Who is gonna win the Amazon Buy Box?

S. Azzolina, D. Usula

Course: SLLD (Prof. F. Chiaromonte)
PhD Economics - Sant'Anna School of Advanced Studies
14/06/2023

Thanks to Emanuel Weitschek, Manuel Razza,
Roger Voyat, and Sergio Meligrana

Does a product sold and/or shipped by Amazon have a higher probability of winning the BuyBox?

Empirically assessing whether Amazon adopt self-preferencing when competing for the Buy Box against third-party sellers.

Motivation

- Amazon as leader in e-commerce
- Concerns on self-preferencing and imitation among
  - Economic scholars
  - Antitrust practitioners
  - Policy-makers
- AGCM fined Amazon for abuse of dominance position steering third-party sellers towards FBA services (2021)
- Lack of transparency and accountability

Roadmap

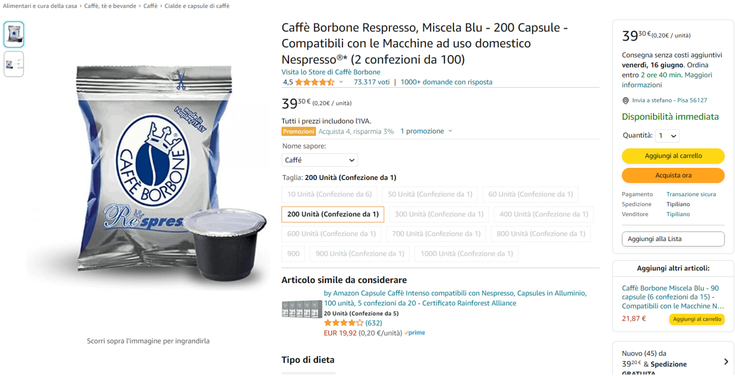
1. Dataset
   - Data Collection
   - Data Cleaning

2. Selection Procedures
   - PCA
   - Stepwise Selection

3. Classification

4. Other directions

The Amazon Buy Box



Alimentari e cura della casa › Caffè, tè e bevande › Caffè › Cialde e capsule di caffè

Scorri sopra l'immagine per ingrandirla

Caffè Borbone Respresso, Miscela Blu - 200 Capsule - Compatibili con le Macchine ad uso domestico Nespresso®* (2 confezioni da 100)

Visita lo Store di Caffè Borbone

4,5 ★★★★☆ 73.317 voti | 1000+ domande con risposta

39^{30} € (0,20€ / unità)

Tutti i prezzi includono l'IVA.

Promozioni Acquista 4, risparmia 3% 1 promozione

Nome sapore: Caffé

Taglia: **200 Unità (Confezione da 1)**

10 Unità (Confezione da 6)
50 Unità (Confezione da 1)
60 Unità (Confezione da 1)
200 Unità (Confezione da 1)
300 Unità (Confezione da 1)
400 Unità (Confezione da 1)
600 Unità (Confezione da 1)
700 Unità (Confezione da 1)
800 Unità (Confezione da 1)
900
900 Unità (Confezione da 1)
1000 Unità (Confezione da 1)

Articolo simile da considerare

by Amazon Capsule Caffè Intenso compatibili con Nespresso, Capsules in Alluminio, 100 unità, 5 confezioni da 20 - Certificato Rainforest Alliance

20 Unità (Confezione da 5)

★★★★☆ (632)

EUR 19,92 (0,20 €/unità) ✓prime

Tipo di dieta

39^{30} €(0,20€ / unità)

Consegna senza costi aggiuntivi **venerdì, 16 giugno**. Ordina entro 2 ore 40 min. Maggiori informazioni

Invia a stefano - Pisa 56127

Disponibilità immediata

Quantità: 1

Pagamento	Transazione sicura
Spedizione	Tipiliano
Venditore	Tipiliano

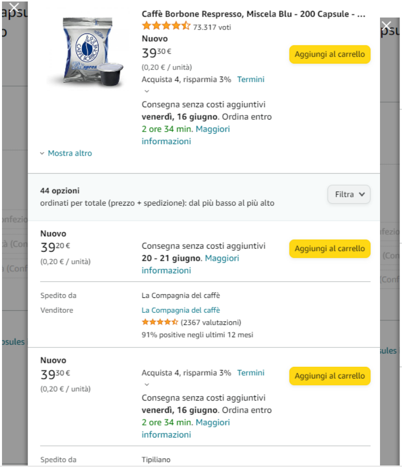
Aggiungi altri articoli:

Caffè Borbone Miscela Blu - 90 capsule (6 confezioni da 15) - Compatibili con le Macchine N...

