July 26

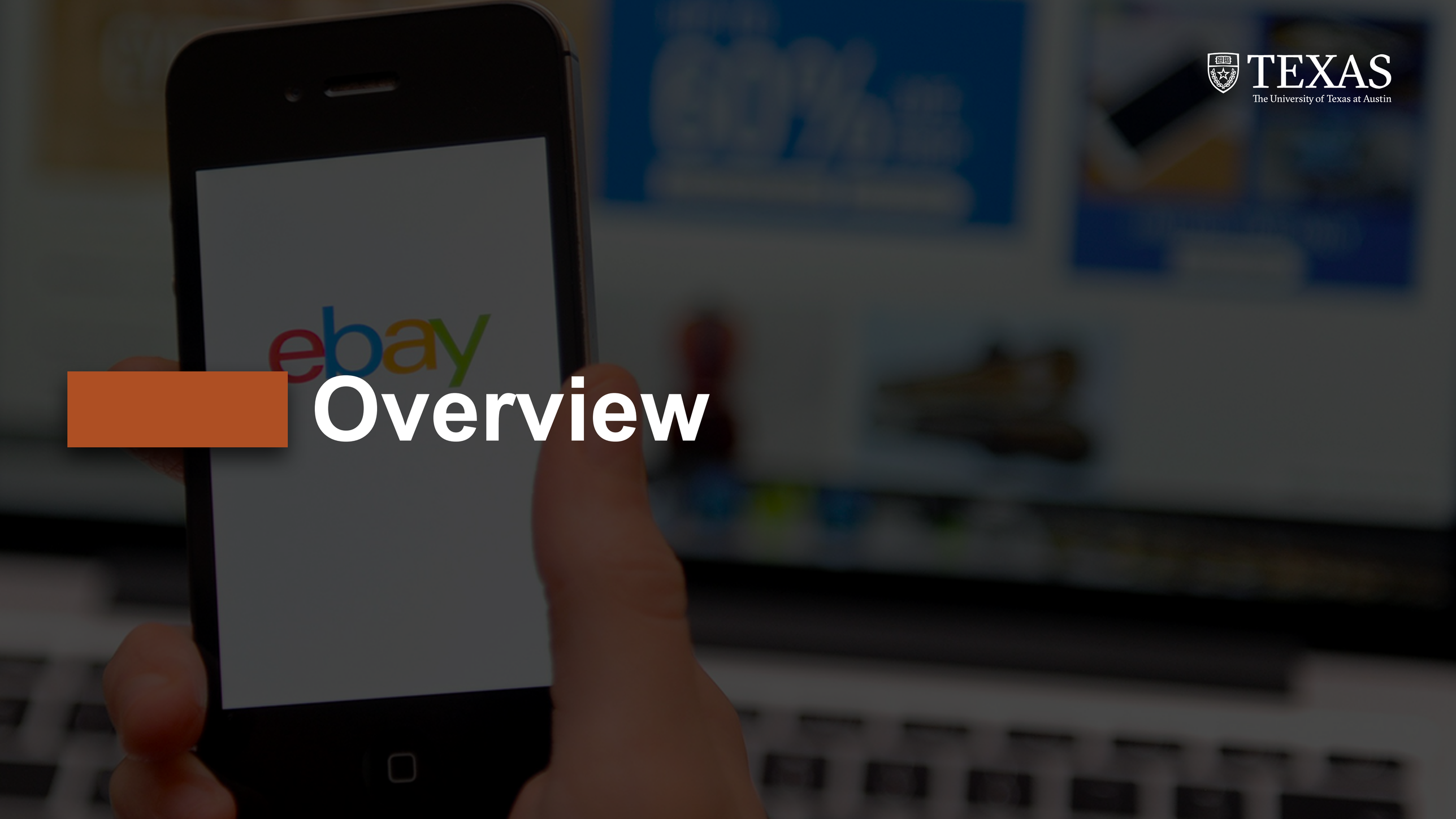
Predictive Modeling

TEXAS
The University of Texas at Austin

iPad on ebay
Sold or not sold

Group 1: Tiffany Sung, Olivia Pan, Lining Jiang, Tammy Huang, Jireh Zhou

Outline

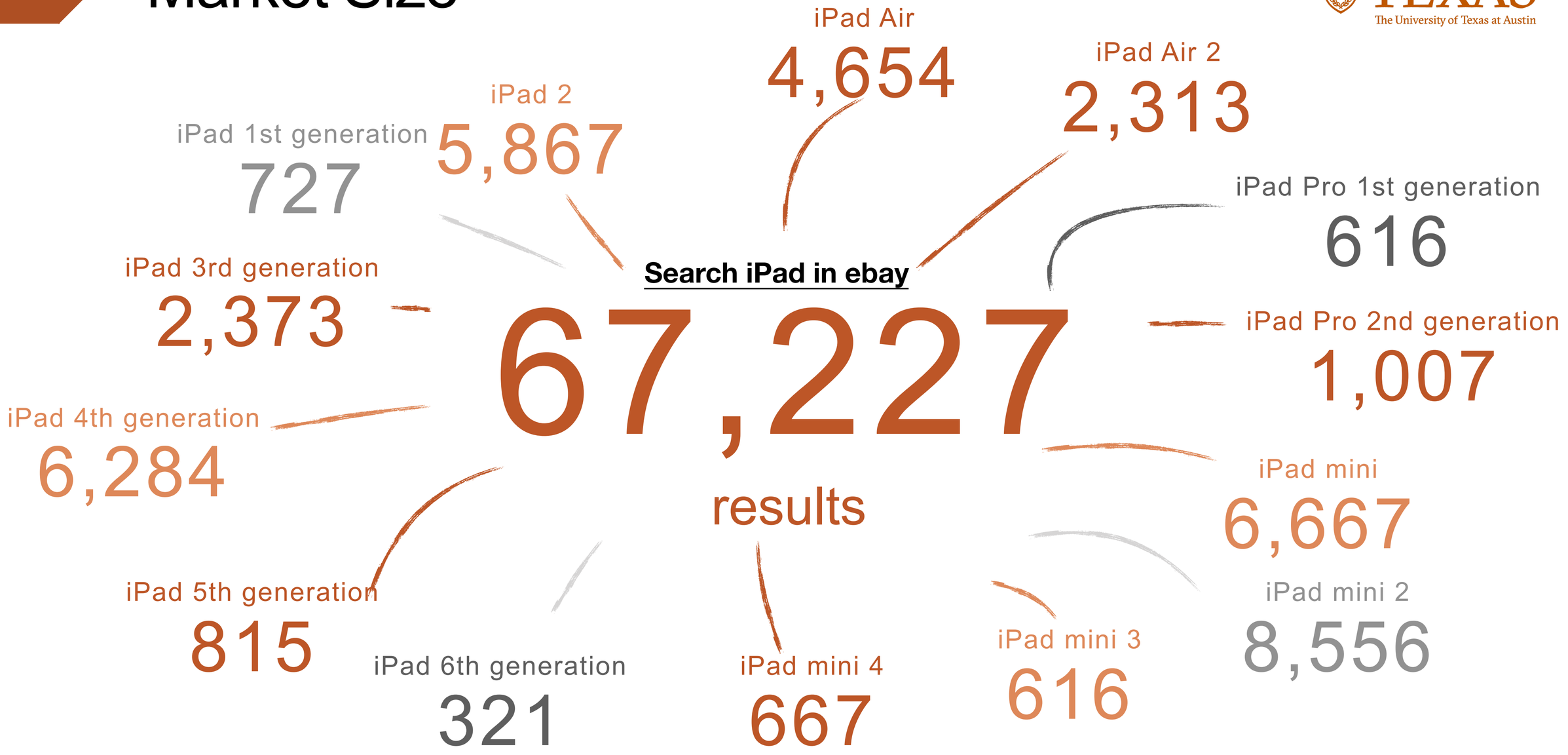


TEXAS
The University of Texas at Austin

Overview

Market Size

TEXAS
The University of Texas at Austin



Features

iPad Air

4,654

Colors

Storage 128GB / 64GB / 32GB / 16GB

Network AT&T verizon T-Mobile

Bid Biddable / Unbiddable

Condition Used/ New/ Manufacturer refurbished/ Seller refurbished

Start Price USD$ 0 ~ USD$ 1000

Overview

TEXAS
The University of Texas at Austin

