba is always helpful. However, amazon always have these two properties as seller and it wins the 92% buy boxes. Then in order to distinguish whether this effect is coming from being the platform owner or not, the ratios of the observations where amazon is not winner is checked.

Table 5: Prime and FBA ratios when Amazon did not win IS THIS CORRECT?

	pid	prime_ratio	fba_ratio	N_Competition	BBox_prime_ratio	BBox_fba_ratio
1	B002ZV0OJO	0.06	0.06	3	0.00	0.00
2	B00AMFLZLG	0.00	0.00	1100	0.00	0.00
3	B00DNSO41M	0.00	0.00	351	0.00	0.00
4	B00MVVI1FC	0.00	0.00	1105	0.00	0.00
5	B00VSIT5UE	0.19	0.49	613	0.99	1.00
6	B00VSITBJO	0.53	1.00	243	1.00	1.00
7	B00YR6BMS2	0.03	0.09	235	0.08	0.10

Previously, it has seen that amazon wins all the B002ZV0OJO and B00DNSO41M buy boxes. This table indicates for these two there are some cases that amazon was not an participant. When the ratios are checked if having the prime seems to work for the last three products and for others it is hard to tell something since the ratios are too small. For fba B00VSIT5UE sellers seem to get advantage.

For the rank it is observed that amazon can occupy different places therefore making an analysis by excluding amazon is not necessary. When the win ratios wrt. ranks are plotted it is noticed that products have different patterns.

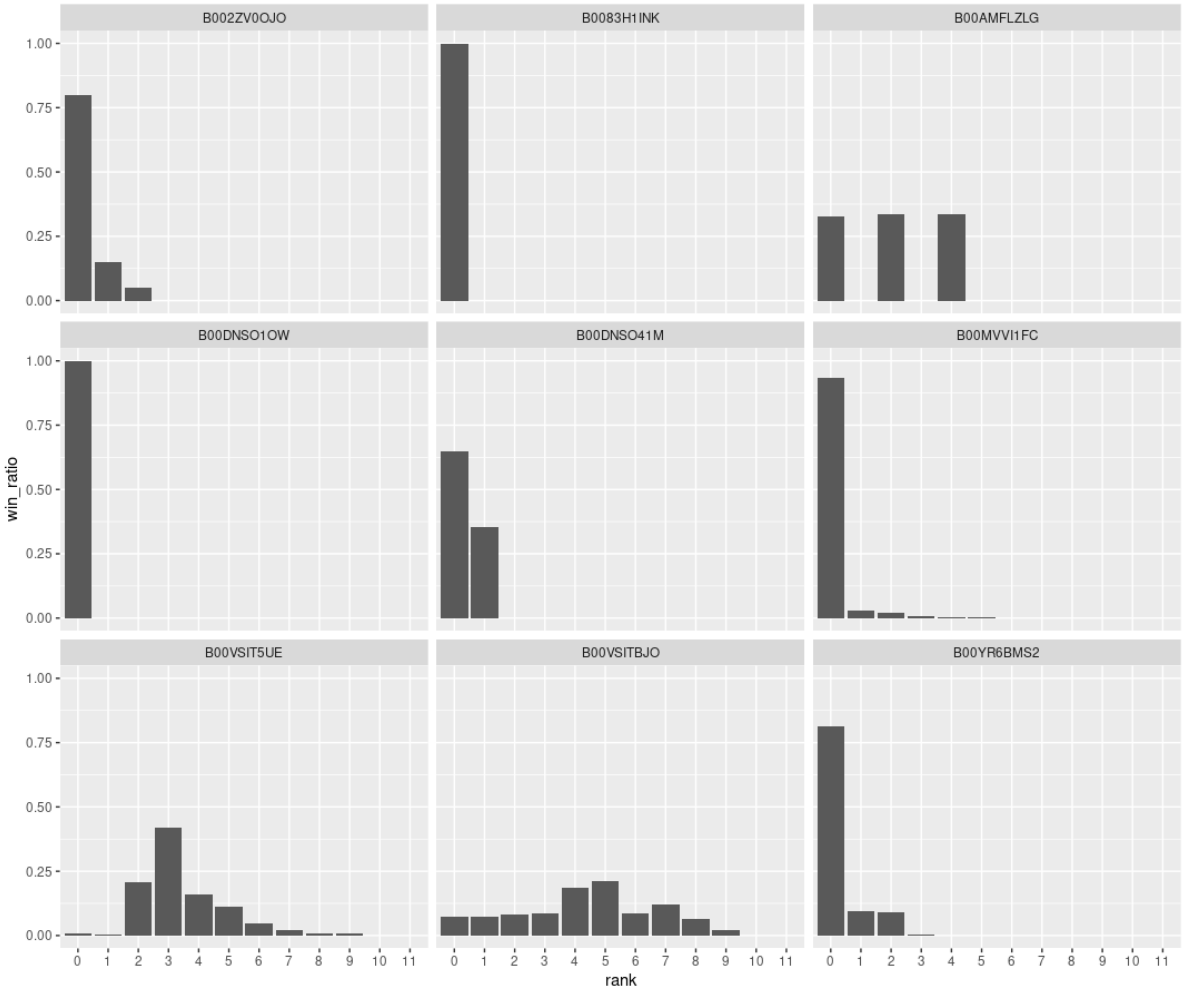

Figure 7: Win ratio of ranks by product

Therefore, it is concluded that rank is important for winning the buy box and it behaves very different for each product.

