Math 189 Final Proejct: Avito Demand Prediction

Ailing Pi (A92061788 Math 189) , Haotian Wu (A92043224 Math189), Wenqi Luo(A92038031), Xiaohang Mao(A92121381 Math189), Yingqi Gao(A92048452 Math189), Yunxiao Xiang (A92087558 Math189),

---

1. Introduction

With the development of technology, online trading has become more and more prevailing. Like traditional trading, online advertisement plays an important role also in determine the sale of products. The situation is even clear in the market for selling used products online. As the largest classified advertisements website in Russia, Avito noticed the effect of advertisements on product demand. It seems like even with optimized product listing, certain characteristics of advertisements also have influence on customer's decision. In order to maximize their sellers' surplus, Avito believes it is beneficial if they can predict demand of specific product with specific advertisement [2]. Therefore, sellers would be able to sell their products at the most proper and satisfying prices. This research aims to help Avito in predicting demand based on description and context of advertisement.

2. Data

Data source: The research will be conducted by using data provided by Avito. The data set contains information regarding details of advertisements and corresponding probability to sale (deal probability). Deal Probability is considered the indicator of demand for a specific advertisement in this research. The research faced two challenges on its way to process data. First, all data information was in Russian. It is inconvenient to discover characteristics in data in this way. Therefore, the data set was converted into English for convenience. [3] Second, aside from background data and language use (title and description) of advertisements, image is another important section in our data set. In order to understand aspects of image that can result in difference in deal probability, the research will prepare the data by converting images into numerical data. The conversion will get blurriness, size, width, height, dimension, dullness, and whiteness data of images. [4]

Data process: The original data set from Avito contains 1503424 products' information.

The research randomly selects 1000 of them for easy processing and inference to investigate the relationship between aspects of advertisement and deal probability. The first step is to find important advertisement properties that can affect deal probability. Test of independence is conducted for features including price, date (including month day, week day or weekend), user type, image existence, text (including word count, capital letter count and digit count for both advertisement description and its title), region, city, different levels of product categories and image features in search for their respective effect on deal probability. Then, bootstrap method is used to estimate sample mean distribution in checking on average of specific feature's influence on deal probability. Price, user type, image existence, word count for both advertisement description and title, parent category, image width, image height, image dimension and image whiteness are tested to be significant in determining deal probability. The second step is to figure out the multinomial function for deal probability based on features that are considered significant in the first step. Finally, the effectiveness of the model is tested through cross validation and calculation of intervals.

3. Background

Online shopping nowadays in the form of e-commerce is one of the fastest growing industries in the world. It can be achieved easily by a mobile device owing to the breakthrough of technologies. Online shopping becomes a trendy life style and increasing number of people trade through online platforms. As for the history of online shop evolution, Michael Aldrich invented the first online shopping system to enable online transaction processing between B2B&B2C. In 1990 Tim Berners- Lee created first World Wide server and browser and E-commerce sales topped $1 trillion for the first time in history. Being as one of the 10 biggest markets by global e-commerce sales in 2015, Russia achieved the online sale amount of $20.30 billion. [6]

Avito is the most popular classified advertisements website in Russia. It has around 23.9 million users in December 2013. More than 500,000 advertisements are posted by

users daily on average. [2] Many sellers meet problems brought by too much (products were underpriced) or too low demand (something wrong with the product listing).[2] Optimization of the product listing may be a helpful way to increase the deal probability.

Online advertisements play an important role in attracting customers and influencing the final deal probability. Combination of tiny factors (the context, image, price listed on the ads, category etc.) can have large effects on buyer's decision. [2] Consequently, researching on the relation between the full description of items and the deal probability can offer an optimized way for sellers to improve the product listing and provide sellers the realistic expectation they should receive[3] The following investigation uses test of independence, logistic regression, linear regression, goodness of fit test and bootstrap to study the relationship between each individual parameter and the deal probability.

4. Understanding the structure of the selected data:

This section will display the distributions of each variable given in the selected data mainly by graphs. The main goal is to roughly visualize the structure of the selected data. Hopefully, understanding the structure will help construct a blue print for further analysis.

4.1 What is the distribution of Deal Probability?

This subsection will display the distribution of the target variable, deal probability. Deal probability is the likelihood that an ad actually makes a deal of something. Since it is impossible to track every transaction with certainty, the value of this variable can be any float between 0 and 1 (including 0 and 1).

Based on the Table 4.1.1, 0 takes a large percentage of the whole data. There are two concentrated area after deleting all 0 data. The Figure 4.1.1 points out that data concentrates in the comparatively small deal probability range. Since the distribution of data is excessively biased, data processing procedure is necessary. One way is to divide the data as zero and nonzero group. The other possible way is to study the two nonzero concentrated region shown in Table 4.1.1.

Table 4.1.1 Location summary statistics of deal probability

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	0.0000	0.1390	0.1679	1.0000

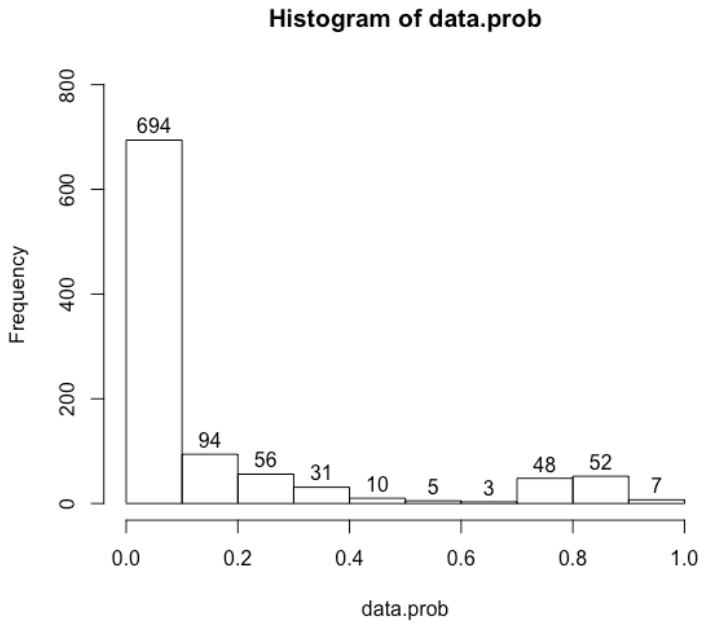

Figure 4.1.1 Histogram of deal probability

4.2 What is the distribution of Price?

The distribution of price is also unevenly. Based on Table 4.2.1 it is notable that the price range is extremely large and high prices have great influences on the mean. Consequently, mean is very large. The histogram of price does not offer much practical information. Thus, further steps like taking log is needed. Since categories and types of products both can influence the price, the research can do individual analysis on how each factor (like category or brand) influences the price separately.