21,87 €

Nuovo (45) da 39^{20} € & **Spedizione GRATUITA**

The Amazon Buy Box

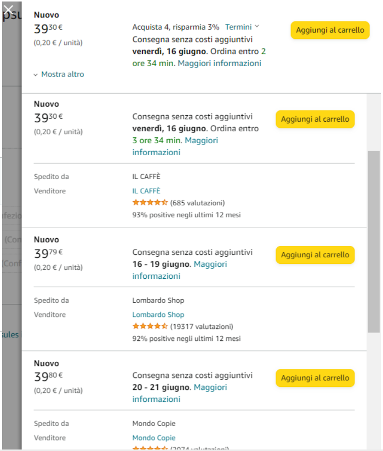


Caffè Borbone Respresso, Miscela Blu - 200 Capsule - ...
★★★★☆ 73.317 voti
Nuovo
39^{30}€
(0,20 € / unità)
Acquista 4, risparmia 3% Termini
Consegna senza costi aggiuntivi **venerdì, 16 giugno**. Ordina entro 2 ore 34 min. Maggiori informazioni
Mostra altro

44 opzioni
ordinati per totale (prezzo + spedizione): dal più basso al più alto

Nuovo
39^{20}€
(0,20 € / unità)
Consegna senza costi aggiuntivi **20 - 21 giugno**. Maggiori informazioni
Spedito da: La Compagnia del caffè
Venditore: La Compagnia del caffè
★★★★☆ (2367 valutazioni)
91% positive negli ultimi 12 mesi

Nuovo
39^{30}€
(0,20 € / unità)
Acquista 4, risparmia 3% Termini
Consegna senza costi aggiuntivi **venerdì, 16 giugno**. Ordina entro 2 ore 34 min. Maggiori informazioni
Spedito da: Tipiliano



Nuovo
39^{30}€
(0,20 € / unità)
Acquista 4, risparmia 3% Termini
Consegna senza costi aggiuntivi **venerdì, 16 giugno**. Ordina entro 2 ore 34 min. Maggiori informazioni
Mostra altro

Nuovo
39^{30}€
(0,20 € / unità)
Consegna senza costi aggiuntivi **venerdì, 16 giugno**. Ordina entro 3 ore 34 min. Maggiori informazioni
Spedito da: IL CAFFÈ
Venditore: IL CAFFÈ
★★★★☆ (685 valutazioni)
93% positive negli ultimi 12 mesi

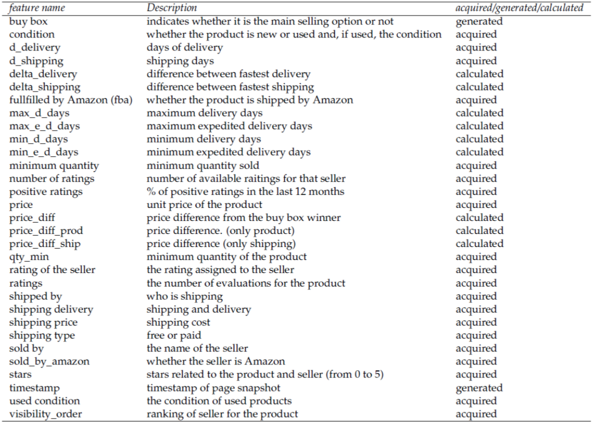
Nuovo
39^{79}€
(0,20 € / unità)
Consegna senza costi aggiuntivi **16 - 19 giugno**. Maggiori informazioni
Spedito da: Lombardo Shop
Venditore: Lombardo Shop
★★★★☆ (19317 valutazioni)
92% positive negli ultimi 12 mesi