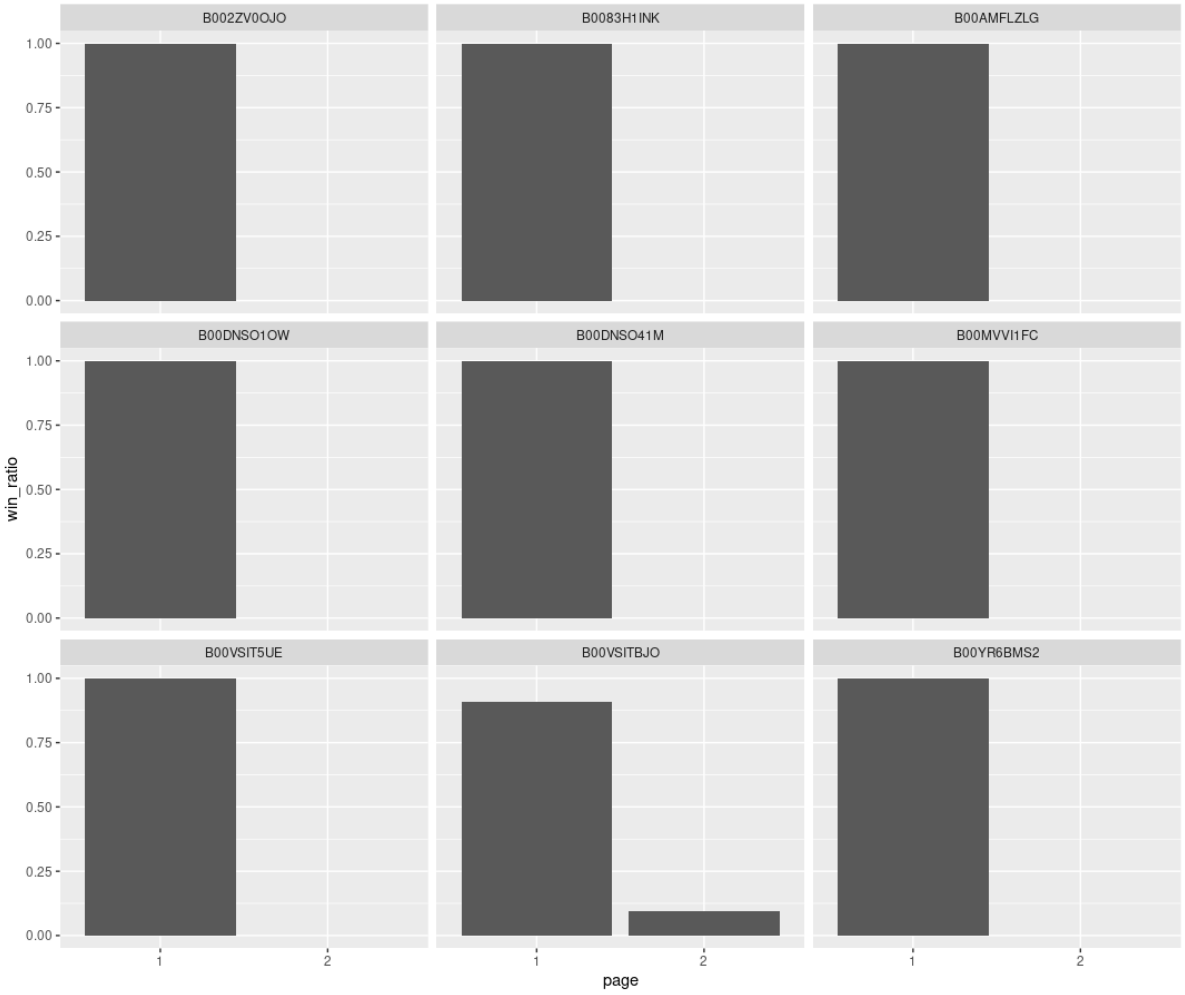

Figure 8: Win ratio of pages by product

Also for the page it seems being in the first is a huge advantage.

1.6 Returning to Rating, Rating Count and Positive Feedback

The observation about amazon and its effect on FBA and Prime analysis crated some questions about the observed effects of rating, rating count and positive feedback.Also it is noticed that amazon has the highest rating, rating count and positive feedback all the time. Therefore it is decided to make analysis for the cases where amazon is not winner of the buy box.

	pid	avg_sid_rating	max_sid_rating	min_sid_rating	N_Competitions	BBox_avg_sid_rating	BBox_max_sid_rating	BBox_min_sid_rating
1	B002ZV0OJO	4.63	5.00	0.00	3	5.00	5.00	5.00
2	B00AMFLZLG	2.68	5.00	0.00	1100	5.00	5.00	5.00
3	B00DNSO41M	4.14	5.00	0.00	351	4.33	5.00	4.00
4	B00MVVI1FC	4.77	5.00	0.00	1105	4.53	5.00	4.50
5	B00VSIT5UE	4.85	5.00	0.00	613	4.96	5.00	4.00
6	B00VSITBJO	4.46	5.00	0.00	243	4.98	5.00	4.50
7	B00YR6BMS2	4.37	5.00	0.00	235	4.72	5.00	4.50

It is good to see that amazon is not the only one carrying buy box ratings to a higher position.