Table 4.2.1 Location summary statistics of price

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	500	1500	362671	10000	36000000

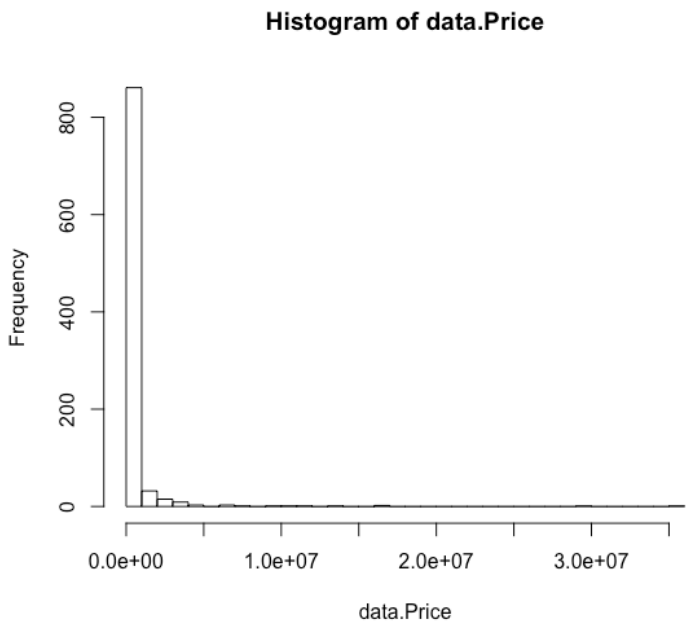

Figure 4.2.1 Histogram of price

4.3 What is the distribution of Activation Date?

The original data contains information of year, month and day and most of them are data of March. There are three ways to analyze activation date data:

4.3.1 Month Day

Data is concentrated on 15th to 28th of a month, indicating that the data may be intendedly collected from the fourteen-day period. This way promises randomness of data and also controls the potential influence on properness of data brought by different time periods. Some special events like promotional sales happen in festivals and it may influence the deal probability. Thus, the research intends to find if there is any special date that has more online advertisements, which may further affect the deal probability. According to

Table 4.3.1.1, the number of online advertisements reached the highest point on 18th and the lowest on 21st. Fluctuation exists but it is not significant on the graph. Competition becomes severer among sellers as Thus, further test is needed to verify if the month day actually influences the deal probability.

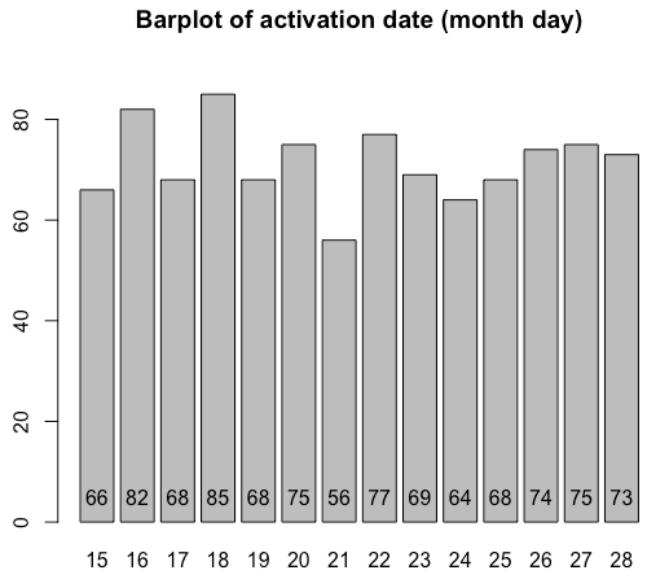

Figure 4.3.1.1 Barplot of activation date (month day)

4.3.2 Week Day

From Figure 4.3.2.1, it has been shown that fluctuation also exists from Monday to Sunday while it is not that significant. The number of advertisements is the largest on Monday while it is comparatively least on Tuesday. The difference among ad numbers on Monday, Thursday and Saturday and between that on Friday and Tuesday are slight. The barplot of activation day (week day) indicates the necessity of further testing.

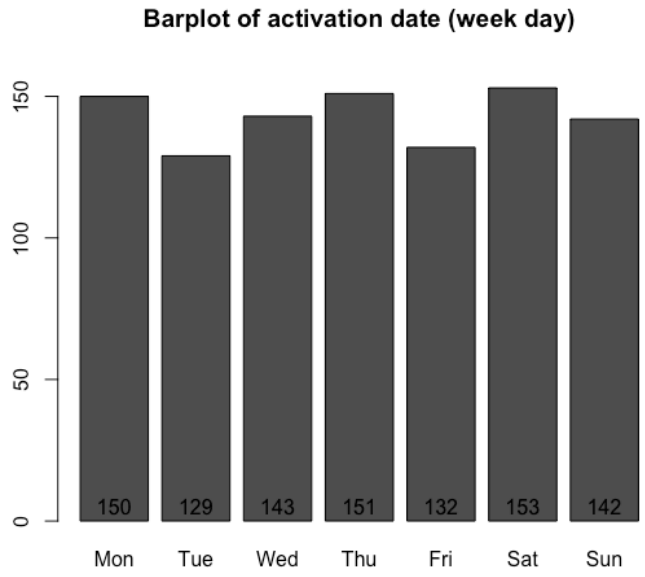

Figure 4.3.2.1 Barplot of activation date (week day)

4.3.3 Weekend/Weekdays

Weekends includes Saturday and Sunday and the rest of days are weekdays. The research wants to analyze how weekends influence the total number of advertisements. Sellers can spend more time posting advertisements online and buyers have more time to shop online on weekends (although Friday can also have some influences). Since the ratio of the number of weekdays to weekend days is 5:2, it is reasonable to assume the distribution of barplot is also 5:2. However, according to Figure 4.3.3.1 the percentage of number of advertisements posted on the weekend is greater than the ratio 5:2. Thus, further testing is needed.

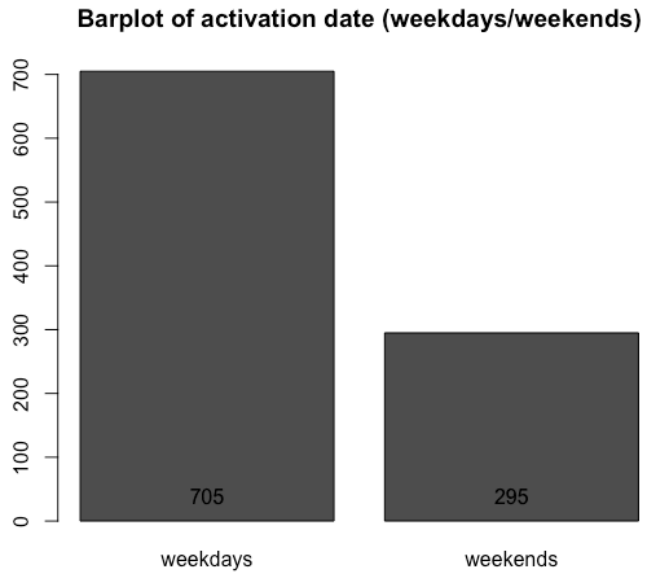

Figure 4.3.3.1 Barplot of activation date (weekdays/weekends)

4.4 What is the distribution of User Type?

There are three types of users: private, company and shop. According to Figure 4.4.1, most users are private users. Company users are much more than shop users. Online platforms with low barriers provide a convenient way for sellers to demonstrate items they want to sell. This fact can offer an explanation for large percent of private uses in Avito. On the other hand, shop users usually have physical stores and online advertising is not the only for them to sell products. Meanwhile, due to large number of private users, the completion among sellers is severe, which can lower the product price, while there are fewer shop and company users and therefore the markets are less competitive. The research assume that the user type has some influences on price and the deal probability and the assumption needs further testing.