Nuovo
39^{80}€
(0,20 € / unità)
Consegna senza costi aggiuntivi **20 - 21 giugno**. Maggiori informazioni
Spedito da: Mondo Copie
Venditore: Mondo Copie
★★★★☆ (2074 valutazioni)

Data Collection

- Develop a web scraping algorithm;
- Collect data on four different products selected among the Italian best-selling categories in the Amazon marketplace:
  - Sport and Leisure: Smartwatch Xiaomi Mi Smart Band 6
  - Food: Coffee capsules
  - Office: Moleskine diary
  - Lighting: Philips light bulbs
- Two different periods of data scraping:
  - February 7 - March 9, 2022 (twice a day)
  - October 20 - November 10, 2022 (once a day)
- Size: 6990 x 32

Initial Dataset



feature name	*Description*	*acquired/generated/calculated*
buy box	indicates whether it is the main selling option or not	generated
condition	whether the product is new or used and, if used, the condition	acquired
d_delivery	days of delivery	acquired
d_shipping	shipping days	acquired
delta_delivery	difference between fastest delivery	calculated
delta_shipping	difference between fastest shipping	calculated
fullfilled by Amazon (fba)	whether the product is shipped by Amazon	acquired
max_d_days	maximum delivery days	calculated
max_e_d_days	maximum expedited delivery days	calculated
min_d_days	minimum delivery days	calculated
min_e_d_days	minimum expedited delivery days	calculated
minimum quantity	minimum quantity sold	acquired
number of ratings	number of available raitings for that seller	acquired
positive ratings	% of positive ratings in the last 12 months	acquired
price	unit price of the product	acquired
price_diff	price difference from the buy box winner	calculated
price_diff_prod	price difference. (only product)	calculated
price_diff_ship	price difference (only shipping)	calculated
qty_min	minimum quantity of the product	acquired
rating of the seller	the rating assigned to the seller	acquired
ratings	the number of evaluations for the product	acquired
shipped by	who is shipping	acquired
shipping delivery	shipping and delivery	acquired
shipping price	shipping cost	acquired
shipping type	free or paid	acquired
sold by	the name of the seller	acquired
sold_by_amazon	whether the seller is Amazon	acquired
stars	stars related to the product and seller (from 0 to 5)	acquired
timestamp	timestamp of page snapshot	generated
used condition	the condition of used products	acquired
visibility_order	ranking of seller for the product	acquired

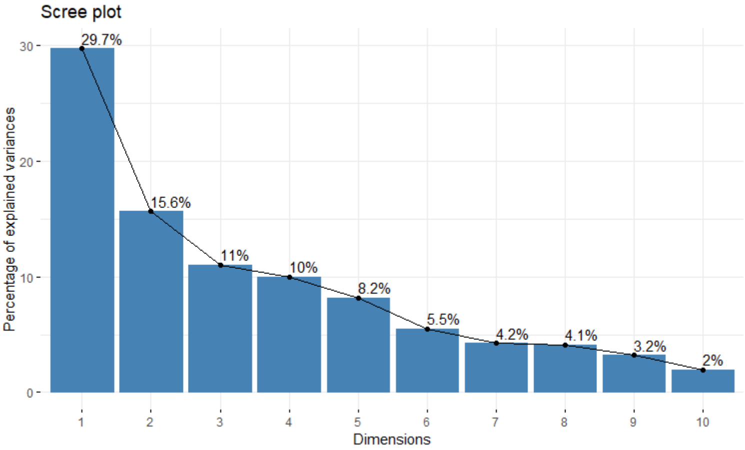
Table 1. Glossary of the features: the considered features in the system.