	pid	avg_sid_pfb	max_sid_pfb	min_sid_pfb	N_Competitions	BBox_avg_sid_pfb	BBox_max_sid_pfb	BBox_min_sid_pfb
1	B002ZV0OJO	9.14	10.00	0.00	3	9.80	9.80	9.80
2	B00AMFLZLG	5.29	10.00	0.00	1100	9.50	9.50	9.50
3	B00DNSO41M	7.97	10.00	0.00	351	8.58	10.00	8.20
4	B00MVVI1FC	9.57	10.00	0.00	1105	8.95	10.00	8.40
5	B00VSIT5UE	9.55	10.00	0.00	613	9.72	10.00	7.50
6	B00VSITBJO	8.75	10.00	0.00	243	9.82	10.00	9.20
7	B00YR6BMS2	8.63	10.00	0.00	235	9.39	9.90	8.70

The same situation applies to the positive feedback which is actually very correlated to rating.

	pid	avg_sid_cnt	max_sid_cnt	min_sid_cnt	N_Competitions	BBox_avg_sid_cnt	BBox_max_sid_cnt	BBox_min_sid_cnt
1	B002ZV0OJO	6.40	100.00	0.00	3	0.02	0.02	0.02
2	B00AMFLZLG	0.03	0.14	0.00	1100	0.14	0.14	0.14
3	B00DNSO41M	0.05	1.46	0.00	351	0.59	1.46	0.00
4	B00MVVI1FC	0.25	100.00	0.00	1105	0.62	1.57	0.00
5	B00VSIT5UE	2.27	100.00	0.00	613	1.57	2.08	0.00
6	B00VSITBJO	4.89	100.00	0.00	243	0.76	2.08	0.00
7	B00YR6BMS2	1.35	100.00	0.00	235	0.04	0.15	0.00

However for the counts the situation changes. It appears that having more than average count does not help much. This make sense since the bad ratings are also in the count and it seems like people checking rating and positives to give decision.

1.7 Other Descriptive Statistics and New Features

1.7.1 Percentage Deviation from Min Price

When a shopping event is considered, a customer tries to get the same product as cheap as possible and it looks like a factor that amazon considers in its buy box algorithm. However price is alone not enough. In the shopping event the customer only sees the offers that are available in that time and it compares them relatively. Therefore a relative feature about is considered, First of all this price should be relative to the min price and the min price should be the minimum price available at that time. Also, there is having a 1$ product for 2$ and having 100$ product for 101$ has different effects so the new feature should be dependent on the scale of the prices. As a result the following statistic is defined to be used instead of raw price.

$$\textit{Percentage Deviation from Min Price} = \frac{\textit{Price offered by seller for the product}}{\textit{Min price given for the product at that time}} - 1$$

In addition, in order to keep the models simpler it is assumed that the customers evaluates prices with the shipping and shipping cost is added to the price and another version of the new statistic is defined:

$$\textit{Percentage Deviation from Min Price+Shipping} = \frac{\textit{Price offered by seller for the product + shipping}}{\textit{Min price + shipping given for the product at that time}} - 1$$

1.7.2 New rank

In order to decrease the number of variables in the models the rank and page information fused together by the fact that rank is special to its page and the maximum rank is 12.

$$\textit{New rank} = 12 * (\textit{Page} - 1) + \textit{Rank}$$

1.7.3 Amazon's Participation

Since amazon has huge effect in the competition, it is interesting to see how many buy box races it participates in each product. If there are products that has some moments that amazon not in the race different features may be gain different importance in the absence of it.

	pid	total_races	amazon_races	wins	amazon_participation	amazon_win_over_total
1	B002ZV0OJO	1109	1105	1102	1.00	0.99
2	B0083H1INK	1113	956	956	0.86	0.86
3	B00DNSO1OW	1108	1108	1108	1.00	1.00
4	B00DNSO41M	474	123	123	0.26	0.26
5	B00VSIT5UE	1109	680	490	0.61	0.44
6	B00VSITBJO	1102	1070	859	0.97	0.78
7	B00YR6BMS2	1025	802	785	0.78	0.77

As the table shows, there are cases where the probabilities expected to change case in the same product because of amazon's participation status. Therefore it may make sense to build a model for the cases where amazon is a contestant and another model otherwise. Creating a variable for this also seems like a solution however being amazon and amazon is not participating is an impossible event and can not be handled in the same model.

1.8 Scales of Data

	Feature	Type	Scale
1	pid	Nominal	-
2	epoc	Interval	[2015-08-11 09:04:13 , 2015-09-02 12:00:56]
3	sid	Nominal	-
4	Price	Ratio	[74.99,1633.60]
5	sid_rating	Ratio	[0,5]
6	sid_pos_fb	Ratio	[0,10]
7	sid_rating_cnt	Ratio	[0,100]
8	shipping	Ratio	[0,59.17]
9	page	Ordinal	{0,1}
10	rank	Ordinal	{0, ... ,12}
11	pid_rating	Ratio	[0,4.5]
12	pid_rating_cnt	Ratio	[0,0.9674]
13	is_fba	Nominal	{yes, no}
14	is_prime	Nominal	{yes, no}

1.9 Outliers

During the descriptive analysis two things are attracted the attention as outliers. First one is the cases where the buy box price is under the minimum price which should be impossible since buy box winner should be included in the set that min price is calculated. That indicates something is wrong about at that time for that product and they are eliminated. The second thing is about the prices. The maximum prices are too away from the average prices and the sellers that offers this price is assumed to be there because they entered the price wrong or they don't update their prices properly and should not be evaluated as a competitor.