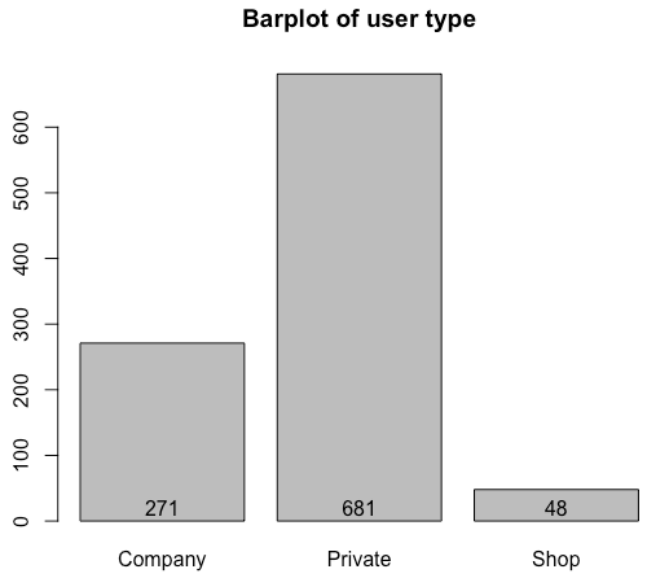

Figure 4.4.1 Barplot of user type

4.5 What is the distribution of advertisements with/without Image?

Based on Figure 4.5.1 most of advertisements use images. Whether using images can affect deal probability or not needs further testing.

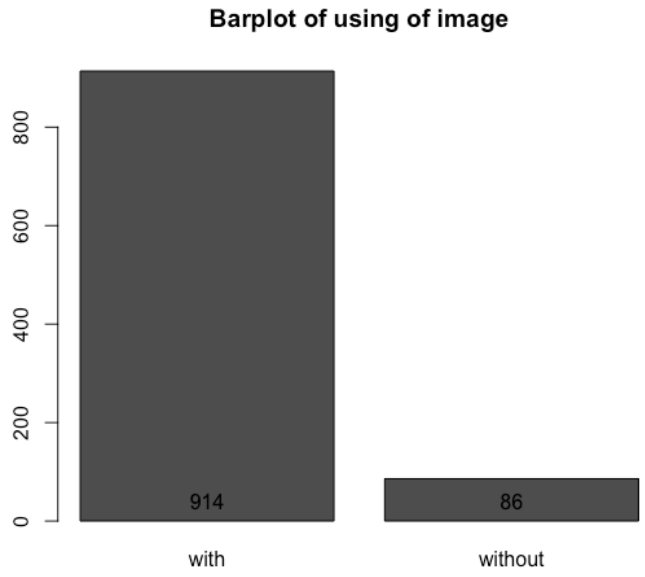

Figure 4.5.1 Barplot of using of image

4.6 What is the summary statistics of Description?

4.6.1 Word Count

From Table 4.6.1.1, the word count ranges from 0 to 384. Figure 4.6.1.1 shows that most word count data are focused on the area within 100. Only few have word account more than 20 or equal to 0. The research wants to test whether the existence of word description and the length of description have influence on the deal probability. Meanwhile, studying the data in the concentrated area i.e.0 to 50 in the histogram is also applicable.

Table 4.6.1.1 Location summary statistics of work count for description

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	7.00	17.00	31.97	36.00	384.00

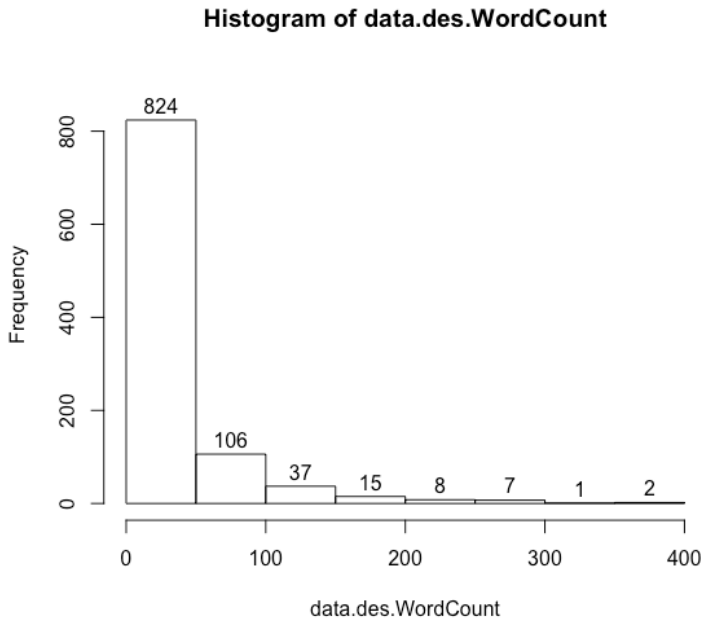

Figure 4.6.1.1 Histogram of word count for description

4.6.2 Capital Letter Count

According to Figure 4.6.2.1, most of data have capital letter count fewer than 10. The 3rd quantile is 0. Almost no sellers will choose to use only upper letters in sentences. Since the max is 39 and it increases the mean to a large extent. The distribution of data is excessively biased, data processing procedure is necessary. One way is to study the data within the first column in Table 4.6.2.1.

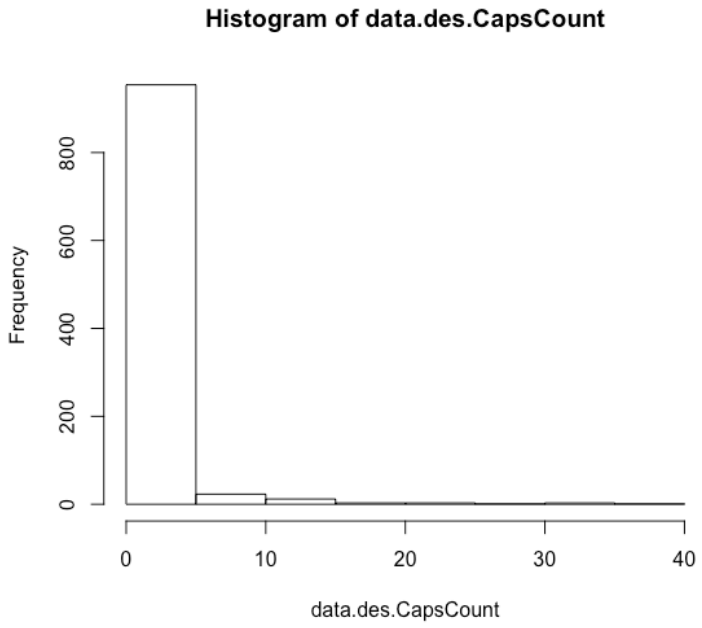

Figure 4.6.2.1 Histogram of capital count for description

Table 4.6.2.1 Location summary statistics of capital letter count for description

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	0.905	0.000	39.000

4.6.3 Digit Count

The histogram of digit count for description indicates that most of data are concentrated in the range from 0 to 10. It has almost the same shape as capital count above. The distribution of data is excessively biased. Further testing is needed. One way is to study the data within the first column in Table 4.6.3.1.

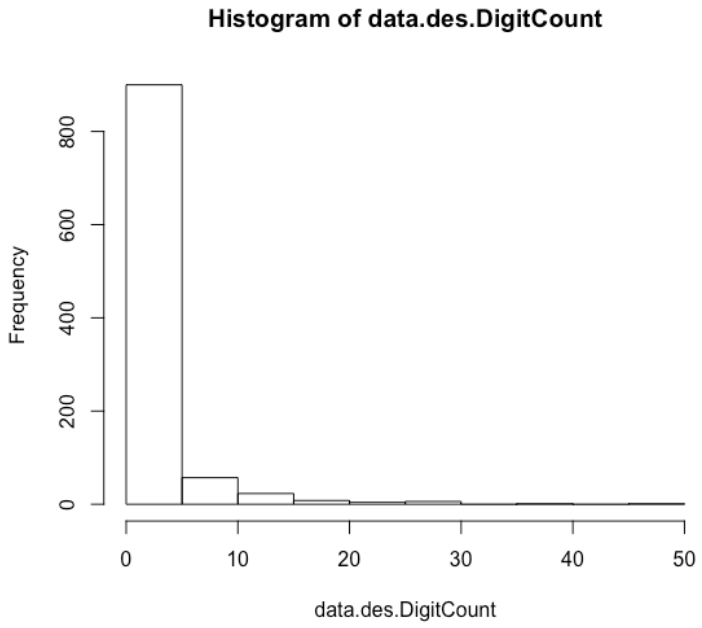

Figure 4.6.3.1 Histogram of digit count for description

Table 4.6.3.1 Location summary statistics of digit count for description

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	1.000	2.151	2.000	50.000

4.7 What is the summary statistics of Title?

4.7.1Word Count

The title word count ranges from 1 to 12. According to Figure 4.7.1.1, it shows a tendency that as the number of word count increases, the corresponding number of advertisements decreases, which shows a negative relationship. To investigate whether the title word count influences the deal probability.