Getting data ready for the analysis

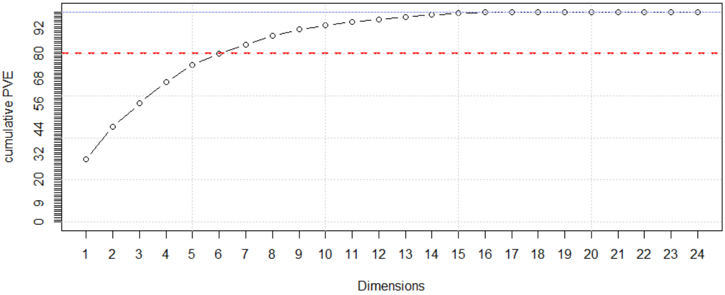
Cleaning of the dataset:

- transformation of the type of variable, from categorical to numerical.
- make the following variables categorical:
    - *condition*
    - *type shipping*
    - *fba*
    - *sold by amazon*

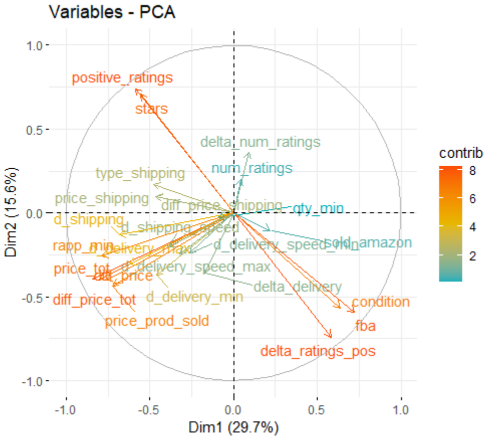
PCA


Scree plot
Percentage of explained variances
0
10
20
30
29.7%
15.6%
11%
10%
8.2%
5.5%
4.2%
4.1%
3.2%
2%
1
2
3
4
5
6
7
8
9
10
Dimensions

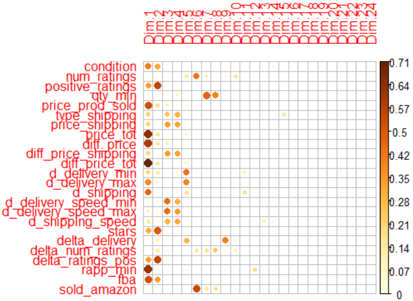
PCA


cumulative PVE
0
9
20
32
44
56
68
80
92
Dimensions
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

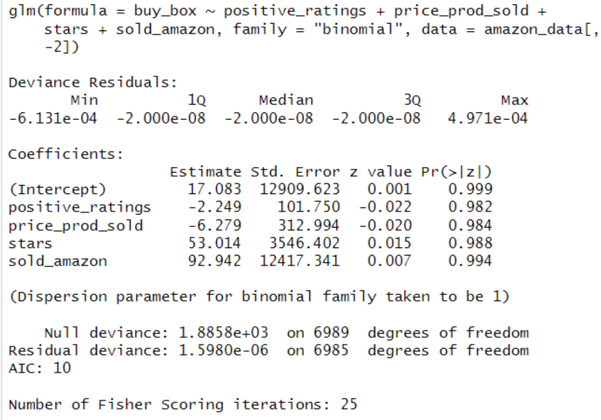
PCA


Variables - PCA
positive_ratings
stars
delta_num_ratings
num_ratings
type_shipping
price_shipping
diff_price_shipping
qty_min
d_shipping
d_shipping_speed
rapp_min
d_delivery_max
d_delivery_speed_min
sold_amazon
price_tot
diff_price
d_delivery_speed_max
delta_delivery
diff_price_tot
d_delivery_min
condition
price_prod_sold
fba
delta_ratings_pos
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
Dim1 (29.7%)
Dim2 (15.6%)
contrib
8
6
4
2