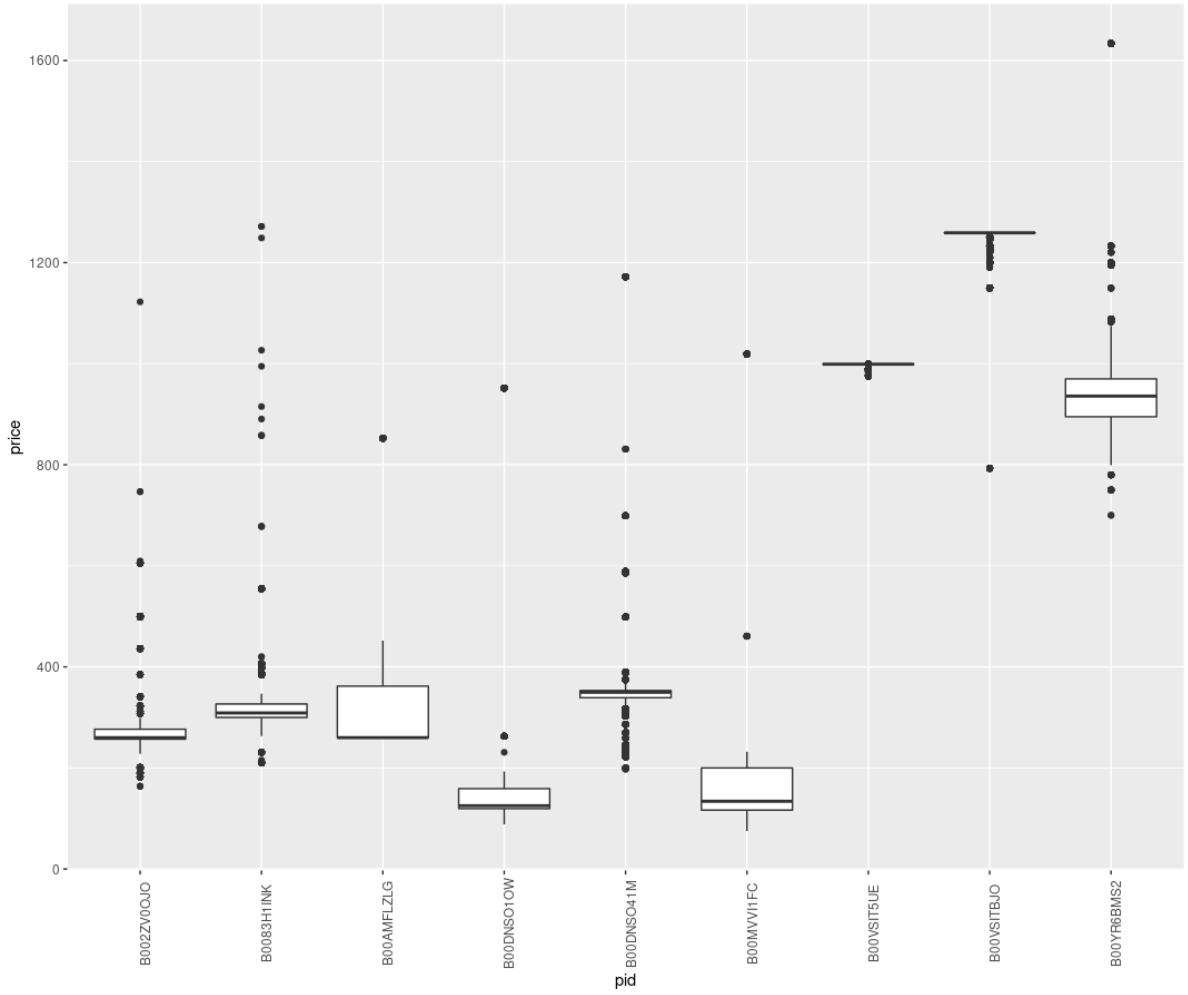

Figure 9: Box plot of prices by product

As the box plot shows, in some products such cases exists and these prices appear too far away from the boxes. In order to not to hurt the training, it is decided to remove these observations from training data. As a rule the observations inside three standard deviation is kept. After removing the outliers, the data is checked again and it is observed that no buy box winner is removed.

1.9.1 Discretization of Existing Features

There are two reasons to discretize the features. First one is it helps to construct a simpler model. The other one is, winning the buy box will be a discrete node in the models and model assumptions does not allow to put any continuous observation above it. Therefore the possible non discrete candidates for the model is gathered into bins. For **% deviation from min** 5 bins with equal observation is constructed and **sid_rating** is divided to 5 bins with intervals $[5, 4.5), [4.5, 4)$, $[4, 3.5), [3.5, 0.5], (0.5, 0]$.

2 Discussion of Descriptive Analysis

During the analysis, some useful information that helps to construct a better model is collected. As a summary:

- The products have different number of sellers and the number of sellers changes in different patterns. This created an expectation that products will have different effects on buy box winning probabilities. Also while investigating the all other metrics different behavior for each product is observed. Thus having the product effect in model is preferable.
- The avg buy box price is very close to the minimum price but not exactly it. It is expected that % deviation from min price will be effective.
- Free shipping is a candidate to model however the effect of free shipping mainly comes from the amazon.
- Rank has a pattern for each product for buy boxes and it can be combined with page to create a single feature which simplifies the models.
- Rating and positive feedback are observed as a useful features for buy box winning however they are highly correlated since positive feed backs contributes to rating calculation, so using one of them is preferable for simplicity.
- Amazon is a game changer. It does not participate in every buy box contest however when it does it wins especially in certain products. Therefore, for a seller other than amazon having amazon in the race have an impact on probability of winning. Considering these facts, having a model where amazon is in the race and another one without amazon is sensible.