Table 4.7.1.1 Location summary statistics of word count for title

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	2	3	4	5	12

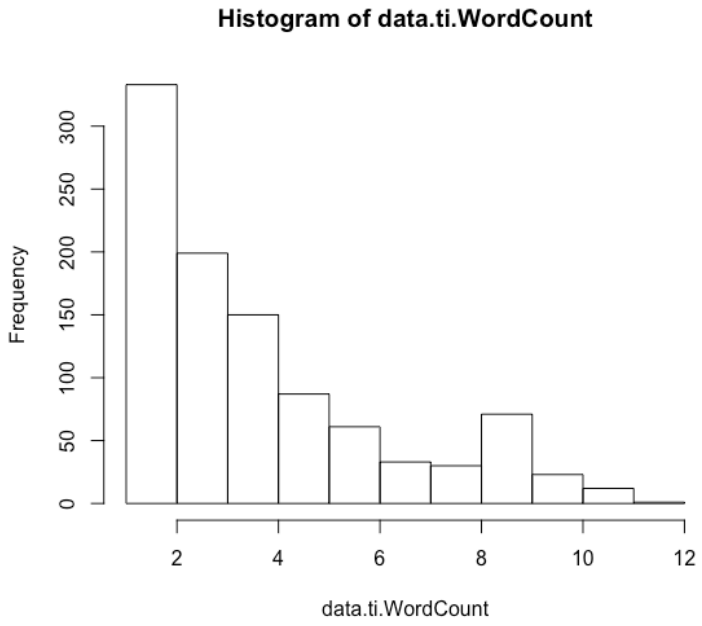

Figure 4.7.1.1 Histogram of word count for title

4.7.2 Capital Letter Count

According to Figure 4.7.2.1 and Table 4.7.2.1, most of sellers did not use capital letters in the title. The max count is 3 and the minimum is 0. The mean is 0.91 since the most of data are 0 but the max is comparatively larger.

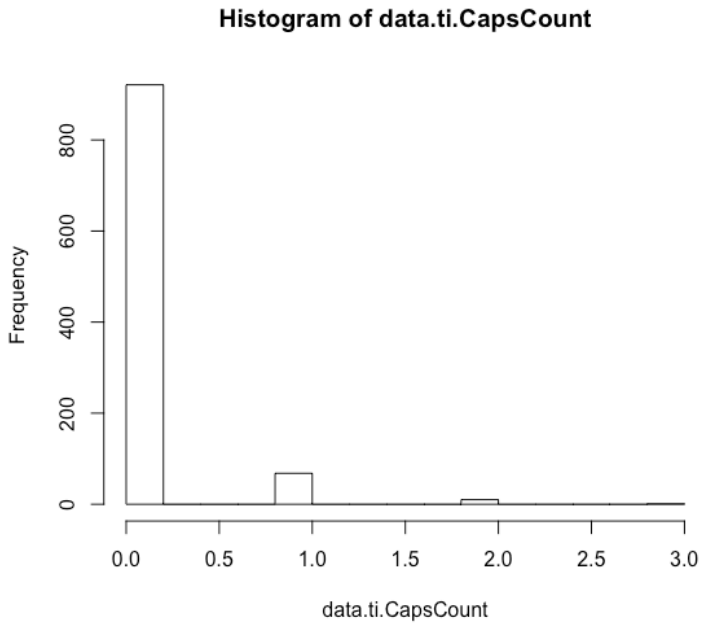

Figure 4.7.2.1 Histogram of capital count for title

Table 4.7.2.1 Location summary statistics of capital letter count for title

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	0.091	0.000	3.000

4.7.3 Digit Count

Based on Figure 4.7.3.1, the number of the digit count and frequency have negative relationship. The max is 6 while more than half of the data is 0. Most of the data are concentrated in the range between 0 and 1. The research will do more testing to investigate whether the digit count of title will influence the deal probability or not.

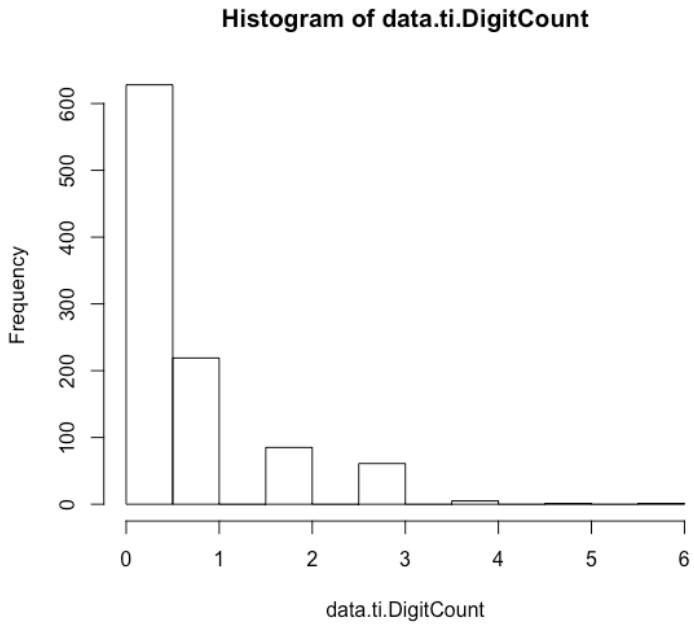

Figure 4.7.3.1 Histogram of digit count for title

Table 4.7.3.1 Location summary statistics of digit count for title

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	0.603	1.000	6.000

4.8 What is the distribution of Region?

There are 28 regions divided based on geographical area. Differences among the number of ads posted in different regions are obvious. Krasnodar region has much more advertisements posted online of than other regions, which implies that the region may have influence on the number of advertisements. The investigation will do more testing to verify the relationship.

Table 4.8.1 Count of ads posted in different regions

Region	Count
Altai region	25
Bashkortostan	50
Belgorod region	18
Chelyabinsk region	57
Irkutsk region	30
Kaliningrad region	23
Kemerovo Region Khanty-Mansiysk	21
Autonomous Okrug	16
Krasnodar region	104
Krasnoyarsk region	39
Nizhny Novgorod Region	48
Novosibirsk region	46
Omsk Region	37
Orenburg region	25
Perm Region	36
Rostov region	56
Samara Region	48
Saratov region	31
Stavropol region	34
Sverdlovsk region	57
Tatarstan	45
Tula region	16
Tyumen region	23
Udmurtia	16
Vladimir region	16
Volgograd region	27
Voronezh region	33
Yaroslavl region	23