PCA

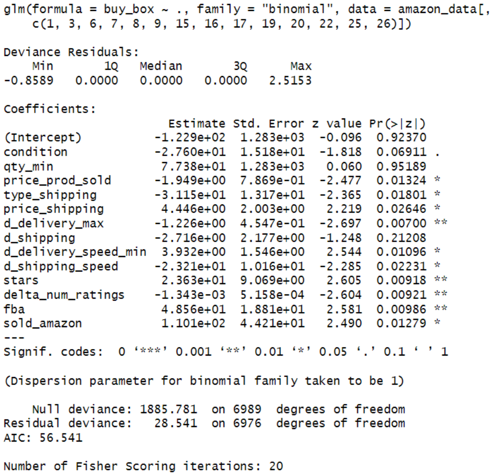

Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
Dim.6
Dim.7
Dim.8
Dim.9
Dim.10
Dim.11
Dim.12
Dim.13
Dim.14
Dim.15
Dim.16
Dim.17
Dim.18
Dim.19
Dim.20
Dim.21
Dim.22
Dim.23
Dim.24
condition
num_ratings
positive_ratings
qty_min
price_prod_sold
type_shipping
price_shipping
price_tot
diff_price
diff_price_shipping
diff_price_tot
d_delivery_min
d_delivery_max
d_shipping
d_delivery_speed_min
d_delivery_speed_max
d_shipping_speed
stars
delta_delivery
delta_num_ratings
delta_ratings_pos
rapp_min
fba
sold_amazon
0.71
0.64
0.57
0.5
0.42
0.35
0.28
0.21
0.14
0.07
0

Backward Selection



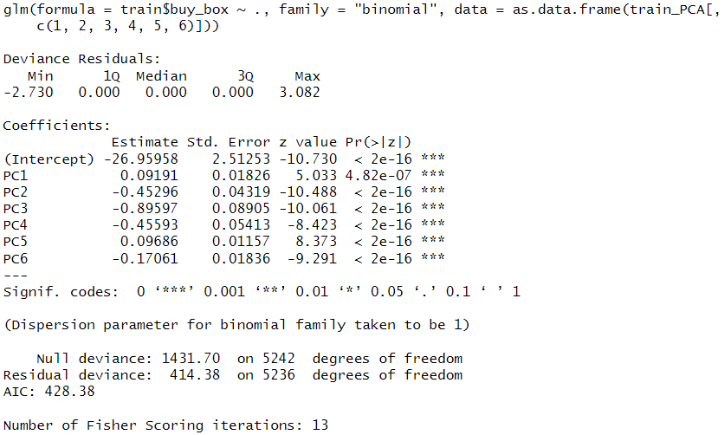
```
glm(formula = buy_box ~ positive_ratings + price_prod_sold +
    stars + sold_amazon, family = "binomial", data = amazon_data[,
    -2])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-6.131e-04  -2.000e-08  -2.000e-08  -2.000e-08   4.971e-04

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         17.083  12909.623   0.001    0.999
positive_ratings    -2.249    101.750  -0.022    0.982
price_prod_sold     -6.279    312.994  -0.020    0.984
stars               53.014   3546.402   0.015    0.988
sold_amazon         92.942  12417.341   0.007    0.994

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.8858e+03  on 6989  degrees of freedom
Residual deviance: 1.5980e-06  on 6985  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 25
```

Forward Selection



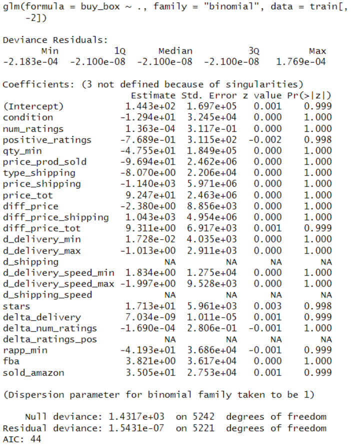
```
glm(formula = buy_box ~ ., family = "binomial", data = amazon_data[,
    c(1, 3, 6, 7, 8, 9, 15, 16, 17, 19, 20, 22, 25, 26)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8589   0.0000   0.0000   0.0000   2.5153

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.229e+02  1.283e+03  -0.096  0.92370
condition            -2.760e+01  1.518e+01  -1.818  0.06911 .
qty_min               7.738e+01  1.283e+03   0.060  0.95189
price_prod_sold      -1.949e+00  7.869e-01  -2.477  0.01324 *
type_shipping        -3.115e+01  1.317e+01  -2.365  0.01801 *
price_shipping        4.446e+00  2.003e+00   2.219  0.02646 *
d_delivery_max       -1.226e+00  4.547e-01  -2.697  0.00700 **
d_shipping           -2.716e+00  2.177e+00  -1.248  0.21208
d_delivery_speed_min  3.932e+00  1.546e+00   2.544  0.01096 *
d_shipping_speed     -2.321e+01  1.016e+01  -2.285  0.02231 *
stars                 2.363e+01  9.069e+00   2.605  0.00918 **
delta_num_ratings    -1.343e-03  5.158e-04  -2.604  0.00921 **
fba                   4.856e+01  1.881e+01   2.581  0.00986 **
sold_amazon           1.101e+02  4.421e+01   2.490  0.01279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1885.781  on 6989  degrees of freedom
Residual deviance:   28.541  on 6976  degrees of freedom
AIC: 56.541

Number of Fisher Scoring iterations: 20
```