3 Learning Structure

We are asked to learn the structure of the DAG with three different methods. As score based method, we used Tabu algorithm. As constraint based, we used Grow-Shrink and we used Max-Min Hill Climbing algorithm as a hybrid method.

We started with using pid, bbox, is_amazon, dc_min_price_dev_perc, dc_sid_rating and new_rank columns in our calculations. We added pid column with a belief of that node may differentiate the characteristic differences of products, like being is_amazon might mean nothing for a product whereas for some products it may bring huge importance. pid, bbox and is_prime columns are used as they exist in the given dataframe. is_amazon, dc_min_price_dev_perc, dc_sid_rating and new_rank

columns are created as we mentioned in **Section 1.7** and **1.9.1**

Before learning the structure the findings from the descriptive analysis is interpreted as black list and white list. For examble the effect of being an amazon to bbox is whitelisted from the beginning as the effect is observed before. **Also in order to obtain more intuitive graph structure with respect to the problem definition direction of arcs are prohibited. Furthermore, improvement in aic values are observed when these** For the model that does not have amazon in competition had the black and white lists without is_amazon feature.

Table 6: Blacklist

	from	to
1	bbox	pid
2	is_amazon	pid
3	dc_min_price_dev_perc	pid
4	new_rank	pid
5	dc_sid_rating	pid
6	bbox	is_amazon
7	bbox	dc_min_price_dev_perc
8	bbox	new_rank
9	bbox	dc_sid_rating
10	dc_min_price_dev_perc	is_amazon
11	new_rank	is_amazon
12	dc_sid_rating	is_amazon

Table 7: Whitelist

	from	to
1	is_amazon	bbox
2	pid	bbox
3	new_rank	bbox

3.1 Model When Amazon Competes

Since, it is observed that having amazon in the competition creates a huge difference and the model assumptions does not allow to construct a model that haves is_amazon and does_amazon_compete at the same time, it is decided to construct two different models.
For the model that amazon competes different feature sets are tried however the model with the the selected features are preferred since its simpler than the others. After that three different dags are learned using Tabu search, Grow-Shrink and Min-Max Hill Climbing Algorithms.

These three dags are compared using cross validation and log-likelihood loss and following table is obtained:

Table 8: Log-Lik Loss of Different Dags for Model with Amazon

	Algorithm	Loss
1	Tabu	4.8549
2	Grow-Shrink	4.8712
3	MMHC	4.9190

As a result following dag from tabu algorithm is decided to be used. With this dag the test data will only use new_rank, is_amazon and pid information since they construct a Markov blanket over bbox. However when the structure is not reduced to these four nodes, one can ask questions about the relationship between other variables and bbox in presence of amazon. So it is decided to keep them.

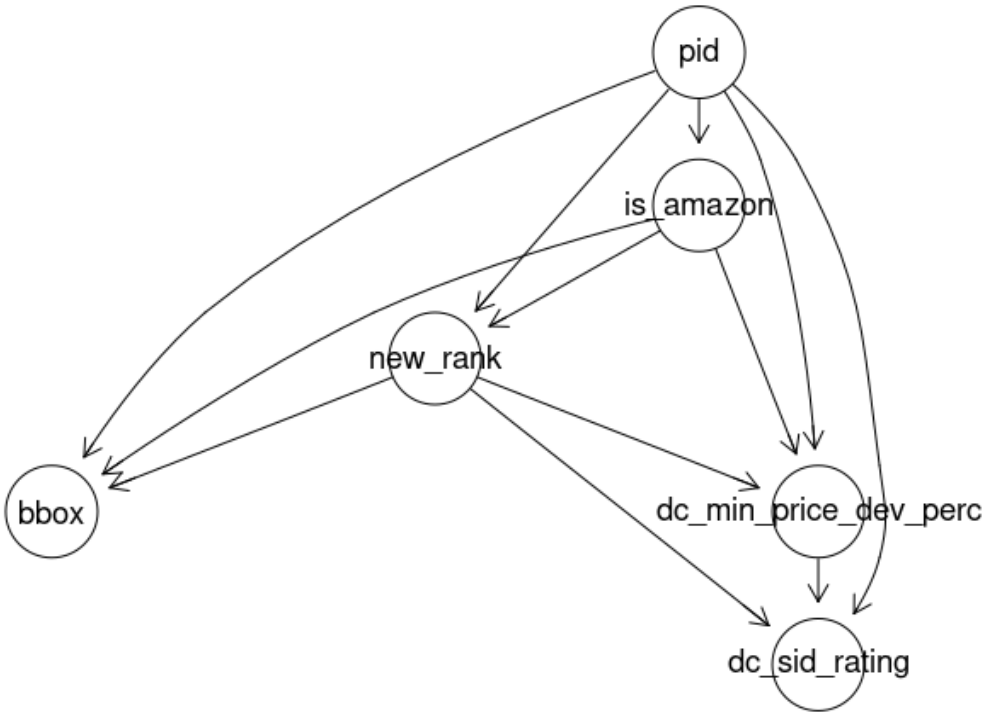

Figure 10: Selected structure for the model that amazon competes

3.2 Model When Amazon does not Compete

When amazon is not inside the race there is no need to have an is_amazon feature therefore it removed from feature set and white/black lists. Then the same procedure is applied by constructing

three different dags are learned using Tabu search, Grow-Shrink and Min-Max Hill Climbing Algorithms.