67,227+

iPad

Problem Sets

TEXAS
The University of Texas at Austin

“Predict

the probability of “sold” or “not sold”
within iPads on ebay


TEXAS
The University of Texas at Austin
Dataset

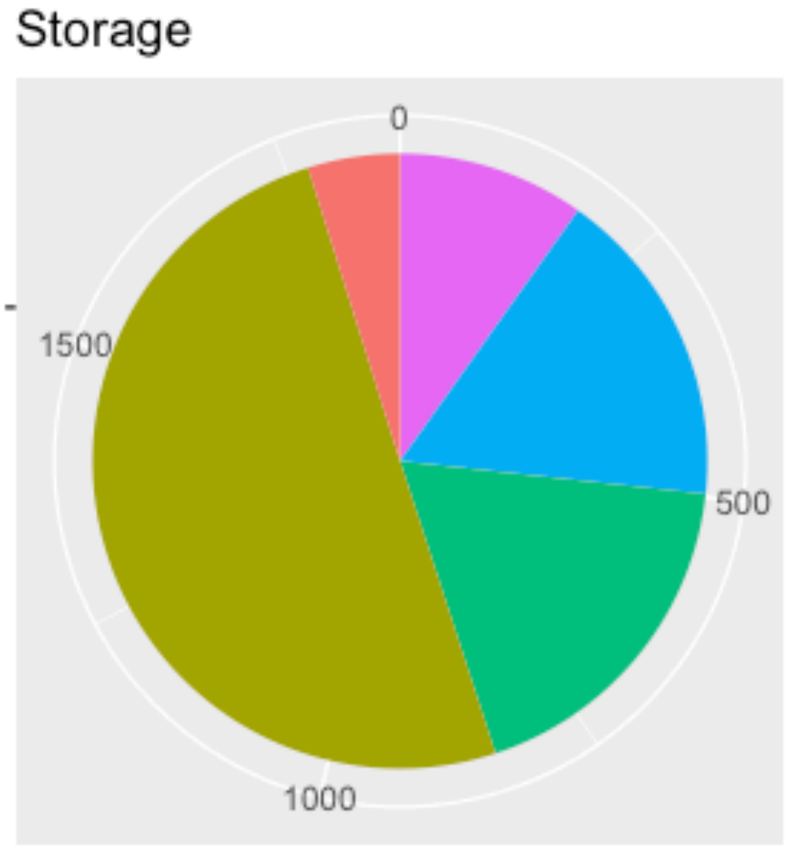
Summary

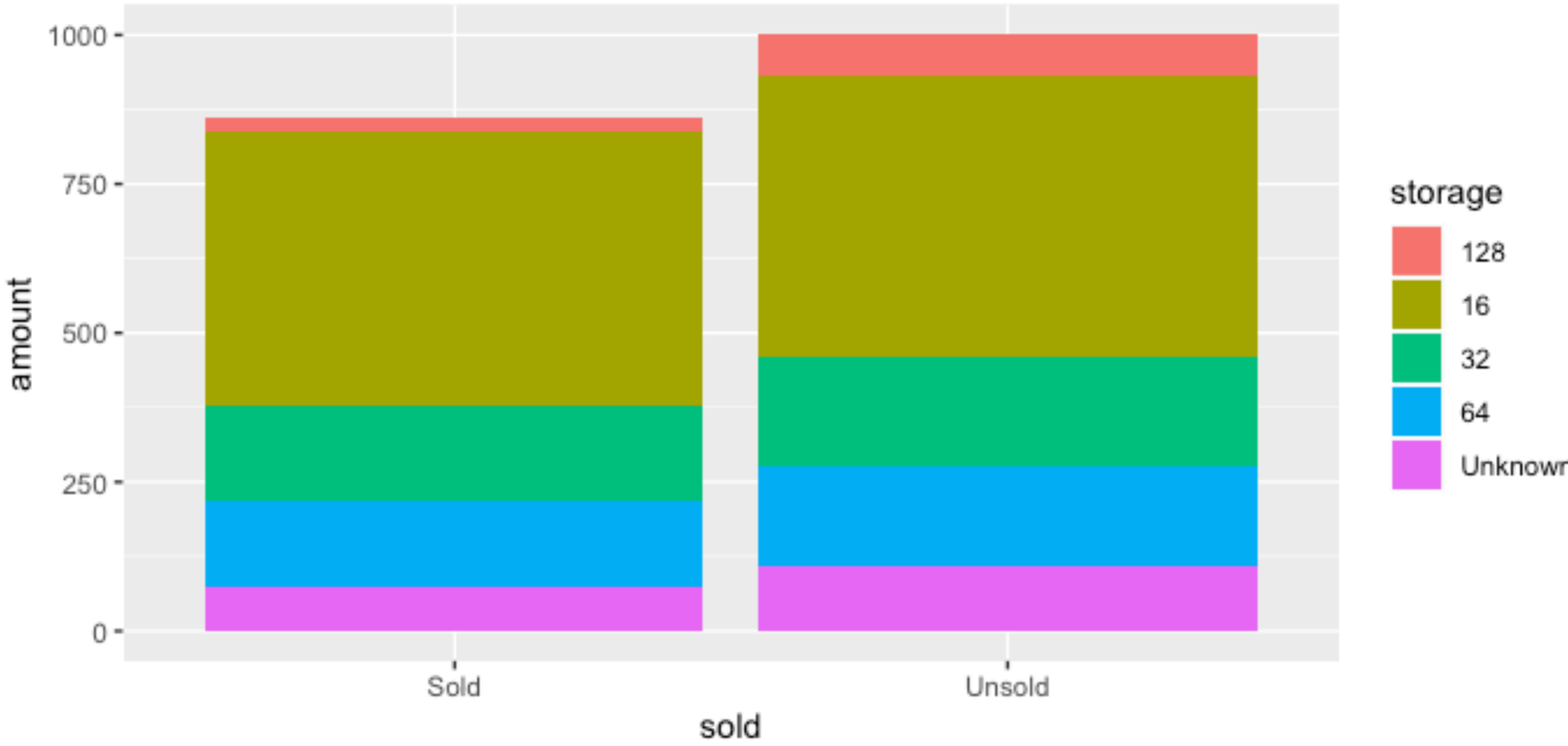
- 1,861 observations
- raw data includes 9 variables:

"description" "biddable" "startprice" "condition" "cellular" "carrier" "color" "storage" “productline"

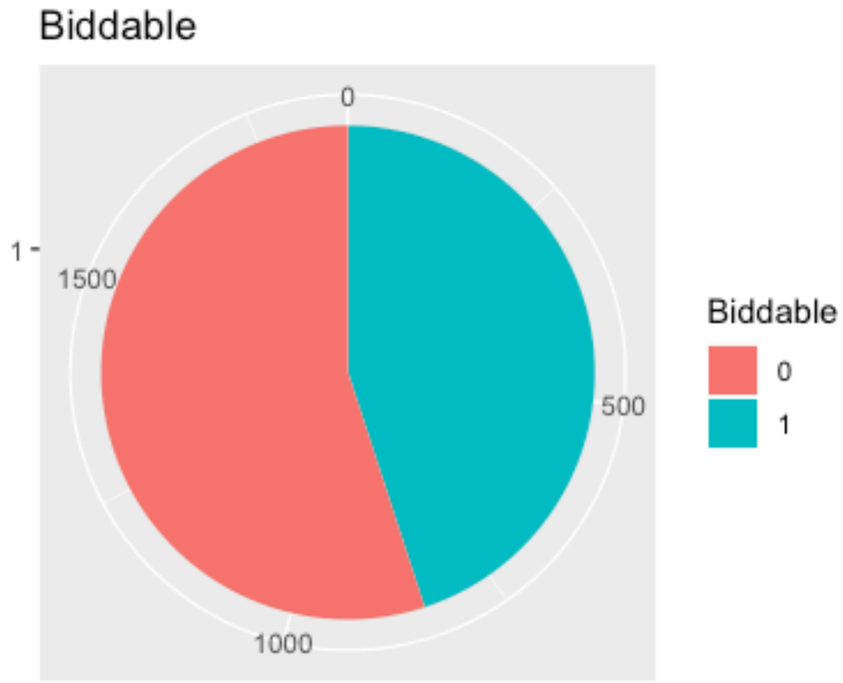
- Turned “description” into word counts “count_des”
- With only “startprice” and “count_des” are continuous data and the others are categorical data