PCA: Logistic Regression



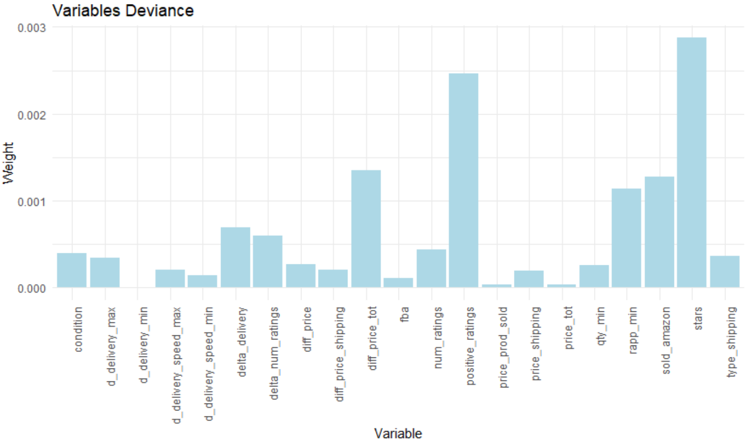
```
glm(formula = train$buy_box ~ ., family = "binomial", data = as.data.frame(train_PCA[,
    c(1, 2, 3, 4, 5, 6)]))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.730   0.000   0.000   0.000   3.082

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -26.95958    2.51253 -10.730  < 2e-16 ***
PC1           0.09191    0.01826   5.033 4.82e-07 ***
PC2          -0.45296    0.04319 -10.488  < 2e-16 ***
PC3          -0.89597    0.08905 -10.061  < 2e-16 ***
PC4          -0.45593    0.05413  -8.423  < 2e-16 ***
PC5           0.09686    0.01157   8.373  < 2e-16 ***
PC6          -0.17061    0.01836  -9.291  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1431.70  on 5242  degrees of freedom
Residual deviance:  414.38  on 5236  degrees of freedom
AIC: 428.38

Number of Fisher Scoring iterations: 13
```

No Selection: Logistic Regression

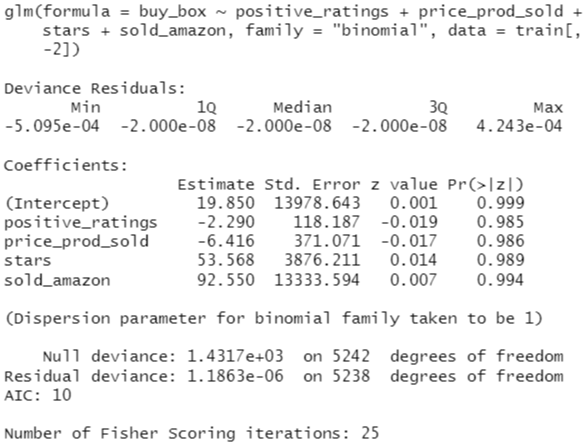


```
glm(formula = buy_box ~ ., family = "binomial", data = train[,
    -2])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.183e-04  -2.100e-08  -2.100e-08  -2.100e-08   1.769e-04

Coefficients: (3 not defined because of singularities)
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.443e+02  1.697e+05   0.001    0.999
condition            -1.294e+01  3.245e+04   0.000    1.000
num_ratings           1.363e-04  3.117e-01   0.000    1.000
positive_ratings     -7.689e-01  3.115e+02  -0.002    0.998
qty_min              -4.755e+01  1.849e+05   0.000    1.000
price_prod_sold      -9.694e+01  2.462e+06   0.000    1.000
type_shipping        -8.070e+00  2.206e+04   0.000    1.000
price_shipping       -1.140e+03  5.971e+06   0.000    1.000
price_tot             9.247e+01  2.463e+06   0.000    1.000
diff_price           -2.380e+00  8.856e+03   0.000    1.000
diff_price_shipping   1.043e+03  4.954e+06   0.000    1.000
diff_price_tot        9.311e+00  6.917e+03   0.001    0.999
d_delivery_min        1.728e-02  4.035e+03   0.000    1.000
d_delivery_max       -1.013e+00  2.911e+03   0.000    1.000
d_shipping                   NA         NA      NA       NA
d_delivery_speed_min  1.834e+00  1.275e+04   0.000    1.000
d_delivery_speed_max -1.997e+00  9.528e+03   0.000    1.000
d_shipping_speed             NA         NA      NA       NA
stars                 1.713e+01  5.961e+03   0.003    0.998
delta_delivery        7.034e-09  1.011e-05   0.001    0.999
delta_num_ratings    -1.690e-04  2.806e-01  -0.001    1.000
delta_ratings_pos            NA         NA      NA       NA
rapp_min             -4.193e+01  3.686e+04  -0.001    0.999
fba                   3.821e+00  3.617e+04   0.000    1.000
sold_amazon           3.505e+01  2.753e+04   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.4317e+03  on 5242  degrees of freedom
Residual deviance: 1.5431e-07  on 5221  degrees of freedom
AIC: 44
```