These three dags are compared using cross validation and log-likelihood loss and following table is obtained:

Table 9: Log-Lik Loss of Different Dags for Model without Amazon

	Algorithm	Loss
1	Tabu	3.9335
2	Grow-Shrink	3.8409
3	MMHC	3.9333

As a result following dag from Grow-Shrink algorithm is decided to be used.

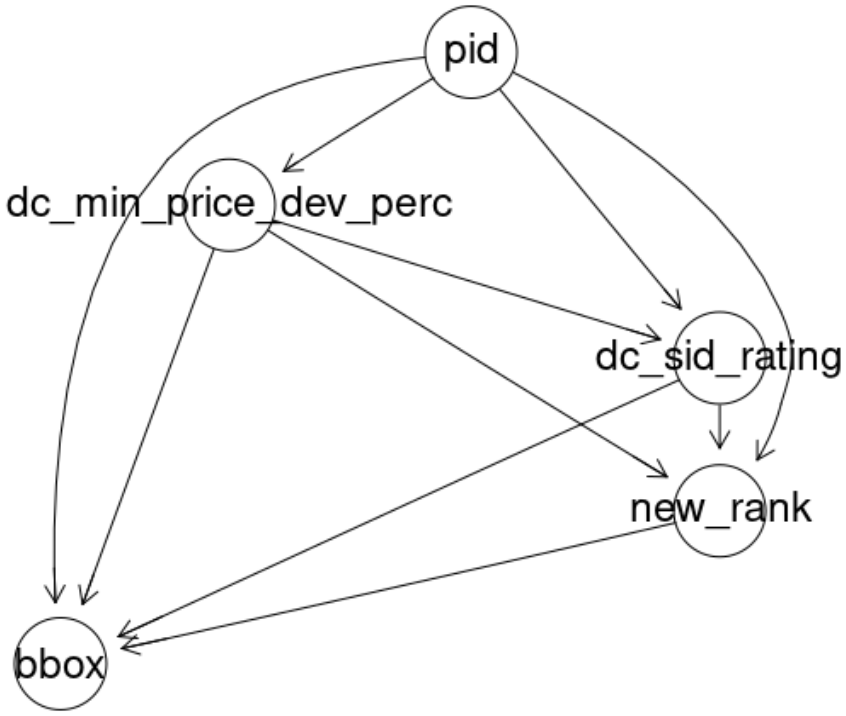

Figure 11: Selected structure for the model that amazon does not competes

4 Learning Parameters

4.1 Parameter Learning

Parameters are learned via bnlearn by using above mentioned models. In the dags there are non observed conditions, for example amazon is not selling two of the products. However it does not

mean that this is impossible. In feature amazon can decide to sell this product too. Therefore, the model that is learned needs to give chance to this event. Mle method does not allow this situation however, Bayes method does. Therefore Bayes method is used for the training. Since parameter space is relatively big, it's not effective to show all the obtained parameters, so it is decided to show a portion of the whole table.

```
  Parameters of node bbox (multinomial distribution)

Conditional probability table:

, , is_amazon = no, new_rank = [1,4]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  9.990364e-01 9.999786e-01 5.000000e-01 9.999833e-01 9.998495e-01 8.317972e-01 9.736842e-01 9.818496e-01 9.929642e-01
  yes 9.635599e-04 2.135839e-05 5.000000e-01 1.671291e-05 1.505117e-04 1.682028e-01 2.631579e-02 1.815038e-02 7.035798e-03

, , is_amazon = yes, new_rank = [1,4]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  2.764931e-03 5.810575e-05 5.000000e-01 5.013537e-05 4.512635e-04 9.931507e-01 2.590013e-01 8.845944e-02 2.126333e-02
  yes 9.972351e-01 9.999419e-01 5.000000e-01 9.999499e-01 9.995487e-01 6.849315e-03 7.409987e-01 9.115406e-01 9.787367e-01

, , is_amazon = no, new_rank = (4,7]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  9.999789e-01 9.999834e-01 5.000000e-01 9.999833e-01 9.998105e-01 8.317972e-01 9.453660e-01 9.980474e-01 9.999765e-01
  yes 2.114701e-05 1.663783e-05 5.000000e-01 1.671291e-05 1.895375e-04 1.682028e-01 5.463401e-02 1.952622e-03 2.349955e-05

, , is_amazon = yes, new_rank = (4,7]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 4.516995e-01 2.019761e-01 5.000000e-01
  yes 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 5.483005e-01 7.980239e-01 5.000000e-01
```