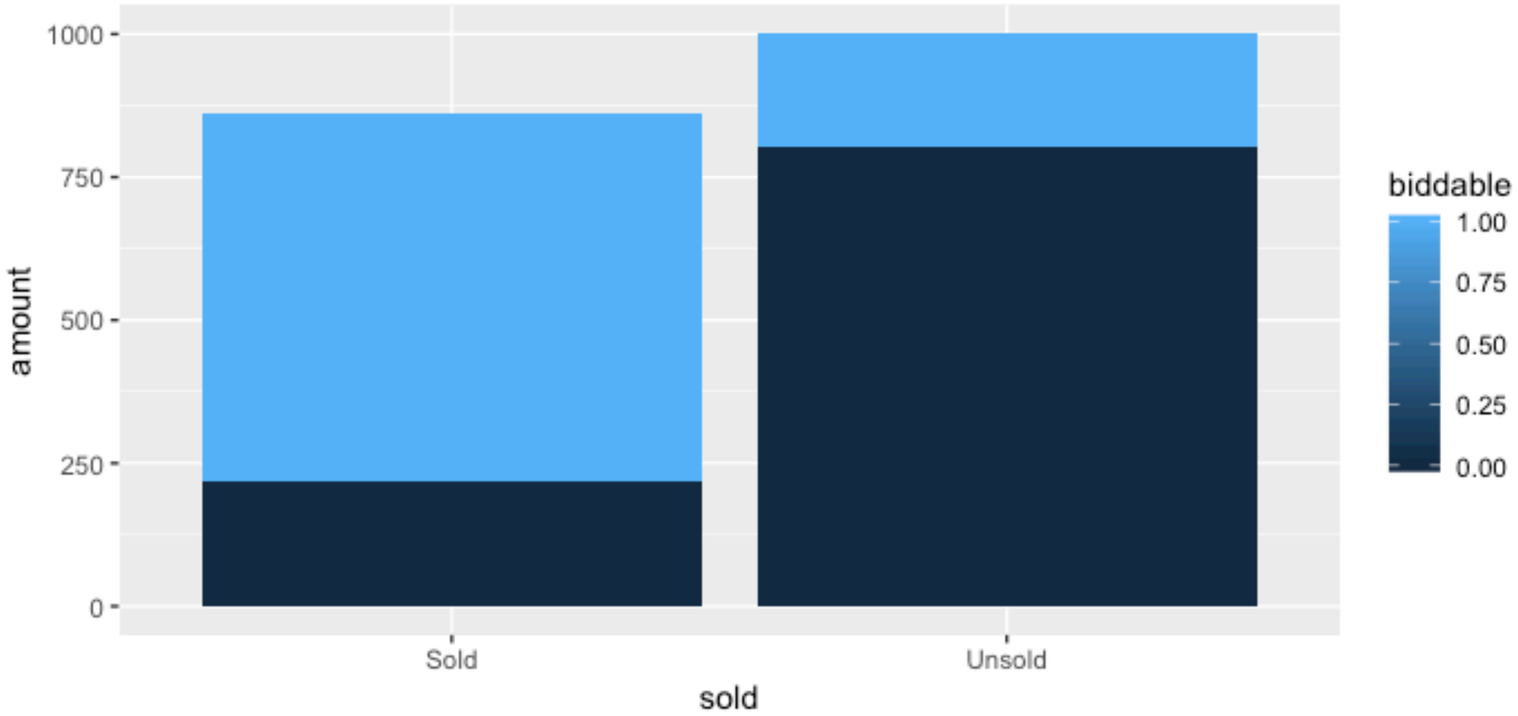
Storage


Storage
0
500
1000
1500

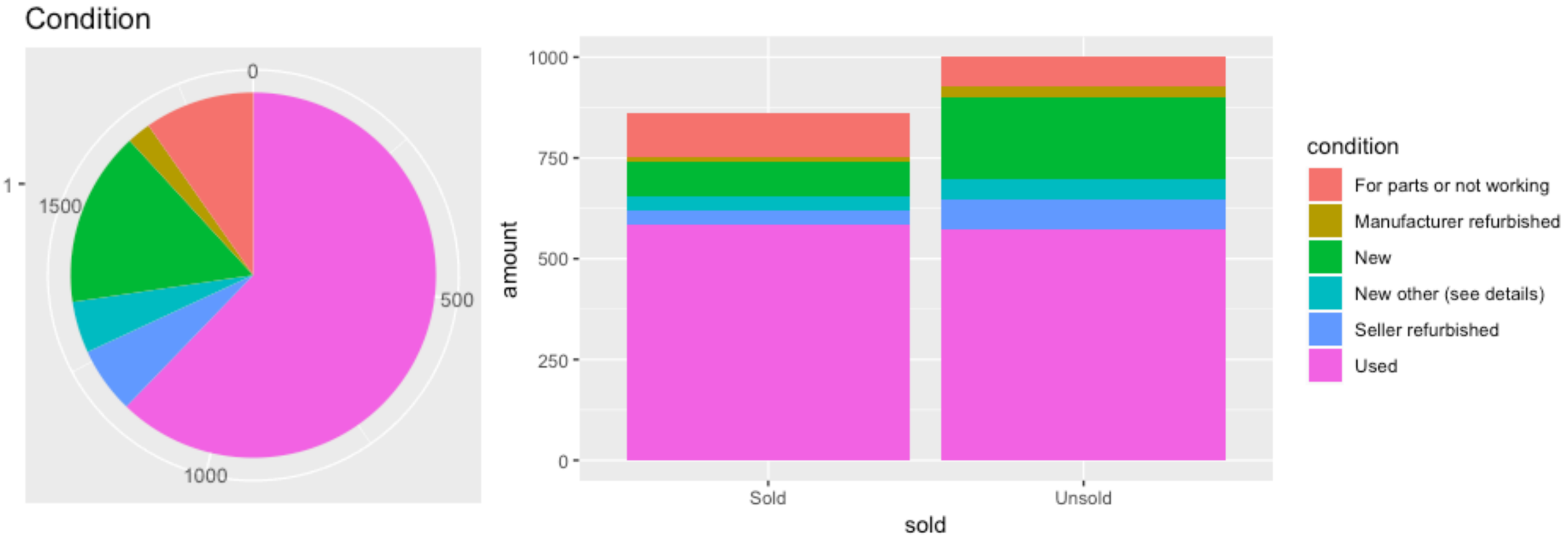

1000
750
500
250
0
amount
Sold
Unsold
sold
storage
128
16
32
64
Unknown

Biddable

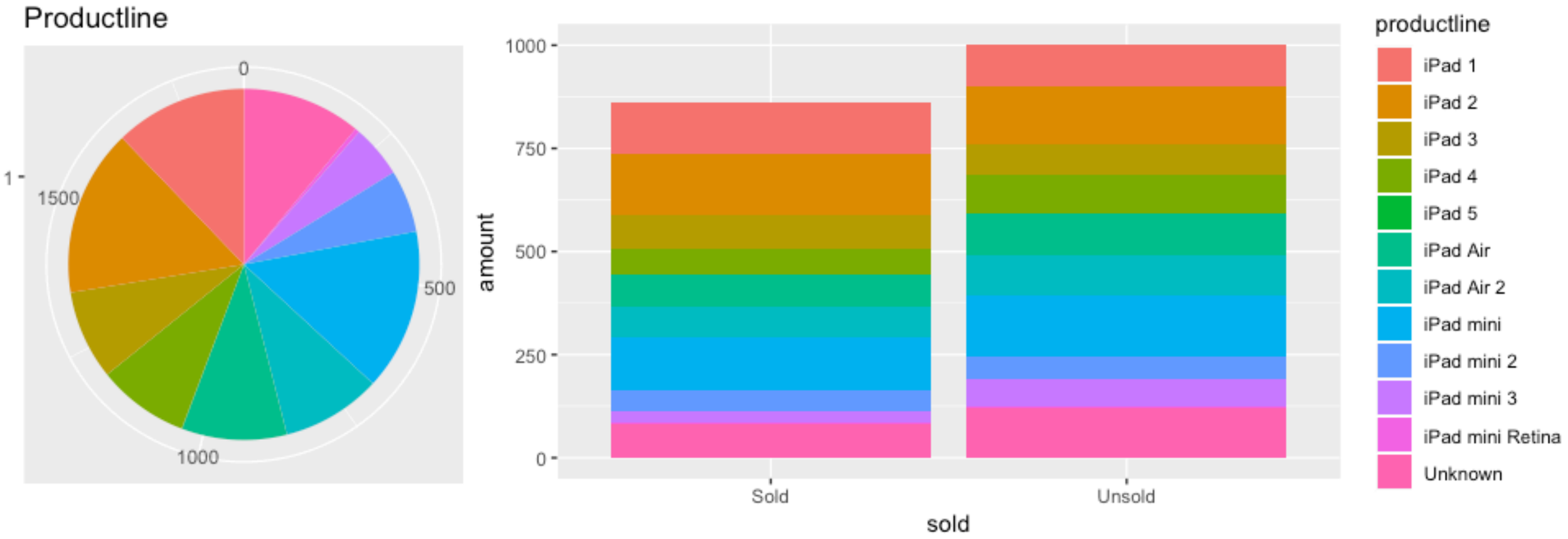

Biddable
0
1
1500
500
1000
Biddable
0
1


1000
750
500
250
0
amount
Sold
Unsold
sold
biddable
1.00
0.75
0.50
0.25
0.00

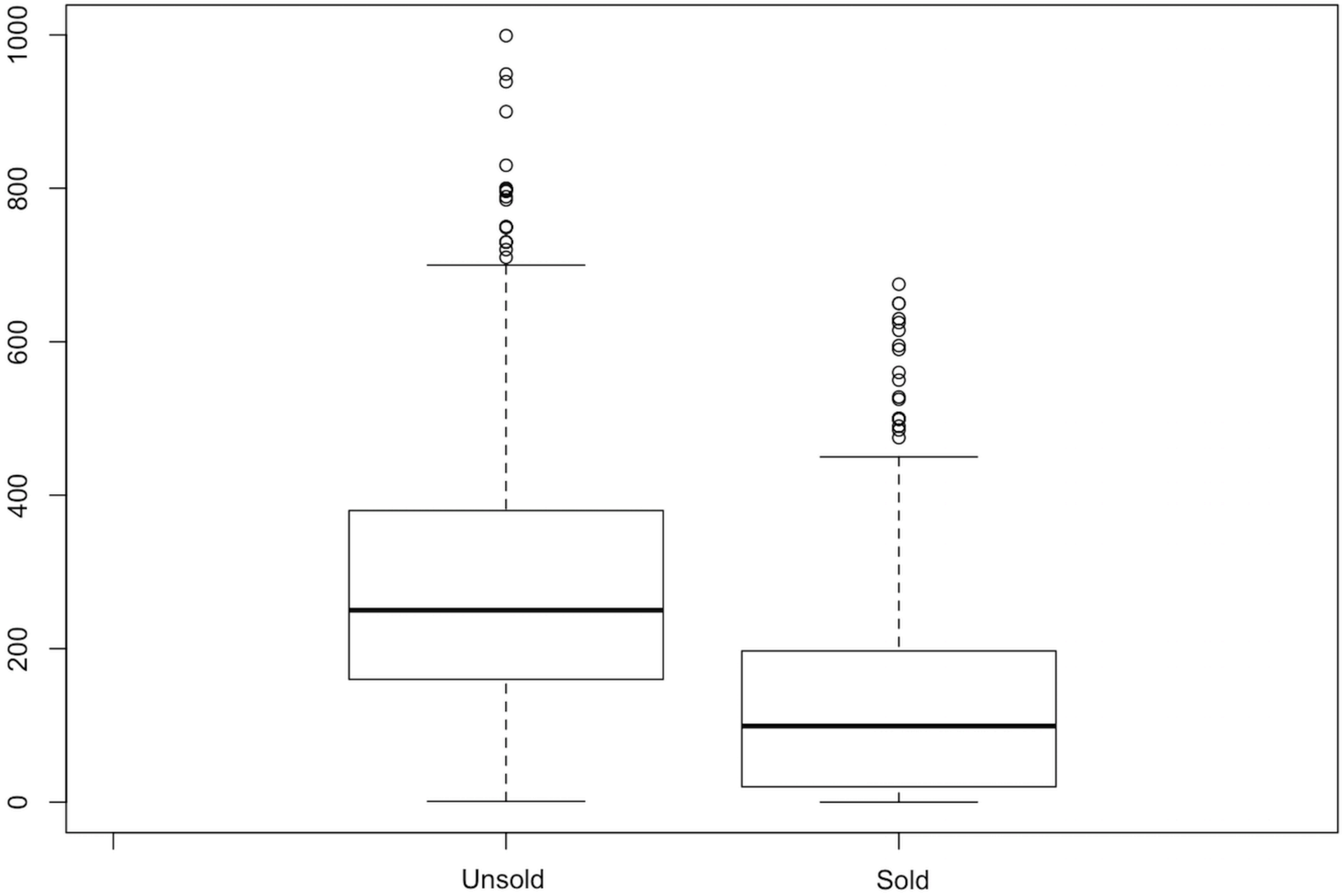
Condition


Condition
0
500
1000
1500
1
amount
1000
750
500
250
0
Sold
Unsold
sold
condition
For parts or not working
Manufacturer refurbished
New
New other (see details)
Seller refurbished
Used

Product line



StartPrice vs Sold



TEXAS
The University of Texas at Austin

Methodology

Logistic Regression/ Classification Tree

Logistic Regression

TEXAS
The University of Texas at Austin

```
glm1=glm(sold~biddable, data = trainingset, family = binomial)

glm2=glm(sold~startprice, data = trainingset, family = binomial)

glm3=glm(sold~productline, data = trainingset, family = binomial)

glm4=glm(sold~biddable + startprice, data = trainingset, family = binomial)

glm5=glm(sold~biddable + productline, data = trainingset, family = binomial)

glm6=glm(sold~startprice + productline, data = trainingset, family = binomial)

glm7=glm(sold~biddable + startprice + productline, data = trainingset, family = binomial)

glm8=glm(sold~biddable + startprice + productline + storage, data = trainingset, family = binomial)

glm9=glm(sold~biddable+ startprice + cellular + carrier + color + storage + productline + description, data =

trainingset, family = binomial)
```

4 influencial variables:
biddable, startprice, productline, storage

Logistic Regression

MODELS	1	2	3	4	5	6
Beta 0	-1.23	1.32	0.33	0.00073	-0.88	1.38
biddable	2.44			1.92	2.42	
startprice		-0.0074		-0.00497		-0.0129
product line			***		***	***
cellular						
carrier						
color						
storage						
count_des						