No Selection: Variance Decomposition

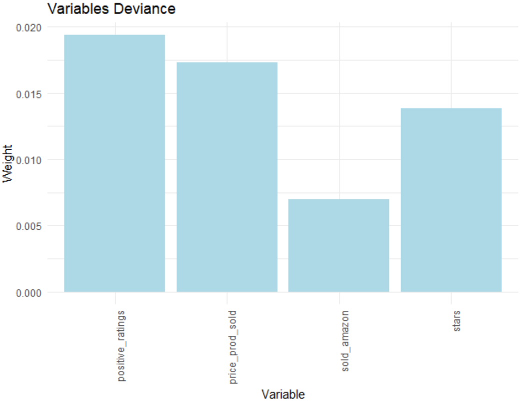


Backward Selection: Logistic Regression

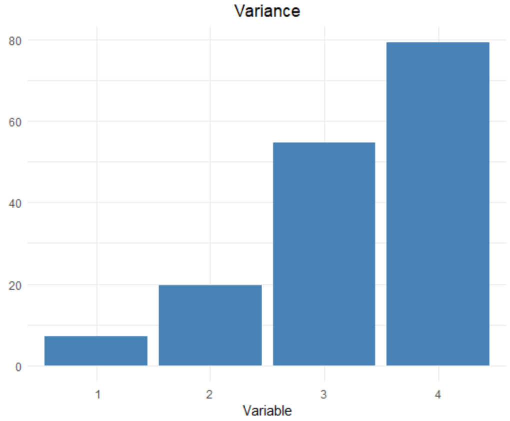


```
glm(formula = buy_box ~ positive_ratings + price_prod_sold +
    stars + sold_amazon, family = "binomial", data = train[,
    -2])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-5.095e-04  -2.000e-08  -2.000e-08  -2.000e-08   4.243e-04

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         19.850  13978.643   0.001    0.999
positive_ratings    -2.290    118.187  -0.019    0.985
price_prod_sold     -6.416    371.071  -0.017    0.986
stars               53.568   3876.211   0.014    0.989
sold_amazon         92.550  13333.594   0.007    0.994

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.4317e+03  on 5242  degrees of freedom
Residual deviance: 1.1863e-06  on 5238  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 25
```

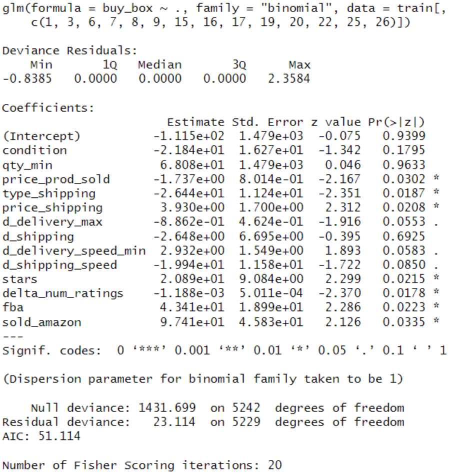
Backward Selection: Variance Decomposition(1)