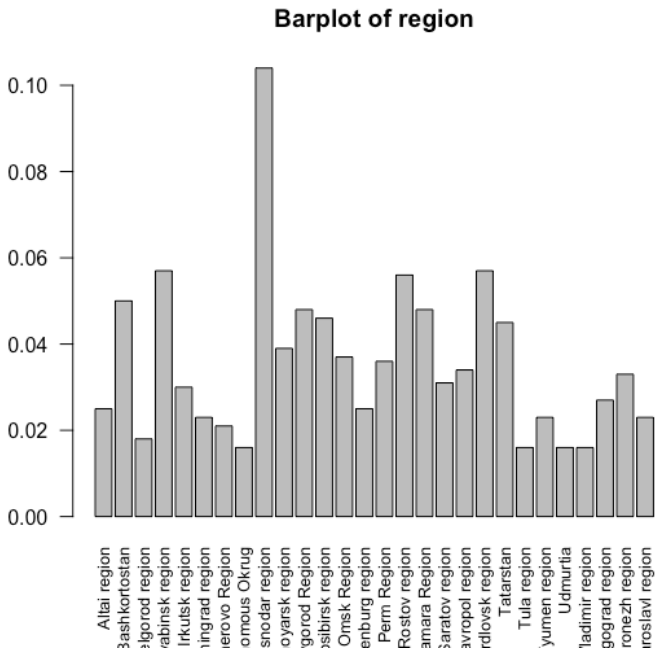

Figure 4.8.1 Barplot of region

4.8.1 City

Data are sampled from 204 cities. Figure 4.7.8.1 shows that the differences among different cities are extremely large, which indicates that it is impractical to research the deal probability based on different regions and therefore cannot provide much useful information for researchers to study.

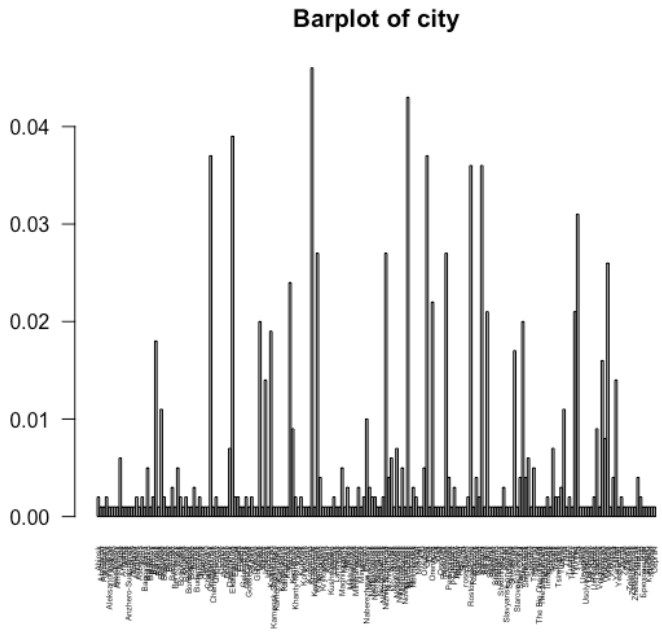

Figure 4.7.8.1 Barplot of city

4.9 What is the distribution of Parent category?

Parent Category is the top-level ad category classified by Avito's ad model. Data are sampled based on 9 parent categories that sellers usually advertise on. According to the barplot of parent category, the advertisement on Avito are the most, much more than other types of products while business only got 9 counts. Owing to huge differences among different parent categories, the investigation assumes that category can have large influence on the deal probability, which can be verified in the following part of the research.

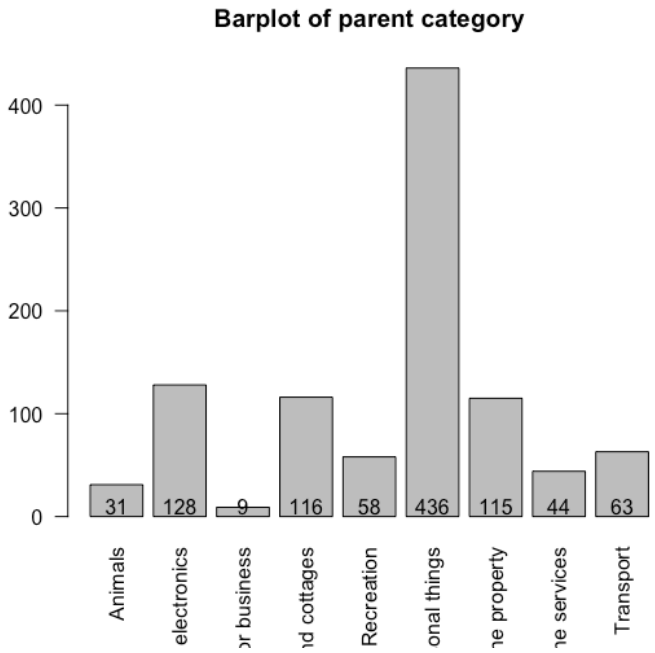

Figure 4.9.1. Barplot of parent category

Table 4.9.1 Count of ads posted in different parent categories

Parent Category	Count
Animals	31
Consumer electronics	128
For business	9
For home and cottages	116
Hobbies and Recreation	58
Personal things	436
The property	115
The services	44
Transport	63

4.10 What is the distribution of Category?

Category is the fine grain ad category as classified by Avito's ad model. According to Figure 4.10.1, the number of categories is too large to analyze. Combining different categories together is not reasonable. Therefore, either test of independence or regression is not applicable. A few categories like clothing and shoes have much higher advertisements. Daily use products tend to have much more advertisements since there are many substitutes products. Thus, classifying data based on Federal District is more proper.

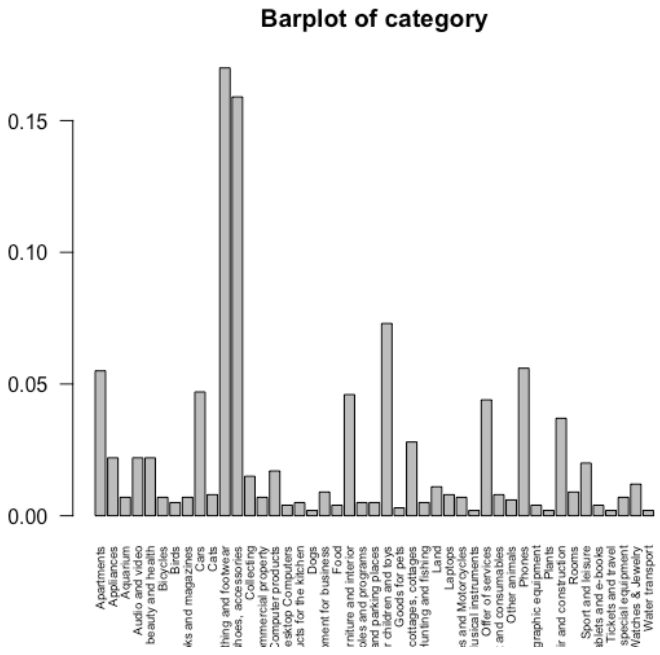

Figure 4.10.1 Barplot of category

4.11 What is the distribution of parameter1, parameter2 and parameter3?

Paremeter1, parameter2 and parameter 3 are optional parameters from Avito's model. Based on Figure 4.11.1, Figure 4.11.2 and Figure 4.11.3, categories are too detailed and such kind of classification cannot help doing the evaluation and offering useful information. The research can ignore this part.