Logistic Regression

glm7=glm(sold~biddable + startprice + productline, data = trainingset, family = binomial)

```
Call:
glm(formula = sold ~ biddable + startprice + productline, family = binomial,
    data = trainingset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2896  -0.7441  -0.2645   0.6482   3.1842

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.2635006  0.2055179   1.282 0.199797
biddable                  1.6396826  0.1412771  11.606  < 2e-16 ***
startprice               -0.0089966  0.0007982 -11.271  < 2e-16 ***
productlineiPad 2         0.1448676  0.2406609   0.602 0.547203
productlineiPad 3         0.7323803  0.2937604   2.493 0.012663 *
productlineiPad 4         0.7675379  0.3202247   2.397 0.016536 *
productlineiPad Air       1.3637772  0.3374355   4.042 5.31e-05 ***
productlineiPad Air 2     2.7280599  0.4094992   6.662 2.70e-11 ***
productlineiPad mini      0.2249164  0.2487482   0.904 0.365893
productlineiPad mini 2    1.1483337  0.3546347   3.238 0.001203 **
productlineiPad mini 3    1.5013492  0.4376725   3.430 0.000603 ***
productlineiPad mini Retina 1.8865622  0.9540091   1.978 0.047984 *
productlineUnknown        0.0263242  0.2647703   0.099 0.920803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2074.0  on 1499  degrees of freedom
Residual deviance: 1415.9  on 1487  degrees of freedom
AIC: 1441.9

Number of Fisher Scoring iterations: 5
```

Logistic Regression

glm8=glm(sold~biddable + startprice + productline + storage, data = trainingset, family = binomial)

```
Call:
glm(formula = sold ~ biddable + startprice + productline + storage,
    family = binomial, data = trainingset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4827  -0.7263  -0.2792   0.6441   3.3433

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.3791136  0.5081351   2.714 0.006646 **
biddable                  1.5513750  0.1444379  10.741  < 2e-16 ***
startprice               -0.0100076  0.0008848 -11.311  < 2e-16 ***
productlineiPad 2         0.2475337  0.2451132   1.010 0.312555
productlineiPad 3         0.8433340  0.3008385   2.803 0.005059 **
productlineiPad 4         0.9383415  0.3315663   2.830 0.004654 **
productlineiPad Air       1.5112818  0.3467771   4.358 1.31e-05 ***
productlineiPad Air 2     2.9012330  0.4285660   6.770 1.29e-11 ***
productlineiPad mini      0.3895729  0.2628067   1.482 0.138246
productlineiPad mini 2    1.3458761  0.3671422   3.666 0.000247 ***
productlineiPad mini 3    1.7692305  0.4621306   3.828 0.000129 ***
productlineiPad mini Retina 2.1692066  0.9635338   2.251 0.024366 *
productlineUnknown        0.2903007  0.3221693   0.901 0.367545
storage16                -1.1193861  0.4427570  -2.528 0.011464 *
storage32                -1.0797054  0.4561020  -2.367 0.017921 *
storage64                -0.6449558  0.4432720  -1.455 0.145672
storageUnknown           -1.2999701  0.5230980  -2.485 0.012950 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2074.0  on 1499  degrees of freedom
Residual deviance: 1405.3  on 1483  degrees of freedom
AIC: 1439.3

Number of Fisher Scoring iterations: 5
```

Logistic Regression

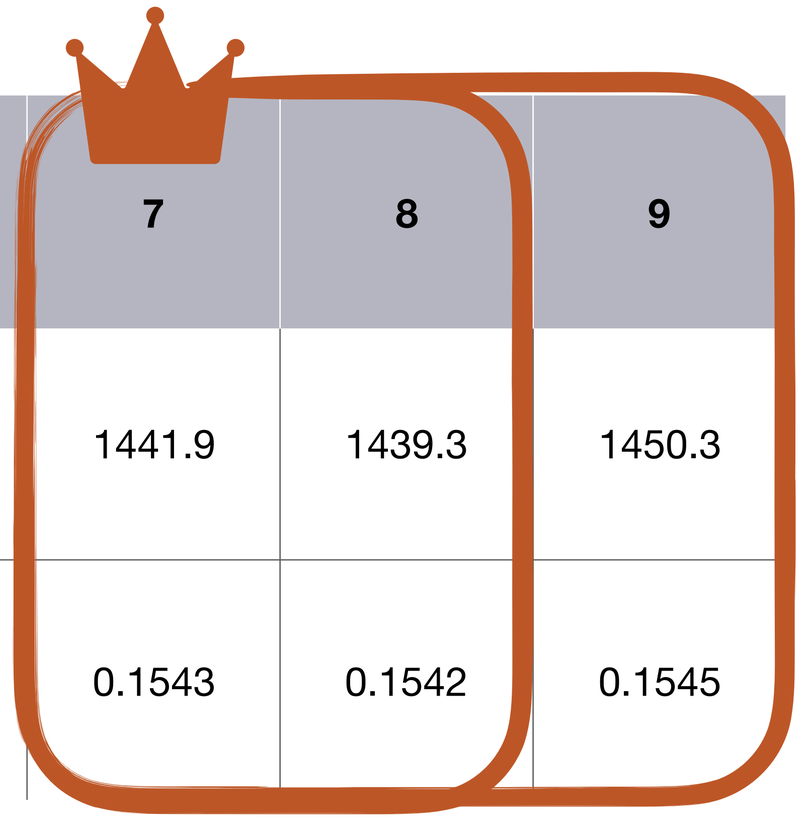
glm9=glm(sold~biddable+ startprice + cellular + carrier + color + storage + productline + description, data =trainingset, family = binomial)

```
Call:
glm(formula = sold ~ biddable + startprice + cellular + carrier +
    color + storage + productline + description, family = binomial,
    data = trainingset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5175  -0.7314  -0.2733   0.6392   3.3937

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              14.62540  624.18948   0.023 0.981306
biddable                  1.56568    0.14844  10.548  < 2e-16 ***
startprice               -0.01001    0.00090 -11.125  < 2e-16 ***
cellular1               -13.19550  624.18921  -0.021 0.983134
cellularUnknown         -13.67902  624.18916  -0.022 0.982516
carrierNone             -13.20298  624.18923  -0.021 0.983124
carrierOther             13.02137  619.73923   0.021 0.983237
carrierSprint             1.14213    0.60379   1.892 0.058546 .
carrierT-Mobile          -0.36619    0.83610  -0.438 0.661406
carrierUnknown           -0.11816    0.35466  -0.333 0.739009
carrierVerizon            0.36805    0.31790   1.158 0.246965
colorGold                -0.57651    0.50306  -1.146 0.251789
colorSpace Gray          -0.05564    0.27440  -0.203 0.839328
colorUnknown             -0.06181    0.18243  -0.339 0.734751
colorWhite               -0.06461    0.20181  -0.320 0.748838
storage16                -1.05150    0.45008  -2.336 0.019477 *
storage32                -1.01166    0.46349  -2.183 0.029059 *
storage64                -0.58933    0.45009  -1.309 0.190415
storageUnknown           -0.75254    0.59103  -1.273 0.202925
productlineiPad 2         0.21401    0.25498   0.839 0.401275
productlineiPad 3         0.85772    0.31126   2.756 0.005857 **
productlineiPad 4         0.92105    0.33574   2.743 0.006082 **
productlineiPad Air       1.51175    0.36812   4.107 4.01e-05 ***
productlineiPad Air 2     2.95805    0.44685   6.620 3.60e-11 ***
productlineiPad mini      0.34853    0.27090   1.287 0.198241
productlineiPad mini 2    1.37229    0.38037   3.608 0.000309 ***
productlineiPad mini 3    1.92651    0.48550   3.968 7.24e-05 ***
productlineiPad mini Retina 2.26517  0.98345   2.303 0.021263 *
productlineUnknown        0.43937    0.34493   1.274 0.202725
description              -0.01370    0.00937  -1.462 0.143749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2074.0  on 1499  degrees of freedom
Residual deviance: 1390.3  on 1470  degrees of freedom
AIC: 1450.3

Number of Fisher Scoring iterations: 13
```

k-fold Cross Validation and AIC

TEXAS
The University of Texas at Austin

MODEL	1	2	3	4	5	6	7	8	9
AIC	1614.5	1707.5	2065.5	1490.4	1622.6	1576.9	1441.9	1439.3	1450.3
Cross Validation Error	0.1764	0.1908	0.2479	0.1602	0.1776	0.1698	0.1543	0.1542	0.1545