Figure 12: Partial Conditional Probability Table for node bbox given pid

```
  Parameters of node bbox (multinomial distribution)

Conditional probability table:

, , dc_min_price_dev_perc = [0,1.01112], dc_sid_rating = 5, new_rank = [1,4]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  5.000000e-01 5.000000e-01 7.610234e-06 5.000000e-01 2.003330e-01 5.096144e-01 9.753459e-01 9.997586e-01 8.078649e-01
  yes 5.000000e-01 5.000000e-01 9.999924e-01 5.000000e-01 7.996670e-01 4.903856e-01 2.465409e-02 2.414293e-04 1.921351e-01

, , dc_min_price_dev_perc = (1.01112,4.21409], dc_sid_rating = 5, new_rank = [1,4]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 9.997356e-01 9.061795e-01 8.167647e-01 9.998576e-01 6.249422e-01
  yes 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 2.644104e-04 9.382052e-02 1.832353e-01 1.424096e-04 3.750578e-01

, , dc_min_price_dev_perc = (4.21409,16.5475], dc_sid_rating = 5, new_rank = [1,4]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 9.999880e-01 9.998971e-01 5.000000e-01 5.000000e-01 6.718601e-01
  yes 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 1.202472e-05 1.028595e-04 5.000000e-01 5.000000e-01 3.281399e-01

, , dc_min_price_dev_perc = (16.5475,39.5791], dc_sid_rating = 5, new_rank = [1,4]

     pid
bbox    B002ZV0OJO   B0083H1INK   B00AMFLZLG   B00DNSO1OW   B00DNSO41M   B00MVVI1FC   B00VSIT5UE   B00VSITBJO   B00YR6BMS2
  no  5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 9.998457e-01 9.999864e-01 5.000000e-01 5.000000e-01 9.999504e-01
  yes 5.000000e-01 5.000000e-01 5.000000e-01 5.000000e-01 1.542734e-04 1.358290e-05 5.000000e-01 5.000000e-01 4.959825e-05
```

Figure 13: Partial Conditional Probability Table for node bbox given dc_min_price_dev_perc, dc_sid_rating and new_rank

As observed in descriptive analysis when the amazon is in the race, and is_amazon is True the conditional probability of winning is high but changes according the pid. For the non observed products it is filled with a prior belief of 0.5 since bayes method is used.

4.2 Amazon winning probability

Amazon winning probability can be calculated as $P(bbox = \text{"}yes\text{"} | is\_amazon = \text{"}yes\text{"})$. Querying this in the model with logic sampling method gives 0.92. If we were to compare this ratio with the empirical value (what is the winning ratio for Amazon in the training data) they are both around 92% which makes totally sense.

```
cpquery(amz_fit,
        event=(bbox=='yes'),
        evidence=(is_amazon=='yes'),
        method="ls")
```

4.3 Prediction

First and last 5 rows of the test data with their predictions can be seen below.

```
> to_test_amazon
             pid bbox is_amazon dc_min_price_dev_perc dc_sid_rating new_rank prediction               epoc
     1: B002ZV0OJO   no        no     (4.21409,16.5475]             5  (10,16]         no 2015-08-11 09:04:13
     2: B002ZV0OJO   no        no     (16.5475,39.5791]             5  (16,23]         no 2015-08-11 09:04:13
     3: B002ZV0OJO   no        no     (16.5475,39.5791]             5  (16,23]         no 2015-08-11 09:04:13
     4: B002ZV0OJO   no        no           [0,1.01112]             0    [1,4]         no 2015-08-11 09:04:13
     5: B002ZV0OJO   no        no     (16.5475,39.5791]           4.5  (10,16]         no 2015-08-11 09:04:13
    ---
101406: B00YR6BMS2   no        no     (4.21409,16.5475]           4.5    [1,4]         no 2015-09-02 12:00:56
101407: B00YR6BMS2   no        no     (16.5475,39.5791]           4.5   (7,10]         no 2015-09-02 12:00:56
101408: B00YR6BMS2   no        no     (16.5475,39.5791]           4.5  (10,16]         no 2015-09-02 12:00:56
101409: B00YR6BMS2  yes       yes           [0,1.01112]             5    [1,4]        yes 2015-09-02 12:00:56
101410: B00YR6BMS2   no        no     (39.5791,1004.85]             5  (10,16]         no 2015-09-02 12:00:56
```

Figure 14: Prediction in the test data when amazon is in the race

```
> to_test_noamazon
             pid bbox dc_min_price_dev_perc dc_sid_rating new_rank prediction is_amazon               epoc
    1: B00AMFLZLG   no           [0,1.01112]             0    [1,4]         no        no 2015-08-11 09:04:13
    2: B00AMFLZLG   no           [0,1.01112]             0    [1,4]         no        no 2015-08-11 09:04:13
    3: B00AMFLZLG   no     (39.5791,1004.85]           4.5    (4,7]         no        no 2015-08-11 09:04:13
    4: B00AMFLZLG  yes           [0,1.01112]             5    [1,4]        yes        no 2015-08-11 09:04:13
    5: B00AMFLZLG   no           [0,1.01112]             0    [1,4]         no        no 2015-08-11 09:35:05
   ---
28428: B00YR6BMS2   no     (39.5791,1004.85]             5    (4,7]         no        no 2015-08-27 16:55:15
28429: B00YR6BMS2   no     (1.01112,4.21409]             5    [1,4]         no        no 2015-08-27 16:55:15
28430: B00YR6BMS2  yes           [0,1.01112]           4.5    [1,4]        yes        no 2015-08-27 16:55:15
28431: B00YR6BMS2   no           [0,1.01112]           4.5    [1,4]        yes        no 2015-08-27 16:55:15
28432: B00YR6BMS2   no     (16.5475,39.5791]             5    (4,7]         no        no 2015-08-27 16:55:15
```