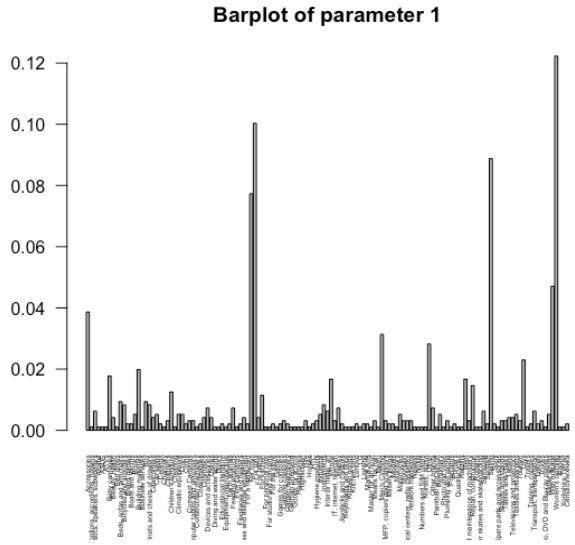

Figure 4.11.1 Barplot of parameter1

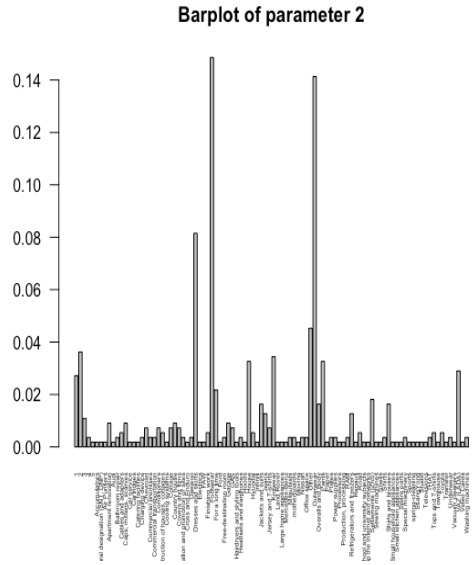

Figure 4.11.2 Barplot of parameter2

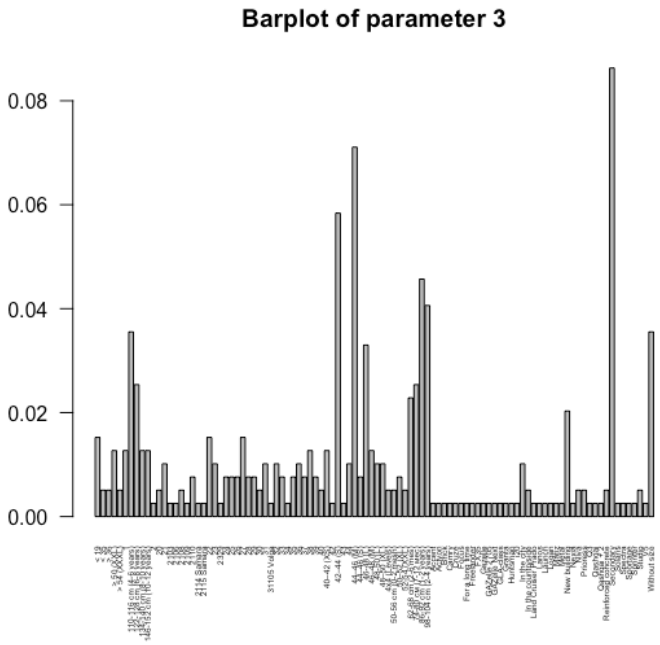

Figure 4.11.3. Barplot of parameter3

5. Test of Independence

Association of each variable that possibly affect deal probabity and binary variable deal probability. This part uses the data collected from 1000 sampled advertisement. NA values are ignored. Settlement of the data can be found in the data section. Main method used here is Pearson Chi-square test of Independence. The null hypothesis H_0 for all test of indendence is, the corresponding variable in that part is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold.

When the corresponding varible is numerical (ie: Price), the quantiles are computed and the numerical variable is converted into categorical variable by counting the number of data in each intervals between every two consequitive quantiles. As the binary varible

deal probabity (whether the probability is 0) is also a categorical varible. Contengiency table can be made and then test of independence can be used.
When the corresponding variable is categorical (ie:), Contengiency table can be directly made by counting number of data in each category and then test of independence can be used.

[Price]

The null hypothesis H_0 is that the price is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The 0.2, 0.4, 0.6, 0.8 quantiles of price variable in sampled data are 400, 1000, 2500,18000 respectively. So five interval are constructed and number of data in each corresponding interval is counted and listed in Table 5.1.1:

Table 5.1.1 Contengiency Table of price and deal probability

Price \ Deal	Deal probability is 0	Deal probability is not 0
<=400	137	56
>400 and <=1000	171	49
>1000 and <=2500	112	40
>2500 and <=18000	108	75
>18000	72	114

The result of test of independence is that X-squared = 82.092, df = 4, p-value < 2.2e-16, indicating that the null hypothesis should be rejected at significance level = 0.05. The price is not independent of whether the deal probability is 0.

[User Type]

The null hypothesis H_0 is that the user type is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since user type is a categorical data itself, the number of data in each corresponding interval can be directly counted and listed in Table 5.2.1:

Table 5.2.1 Contengiency Table of user type and deal probability

User \ Deal	Deal probability is 0	Deal probability is not 0
Company	184	87
Private	420	261
Shop	23	25

The result of test of independence is that X-squared = 7.9213, df = 2, p-value = 0.01905, indicating that the null hypothesis should be rejected at significance level = 0.05. The the user type is not independent of whether the deal probability is 0.

[Dullness of image]

The null hypothesis H_0 is that the dullness is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The numerical variable dullness is converted into categorical variable depending on whether dullness = 0. The category is not decided using quantiles in this case since the majority of data points have image with dullness = 0. So two category for dullness are constructed and number of data in each corresponding interval is counted and listed in Table 5.3.1:

Table 5.3.1 Contengiency Table of dullness and deal probability

Image \ Deal	Deal probability is 0	Deal probability is not 0
Dullness= 0	394	207
Dullness≠ 0	205	108

The result of test of independence is that X-squared = 2.4722e-30, df = 1, p-value = 1, indicating that the null hypothesis should not be rejected at significance level = 0.05. The the dullness is independent of whether the deal probability is 0.

[Whiteness of image]

The null hypothesis H_0 is that the whiteness is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The numerical variable whiteness is converted into categorical variable depending on whether whiteness = 0. The category is not decided using quantiles in this case since the

majority of data points have image with whiteness = 0. So two category for whiteness are constructed and number of data in each corresponding interval is counted and listed in Table 5.4.1:

Table 5.4.1 Contengiency Table of whiteness and deal probability

Image \ Deal	Deal probability is 0	Deal probability is not 0
Whiteness= 0	415	184
Whiteness≠ 0	184	131

The result of test of independence is that X-squared = 10.323, df = 1, p-value = 0.001314, indicating that the null hypothesis should be rejected at significance level = 0.05. The the whiteness is not independent of whether the deal probability is 0.

[Image size]

The null hypothesis H_0 is that the image size is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% quantiles of price variable in sampled data are 20016.6, 24590.4, 28770.8, 32937.0, 37025.5, 40999.4, 44986.2, 50796.4, 59006.3 respectively. So ten intervals are constructed and number of data in each corresponding interval is counted and listed in Table 5.5.1:

Table 5.5.1 Contengiency Table of image size and deal probability

Img size \ Deal	Deal probability is 0	Deal probability is not 0
<=20016.6	61	31
>20016.6 and <=24590.4	56	35
>24590.4and <=28770.8	63	28
>28770.8 and <=32937.0	57	36
>32937.0and <=37025.5	64	26
>37025.5 and <=40999.4	61	30
>40999.4 and <=44986.2	60	32
>44986.2 and <=50796.4	59	32
>50796.4 and <=59006.3	62	29
>59006.3	56	36

The result of test of independence is that X-squared = 4.4713, df = 9, p-value = 0.8777, indicating that the null hypothesis should not be rejected at significance level = 0.05. The image size is independent of whether the deal probability is 0.