k-fold Cross Validation and AIC

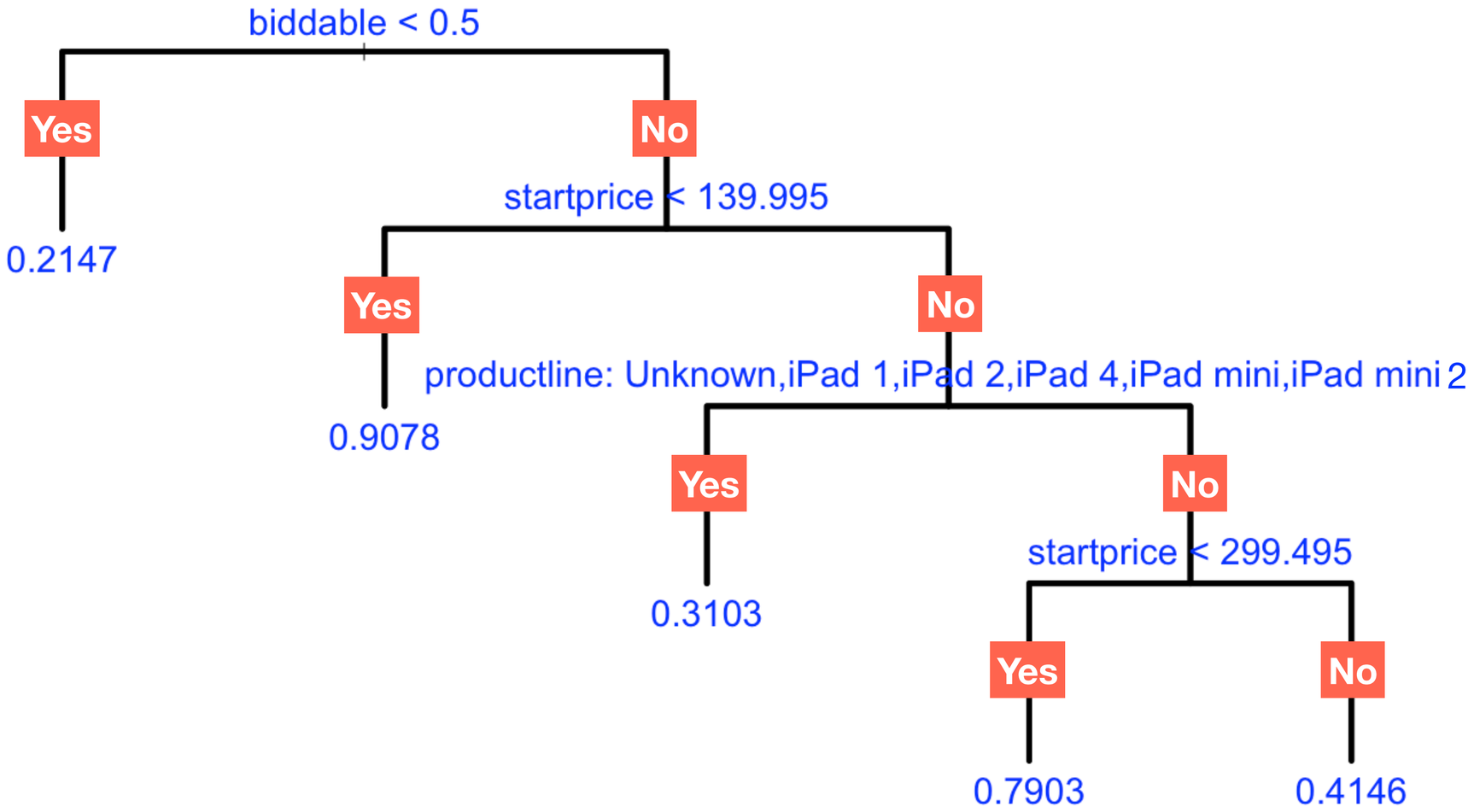
- glm7=glm(sold~biddable + startprice + productline, data = trainingset, family = binomial)

$$\begin{cases} Y = 0 \text{ , not sold} \\ Y = 1, \text{ sold} \end{cases}$$

- Predict SOLD if $Y > 0.5$
- Accuracy: 81.111% = (171+121) / (171+33+35+121)

glm.pred	0	1
0	171	33
1	35	121

Classification Tree



Classification Tree

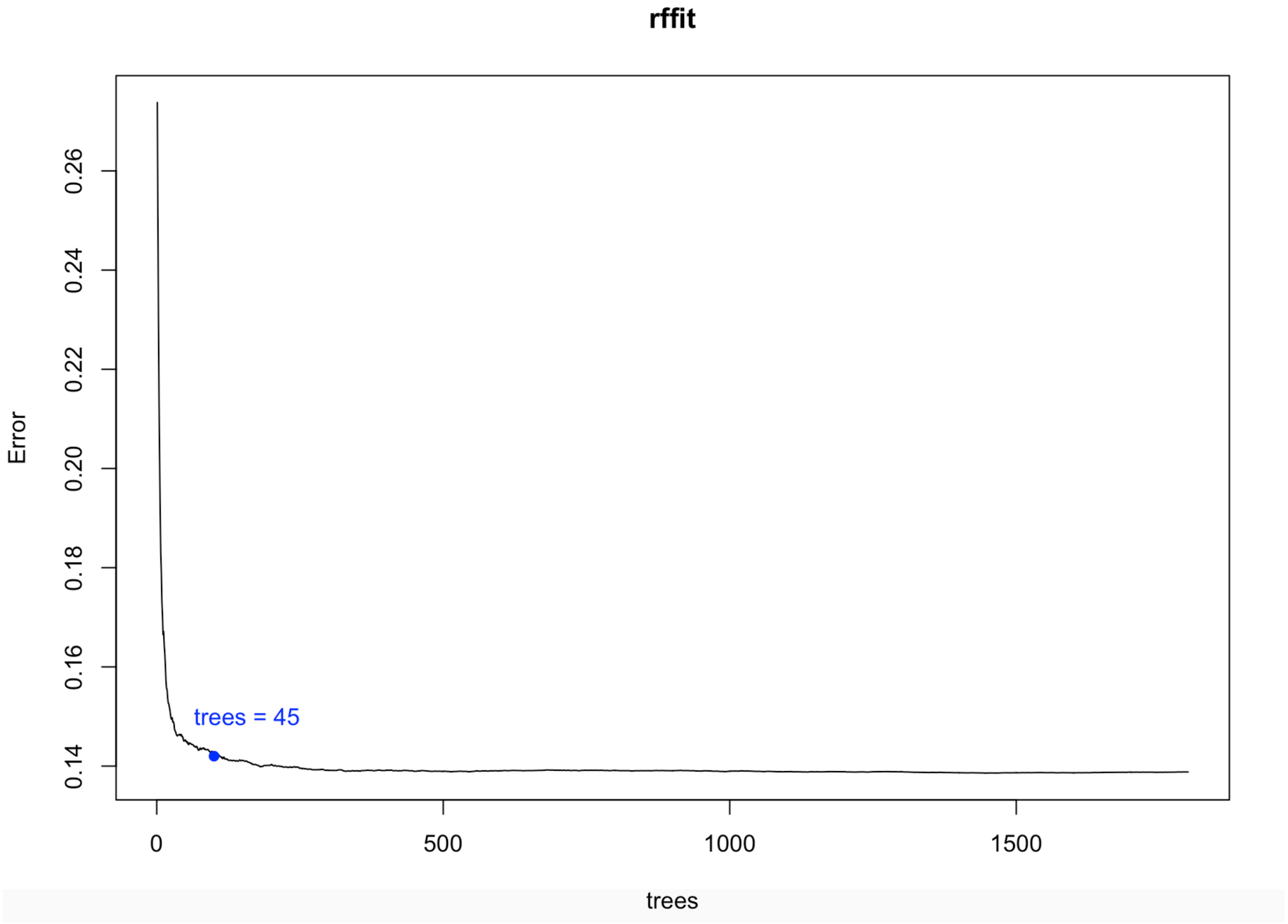
- Variables actually used in tree construction:

  [1] "biddable" "startprice" “productline”

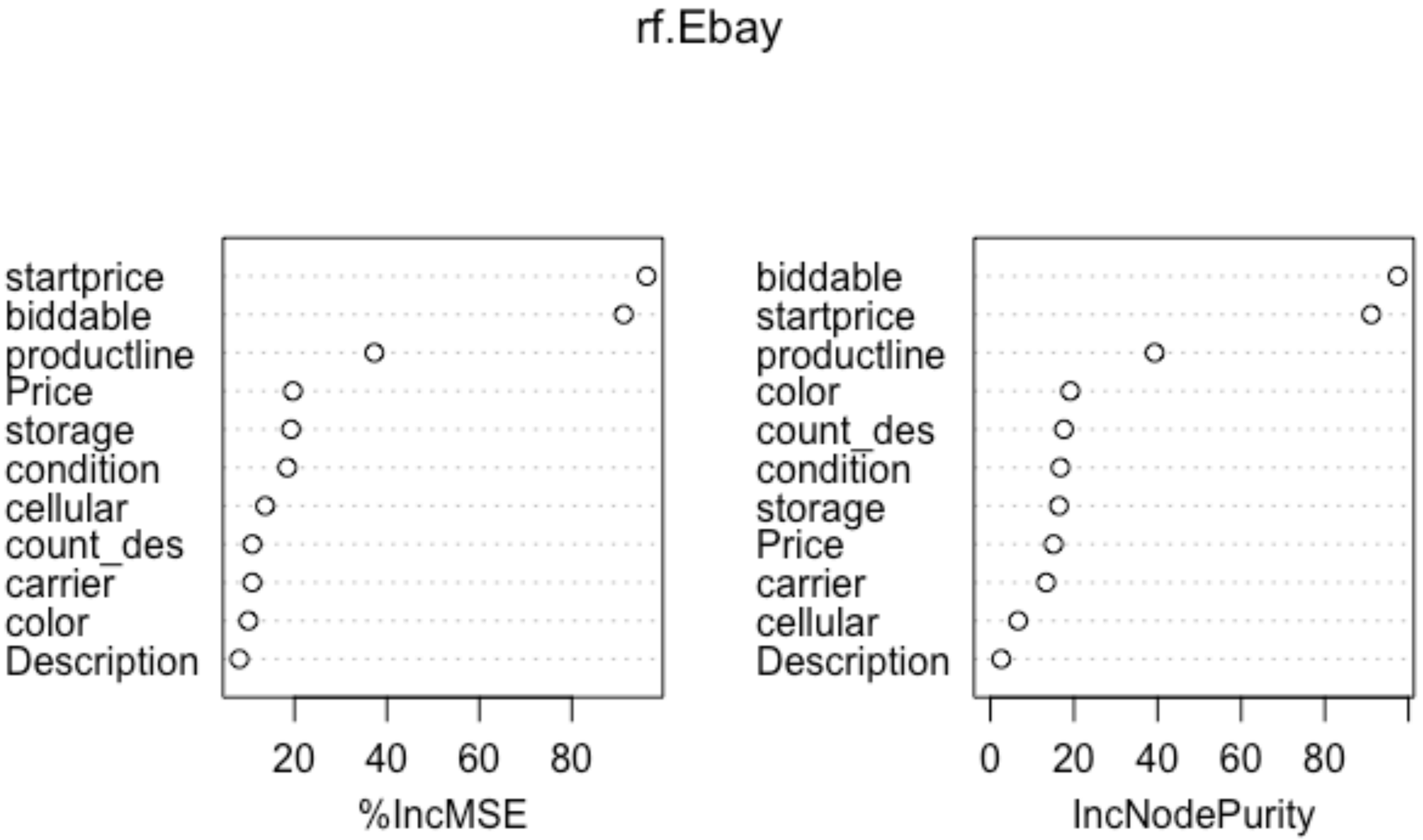
- Residual mean deviance: 0.1499

- Accuracy: 80.06% = (175+114) / (175+50+22+114)