Backward Selection: Variance Decomposition(2)

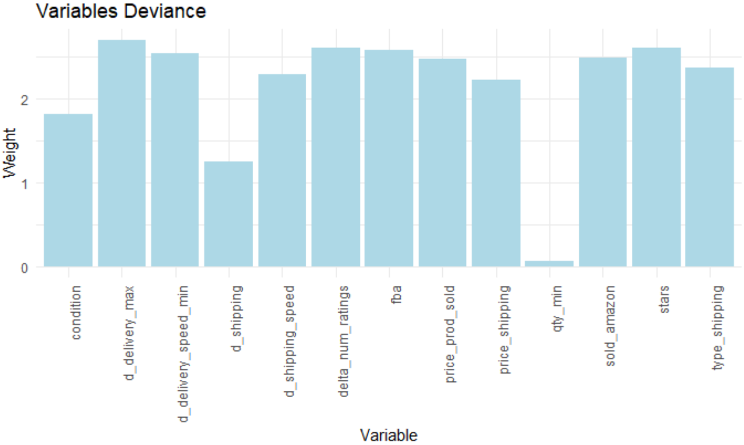


Forward Selection: Logistic Regression

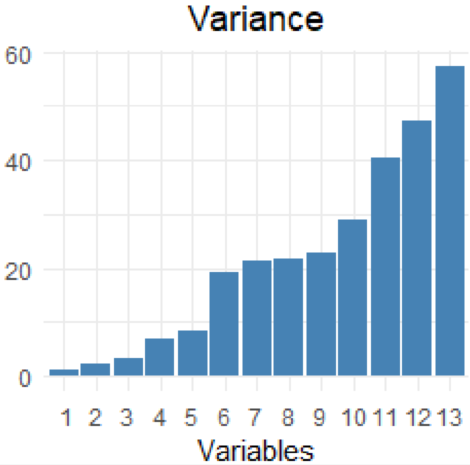


```
glm(formula = buy_box ~ ., family = "binomial", data = train[,
    c(1, 3, 6, 7, 8, 9, 15, 16, 17, 19, 20, 22, 25, 26)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8385   0.0000   0.0000   0.0000   2.3584

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.115e+02  1.479e+03  -0.075   0.9399
condition           -2.184e+01  1.627e+01  -1.342   0.1795
qty_min              6.808e+01  1.479e+03   0.046   0.9633
price_prod_sold     -1.737e+00  8.014e-01  -2.167   0.0302 *
type_shipping       -2.644e+01  1.124e+01  -2.351   0.0187 *
price_shipping       3.930e+00  1.700e+00   2.312   0.0208 *
d_delivery_max      -8.862e-01  4.624e-01  -1.916   0.0553 .
d_shipping          -2.648e+00  6.695e+00  -0.395   0.6925
d_delivery_speed_min  2.932e+00  1.549e+00   1.893   0.0583 .
d_shipping_speed    -1.994e+01  1.158e+01  -1.722   0.0850 .
stars                2.089e+01  9.084e+00   2.299   0.0215 *
delta_num_ratings   -1.188e-03  5.011e-04  -2.370   0.0178 *
fba                  4.341e+01  1.899e+01   2.286   0.0223 *
sold_amazon          9.741e+01  4.583e+01   2.126   0.0335 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1431.699  on 5242  degrees of freedom
Residual deviance:   23.114  on 5229  degrees of freedom
AIC: 51.114

Number of Fisher Scoring iterations: 20
```

Forward Selection: Variance Decomposition(1)



Forward Selection: Variance Decomposition(2)



Possible further developments

- Cross-Validation
- Dataset
  - Number of products
  - Span of time
  - Unbalanced dataset
- Selection procedures
  - Best Subset Selection
  - Cross-Validation Method
  - Shrinkage Methods (Ridge, Lasso)
- Classification algorithms
  - LDA, QDA
  - KNN
  - Random Forest, SVM

Thank You :)