[Image width]

The null hypothesis H_0 is that the image width is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since the width values are concentrated on specific values, The cutoffs of intervals are decided to be 300, 400, 500 respectively. So four intervals are constructed and number of data in each corresponding interval is counted and listed in Table 5.6.1:

Table 5.6.1 Contengiency Table of image width and deal probability

Img width \ Deal	Deal probability is 0	Deal probability is not 0
<=300	91	38
>300 and <=400	293	99
>400 and <=500	137	112
>500	78	66

The result of test of independence is that X-squared = 36.582, df = 3, p-value = 5.64e-08, indicating that the null hypothesis should be rejected at significance level = 0.05. The image width is not independent of whether the deal probability is 0.

[Image height]

The null hypothesis H_0 is that the image height is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since the height values are concentrated on specific values (360 and 480), The cutoff of intervals is decided to be 400. So two intervals are constructed and number of data in each corresponding interval is counted and listed in Table 5.7.1:

Table 5.7.1 Contengiency Table of image height and deal probability

Img height \ Deal	Deal probability is 0	Deal probability is not 0
<=400	266	196
>400	333	119

The result of test of independence is that X-squared = 25.502, df = 1, p-value = 4.419e-07, indicating that the null hypothesis should be rejected at significance level = 0.05. The image height is not independent of whether the deal probability is 0.

[Certain image dimension]

The null hypothesis H_0 is that the certain image dimension is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Firstly two dimensions (270*480 and 360*480) are tried . So two category are constructed and number of data in each corresponding interval is counted and listed in Table 5.8.1:

Table 5.8.1 Contengiency Table of 2 image dimensions and deal probability

Dimension \ Deal	Deal probability is 0	Deal probability is not 0
360*480	171	87
270*480	67	29

The result of test of independence is that X-squared = 0.24862, df = 1, p-value = 0.618, indicating that the null hypothesis should not be rejected at significance level = 0.05. The two specific image (270*480 and 360*480) dimensions are independent of whether the deal probability is 0.

Next, four most common dimensions (270*480, 360*480, 480*360, 640*360) are used . So four category are constructed and number of data in each corresponding interval is counted and listed in Table 5.8.2:

Table 5.8.2 Contengiency Table of 4 image dimensions and deal probability

Dimension \ Deal	Deal probability is 0	Deal probability is not 0
360*480	171	87
270*480	67	29
480*360	112	94
640*360	54	17

The result of test of independence is X-squared = 14.756, df = 3, p-value = 0.002037, indicating that the null hypothesis should be rejected at significance level = 0.05. The four most common specific image (270*480, 360*480, 480*360, 640*360) dimensions are not independent of whether the deal probability is 0.

[Blurness]

The null hypothesis H_0 is that the image blurness is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The 0.25, 0.5, 0.75 quantiles of blurness variable in sampled data are 279.03, 517.9, 935.09 respectively. So four intervals are constructed and number of data in each corresponding interval is counted and listed in Table 5.9.1:

Table 5.9.1 Contengiency Table of blurness and deal probability

Blurness \ Deal	Deal probability is 0	Deal probability is not 0
<=279.03	155	74
>279.03 and <=517.91	155	73
>517.91 and <=935.09	147	81
>935.09	142	87

The result of test of independence is X-squared = 2.448, df = 3, p-value = 0.4848, indicating that the null hypothesis should not be rejected at significance level = 0.05. The image blurness is independent of whether the deal probability is 0.

[Region]

The null hypothesis H_0 is that the region is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since region is a categorical data itself, the number of data in each corresponding interval can be directly counted and listed. However, there are too many region categories. To solve this problem, the "Fedral District" is applied to categorise the regions. Specifically, the Fedral district includes: Siberia:Altai region\ Irkutsk region\ Kemerovo Region\ Krasnoyarsk region\ Novosibirsk region\ Omsk Region; Volga: Bashkortostan\ Nizhny Novgorod Region\ Orenburg region\ Perm Region\ Samara Region\ Saratov region\ Tatarstan\ Udmurtia; Central: Belgorod region\ Tula region\ Vladimir region\ Voronezh region\Yaroslavl region; Ural : Chelyabinsk region\ Khanty-Mansiysk Autonomous Okrug\ Sverdlovsk region/ Tyumen region; Northwest: Kaliningrad region; South: Krasnodar region\ Rostov region\ Volgograd region; North Caucasus: Stavropol region. Then seven intervals are constructed and number of data in each corresponding interval is counted and listed n Table 5.10.1:

Table 5.10.1 Contengiency Table of Region and deal probability

Region \ Deal	Deal probability is 0	Deal probability is not 0
Volga	178	121
Ural	97	56
Siberia	130	68
Central	72	34
South	110	77
Northwest	17	6
North Caucasus	23	11

The result of test of independence is that X-squared = 6.0863, df = 6, p-value = 0.4136, indicating that the null hypothesis should not be rejected at significance level = 0.05. The reigion is independent of whether the deal probability is 0.

[Image/No Image]

The null hypothesis H_0 is that whether the advertisement contains image is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since image/no image is a categorical data itself, the number of data in each corresponding interval can be directly counted and listed in Table 5.11.1:

Table 5.11.1 Contengiency Table of image/no image and deal probability

Image \ Deal	Deal probability is 0	Deal probability is not 0
No image	28	58
With image	599	315

The result of test of independence is that X-squared = 35.156, df = 1, p-value = 3.043e-09, indicating that the null hypothesis should be rejected at significance level = 0.05. Whether the advertisement contains image is not independent of whether the deal probability is 0.

[Parent Category]

The null hypothesis H_0 is that parent category is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since parent category is a categorical data itself, the number of data in each corresponding interval can be directly counted and listed in Table 5.12.1:

Table 5.12.1 Contengiency Table of parent category and deal probability

Parent Cate \ Deal	Deal probability is 0	Deal probability is not 0
Animal	15	16
Consumer	68	60
Home	66	50
Hobby	48	10
Personal	365	71
Property	26	89
Service	7	37
Transport	28	35

The result of test of independence is that X-squared = 231.33, df = 7, p-value < 2.2e-16, indicating that the null hypothesis should be rejected at significance level = 0.05. The parent category is not independent of whether the deal probability is 0.

[Month Day]

The null hypothesis H_0 is that month day is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since month is a categorical data itself, the number of data in each corresponding interval can be directly counted and listed in Table 5.13.1:

Table 5.13.1 Contengiency Table of month day and deal probability

Month day \ Deal	Deal probability is 0	Deal probability is not 0
2017-03-15	43	23
2017-03-16	49	33
2017-03-17	50	18
2017-03-18	58	27
2017-03-19	43	25
2017-03-20	48	27
2017-03-21	39	17
2017-03-22	47	30
2017-03-23	42	27
2017-03-24	41	23
2017-03-25	42	26
2017-03-26	46	28
2017-03-27	42	33
2017-03-28	37	36

The result of test of independence is that X-squared = 12.434, df = 13, p-value =0.4924 indicating that the null hypothesis should not be rejected at significance level = 0.05. The month day is independent of whether the deal probability is 0.

[Whether on Weekend]

The null hypothesis H_0 is that whether the advertisement is on weekend is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. Since whether the advertisement is on weekend is a

categorical data itself, the number of data in each corresponding interval can be directly counted and listed in Table 5.13.1:

Table 5.13.1 Contengiency Table of whether the ad is on weekend and deal probability

IsWeekend \ Deal	Deal probability is 0	Deal probability is not 0
Weekend	189	106
Weekday	438	267

The result of test of independence is that X-squared = 0.25692, df = 1, p-value = 0.6122, indicating that the null hypothesis should be rejected at significance level = 0.05. Whether the advertisement is on weekend is independent of whether the deal probability is 0.