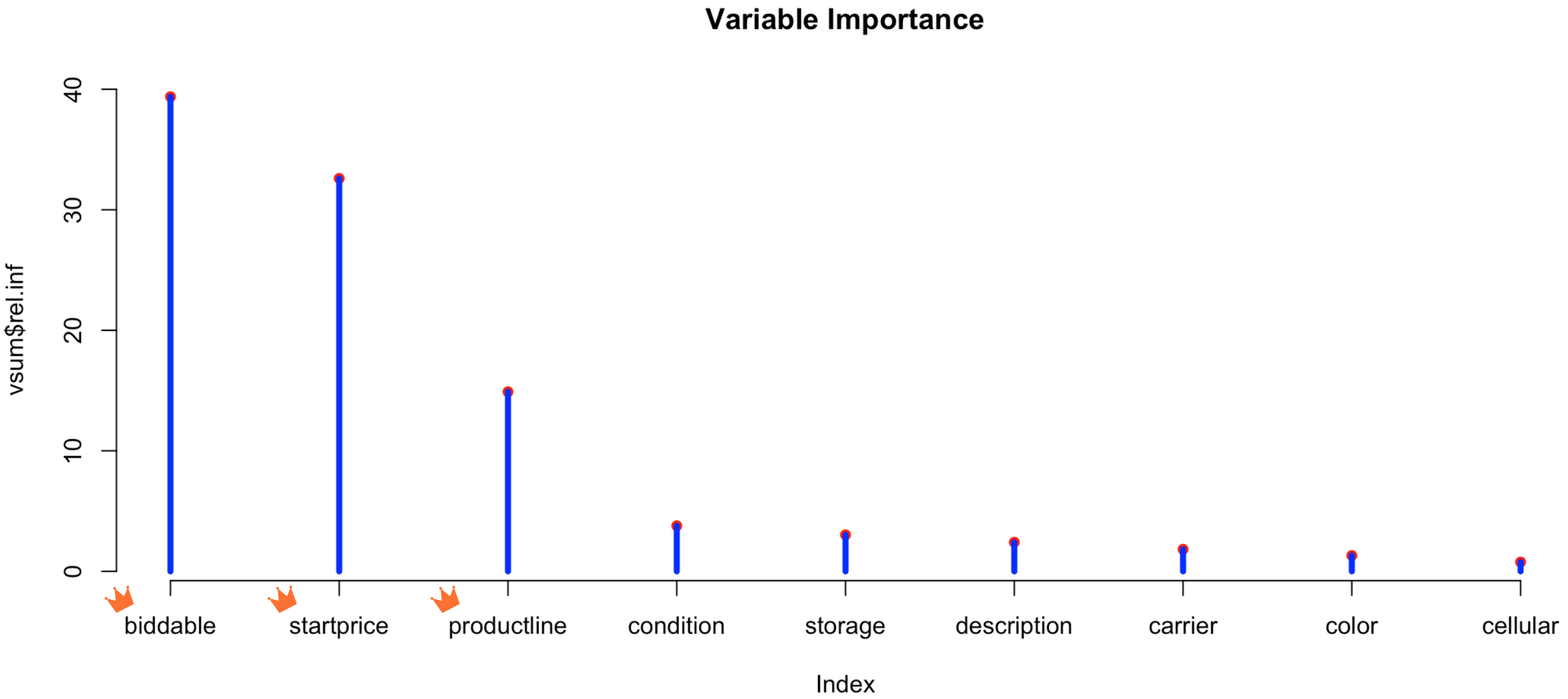
tree.pred	0	1
0	175	50
1	22	114


rffit
Error
0.14
0.16
0.18
0.20
0.22
0.24
0.26
trees = 45
0
500
1000
1500
trees

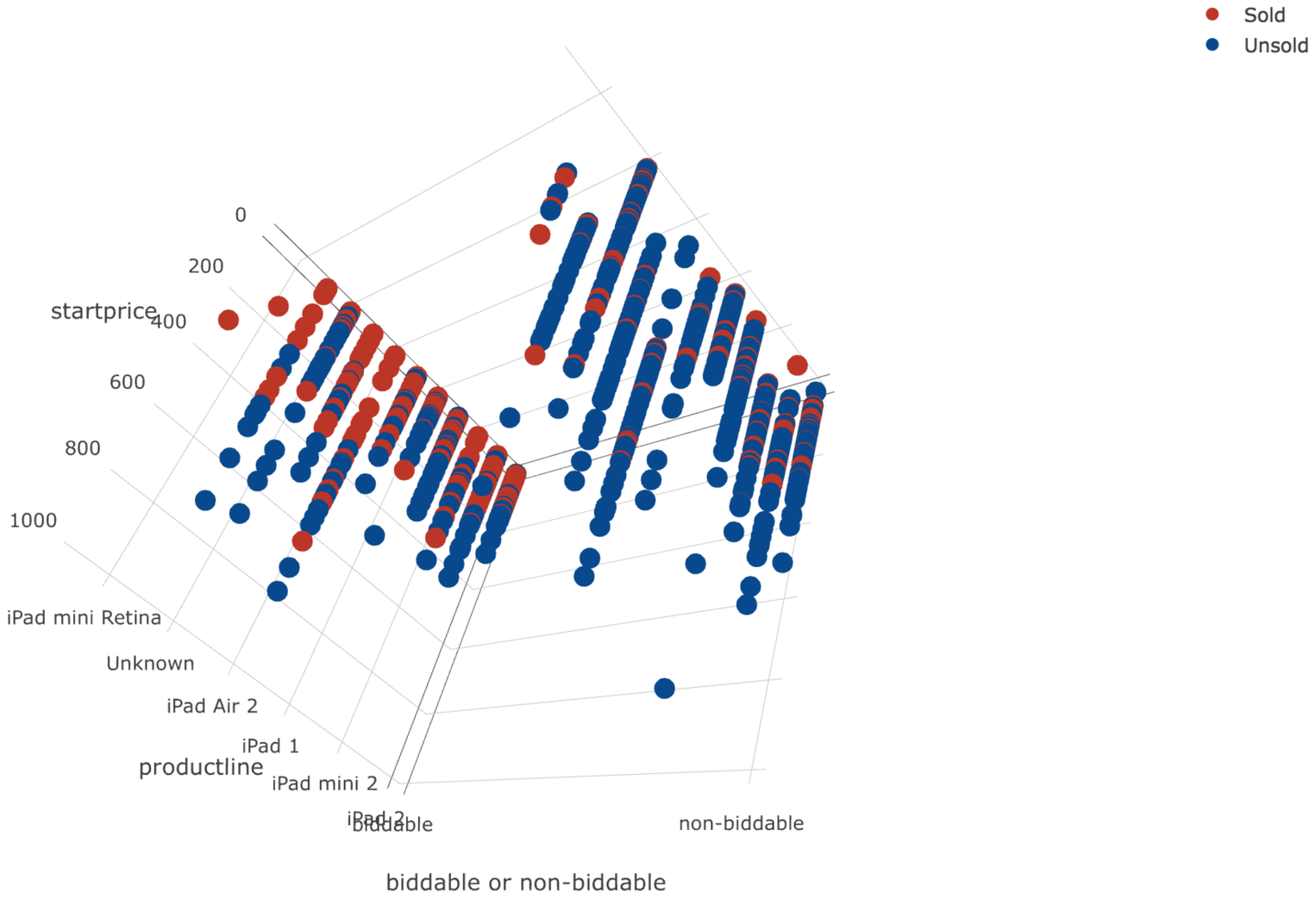
Random Forests

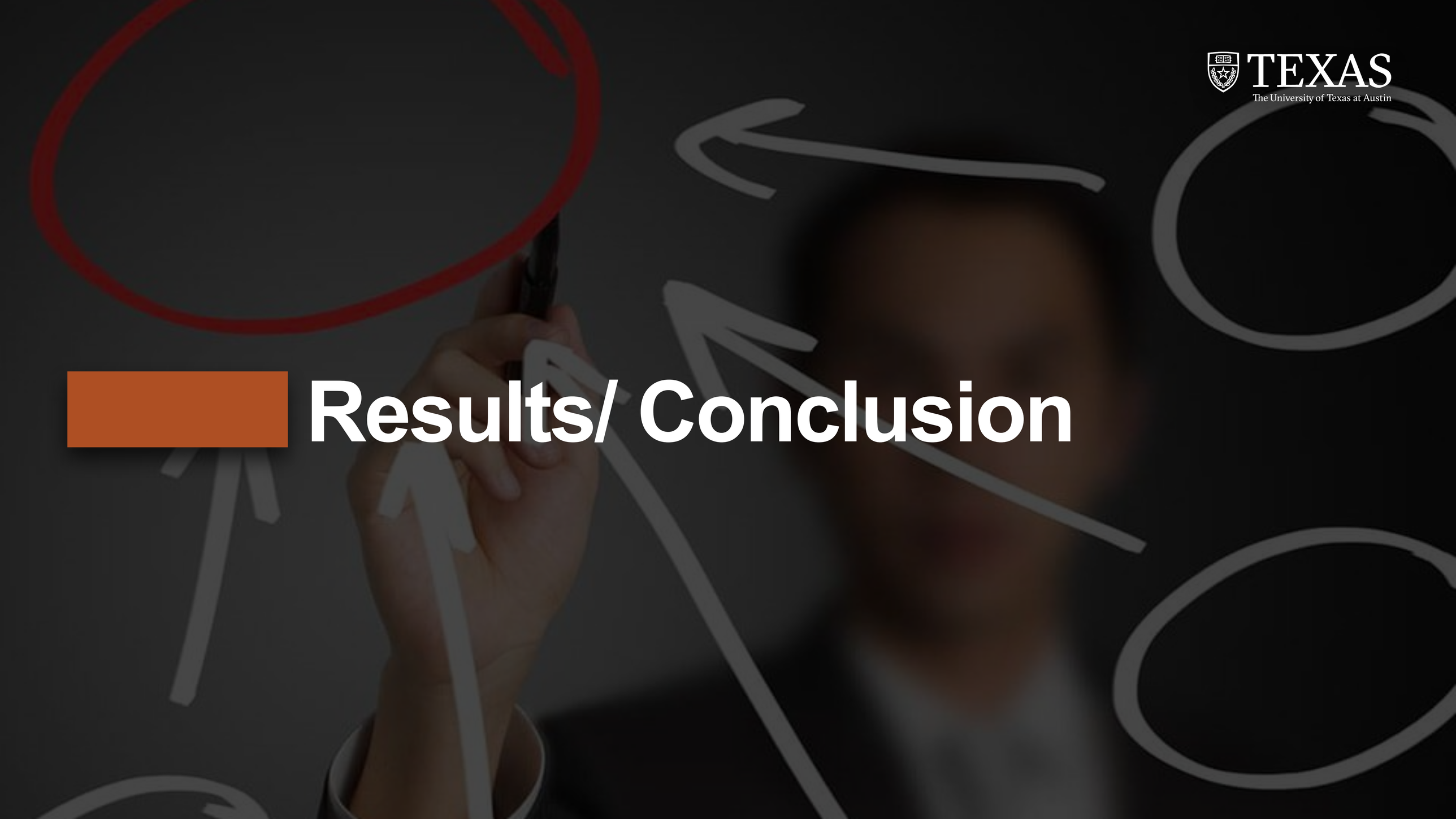


Boosting


Variable Importance
vsum$rel.inf
0
10
20
30
40
biddable
startprice
productline
condition
storage
description
carrier
color
cellular
Index

3D Presentation





Results/ Conclusion

Results

TEXAS
The University of Texas at Austin

	Logistic Model (glm7)	Classification Tree (5 nodes)	Random Forest (45 trees)
Variables Used	**Biddable, StartPrice, Product line**	**Biddable, StartPrice, Product line**	**Unknown**
Accuracy	**81.11%**	**80.03%**	**86.00%**

Conclusion

3 significant variables

Set as biddable, lower your start price and then...

Higher chance to ... SELL!!!!

TEXAS
The University of Texas at Austin

FAQ