Figure 15: Prediction in the test data when amazon is not in the race

4.4 Accuracy

Table 10: Prediction Accuracies

	Train Accuracy	Test Accuracy
Amazon competes	0.9921	0.9923
Amazon doesn't compete	0.9719	0.9717
Total	0.9876	0.9877

Prediction accuracy in the test data is 0.9877 and our training accuracy is 0.9876. If Amazon is competing then our test accuracy is 0.9923, and training accuracy is 0.9921. If Amazon is not competing, then test accuracy is 0.9717 and training accuracy is 0.9719. We can see that our training and test accuracies are highly close to each other, this shows that their distributions are quite similar, these two datasets have similar behaviors and the models does not overfit.

```
Confusion Matrix and Statistics


          no    yes
  no  117857    645
  yes    934   8043

               Accuracy : 0.9876
                 95% CI : (0.987, 0.9882)
    No Information Rate : 0.9318
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.904

 Mcnemar's Test P-Value : 0.0000000000004238

            Sensitivity : 0.92576
            Specificity : 0.99214
         Pos Pred Value : 0.89596
         Neg Pred Value : 0.99456
             Prevalence : 0.06815
         Detection Rate : 0.06309
   Detection Prevalence : 0.07042
      Balanced Accuracy : 0.95895

       'Positive' Class : yes
```

Figure 16: Confusion Matrix for the Training Data

```
Confusion Matrix and Statistics


          no    yes
  no  120205    660
  yes    936   8041

               Accuracy : 0.9877
                 95% CI : (0.9871, 0.9883)
    No Information Rate : 0.933
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9031

 Mcnemar's Test P-Value : 0.000000000005835

            Sensitivity : 0.92415
            Specificity : 0.99227
         Pos Pred Value : 0.89573
         Neg Pred Value : 0.99454
             Prevalence : 0.06701
         Detection Rate : 0.06193
   Detection Prevalence : 0.06914
      Balanced Accuracy : 0.95821

       'Positive' Class : yes
```

Figure 17: Confusion Matrix for the Test Data

5 Discussion and Conclusion

In this study, we investigated how to win Amazon's buy box and constructed a belief network to predict the win with the given information. Firstly, we conducted a descriptive analysis and see that amazon favors itself in the buybox decision and gets the most of the buy boxes. Therefore, having amazon in the buybox race has huge impact on the other stores win chance. Knowing that fact it is decided to have two different models and desperate the effect of amazon. One model is constructed for the races where amazon participates and other one is constructed for the races that amazon does not participate. Another factor observed as important is the product id. The plots we showed indicated that the products behaves very differently in terms of other features. Thus, we kept pid in the feature set. Also, our analysis showed that the rank information is an important metric when it is combined with the page information. For the price the deviation from the min is calculated by putting the shipping information into it. Lastly, some variables are discretized for simplicity and not to have problem with model assumptions.

At the end it is decided to keep pid, bbox, is_amazon, dc_min_price_dev_perc, dc_sid_rating and new_rank as features in the models by whitelisting three important features pid, is_amazon and new_rank. For each model three different structures learned and the structures are evaluated using cross validation and the best ones are selected.

As a result of the model that amazon competes. We see that the white listed features are directly connected to bbox and they construct a Markov blanket together. The other features has an effect via the new rank. Even though the test data has these three columns, in real life one should know how to change its new rank to obtain better buy box winnings. In order to win a race where amazon competes, you may want to decrease the price and increase your rating, however amazon already have these conditions in the most of the cases so there is not so much chance but winning when amazon fails may become important to make profits. This model is also validated by the high accuracy in the test data, and results should be reliable.

For the model that amazon does not compete. Again pid is connected to all nodes that shows its importance once again. All the others also directly connected to the bbox. Therefore, in order to win a buy box one should choose which pid to join the race carefully. Also it should be closer to the minimum price and its rating should be high that results better ranks and higher bbox win probabilities. This model is more straight forward since there is no such an effect as amazon. Finally the validation of this model is made in the test data and high accuracy values are obtained.

In the probability tables, we see 0.5 as probabilities since there are no observation about them and bayesian methods fills with a prior. This might be considered as a draw back for this models however not having this product in the period means it is likely to not to have it much and when it does the probabilities can be updated easily with the new data.

While treating pid, we decided that having separate models for each product by separating amazon and no amazon cases gets too complicated. Thus we introduced product id columns in the model however it would be great to see the pros and cons of this approach over creating separate models for each product. So, as a future work, creating separate models for each product id can be considered.

By using Bayesian networks approach we can predict buy box winners with almost 99% probability. Our final models are easy to implement, we can say that it is a white-box model, we can analyze the effects and correlations of the features, and we can force their relations according to our needs. We can experiment with different conditions easily. So we think that this BN approach suits well to this problem and we believe that we generated good predictions for the test period.

References

[1] Radhakrishnan Nagarajan and Marco Scutari. *Bayesian Networks in R with Applications in Systems Biology*. Springer, New York, 2013. ISBN 978-1-4614-6445-7, 978-1-4614-6446-4.

[2] Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[3] Marco Scutari. Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. *Journal of Statistical Software*, 77(2):1–20, 2017.

[4] Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks with Examples in R*. Chapman and Hall, Boca Raton, 2014. ISBN 978-1-4822-2558-7, 978-1-4822-2560-0.