[Description word count]

The null hypothesis H_0 is that the description word count is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The 0.2, 0.4, 0.6, 0.8 quantiles of description word count variable in sampled data are 6, 12, 21, 45 respectively. So five intervals are constructed and number of data in each corresponding interval is counted and listed in Table 5.14.1:

Table 5.14.1 Contengiency Table of description word count and deal probability

Word count \ Deal	Deal probability is 0	Deal probability is not 0
<=6	170	57
>6 and <=12	135	60
>12 and <=21	104	81
>21 and <=45	112	86
>45	106	89

The result of test of independence is X-squared = 28.646, df = 4, p-value = 9.225e-06, indicating that the null hypothesis should be rejected at significance level = 0.05. The decription word count is not independent of whether the deal probability is 0.

[Title word count]

The null hypothesis H_0 is that the title word count is independent of whether the deal probability is 0. The alternative hypothesis H_1 will be that this independence doesn't hold. The 0.2, 0.4, 0.6, 0.8 quantiles of title word count variable in sampled data are 2, 3, 4, 6 respectively. So five intervals are constructed and number of data in each corresponding interval is counted and listed in Table 5.15.1:

Table 5.15.1 Contengiency Table of title word count and deal probability

Word count \ Deal	Deal probability is 0	Deal probability is not 0
<=2	239	94
>2 and <=3	119	80
>3 and <=4	98	52
>4 and <=6	93	54
>6	78	93

The result of test of independence is X-squared = 37.439, df = 4, p-value = 1.463e-07, indicating that the null hypothesis should be rejected at significance level = 0.05. The title word count is not independent of whether the deal probability is 0.

Short summary: By test of independence, one can see that the variables—including price, user type, image width, image height, image dimension, image whiteness, image/no image, parent category, description word count and title word count—are not independent of the deal probability. So they have higher probability to become candidates of explanatory variables in the regression model. Since the data is skewed and the variables might have affect on each others, the test of independence is not enough to eliminate those variables for which the null hypothesis is not rejected from the regression model in which the study is interested. Further elimination process according to p-value will be applied in a following section to eliminate less relavent variables.

6. Bootstrap

For each of the variables, the null hypothesis H_0 in this section is that the data within category selected has a mean equal to the mean of the whole sample, H_1 is the means are different.
In this section, two method—Boostarp_CI_Method(data1,data2,n) and Boostarp_Zero_Method(data1,data2,n)—are designed to perform bootstrap on the data based on resample size = length(data1) and from data2 without replacement. The precondition of applying those two methods is that data1 has a smaller size than data2.

The underlying idea of the Boostarp_CI_Method(data1,data2,n) is that one firstly choose a category from the sample and compute its mean, then resample with size of that category size from the whole sample, after resampling, compute the resample means to construct 95% confidence interval of sample mean and see whether the sample mean of that category (data1) is statistically significantly different from the sample mean of whole estimated by bootstrap. If the difference is significant, it indicates that the categorical variable has an statistically significant affect on the sample since data with in that category will have a different mean from the whole sample mean.

However, since deal probability data contains lots of zeros, it might to some extent affect the sample mean. So the Boostarp_Zero_Method(data1,data2,n) is designed. The difference is in that this method count the number of zeros from resamples instead of resample mean. Similar decision rule is applied here. This method should intuitively be more accurate than Boostarp_CI_Method(data1,data2,n) since it elimiates the effect of too many zeros. More importantly, data1 and data2 are supposed to be independent in this method, but data1 and data2 are actually from the same sample in this case as they are just different categories from the whole sample. This might cause error in this method. Besides, data1 sample size will affect data2 mean to some extent if the size of data1 is too large compared that of to data. If data1 covers most propotion of sample, data2 mean is supposed to be close to that of data1 (which is also close to the whole sample mean) unless data2 are all outliners (which seems impossible).

For each variable, both method is applied to data and each category is selected to be data1 in the test (so that data2 is the other categories):

[With/without image]

For the categorical variable—with/without image, the test using Boostarp_CI_Method fails to reject that the data from specifc categories has the same mean as the whole sample, Boostarp_Zero_Method reject that the data from specifc categories has the same mean as the whole sample. In this situation, it can be seen that Boostarp_Zero_Method makes a difference and it is intuitively more accurate, so Zero_Method is used as the major decision rule in this case. The test rejects that the data from specifc categories has the same mean as the whole sample, indicating whether the advertisement has an image will affect the deal probability. This test has the same as the test of independence used in section 5.

[Description word count]

For the numerical variable—description word count, The 0, 0.25, 0.5, 0.75, 1 quantiles of description word count variable in sampled data are 2, 8, 17, 37, 386 respectively. So the numerical data is converted into five categories. The majority of the tests using Boostarp_CI_Method rejects that the data from the four categories respectively has the same mean as the whole sample. indicating whether the description word count will affect the deal probability. This test has the same as the test of independence used in section 5.

[Title word count]

For the numerical variable—title word count, The 0, 0.25, 0.5, 0.75, 1 quantiles of title word count variable in sampled data are 1, 2, 3, 5, 12 respectively. So the numerical variable is converted into five categories. The majority of the tests using Boostarp_CI_Method rejects that the data from four categories respectively has the same mean as the whole sample. indicating whether the title word count will affect the deal probability. This test has the same as the test of independence used in section 5.

Short summary: Bootstrap is applied to further test the independence because the data from sample are mostly skewed and Bootstrap can capture the skewness. After applying bootstrap method on three varaibles—With/without image, Description word count and Title word count, it can be concluded from the test that all three variables have effect on the deal probability. Tests in this section all have the same result as the test of independence used in section 5, indicating the outcomes and conclusion in section 5 are to some extent accurate. Besides, As for the problem of the testing method in this part, investiagtions are conducted to solve the problem and it will be further discussed in "extra credit" section.

7. Nonzero deal probability analysis

The distribution of deal probility is shown in Figure 7.0.1:

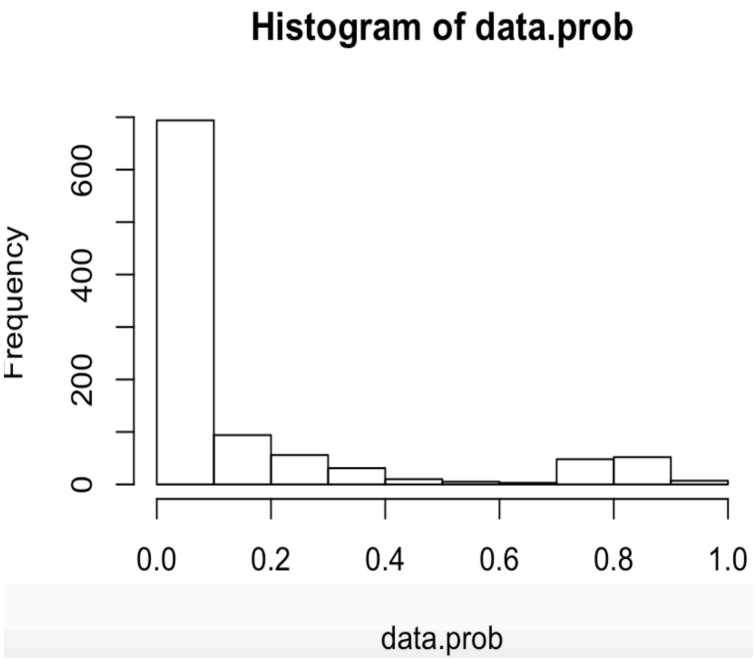

Figure 7.0.1 Histogram of data.prob

The distribution of non-zero deal probability is shown in Figure 7.0.2:

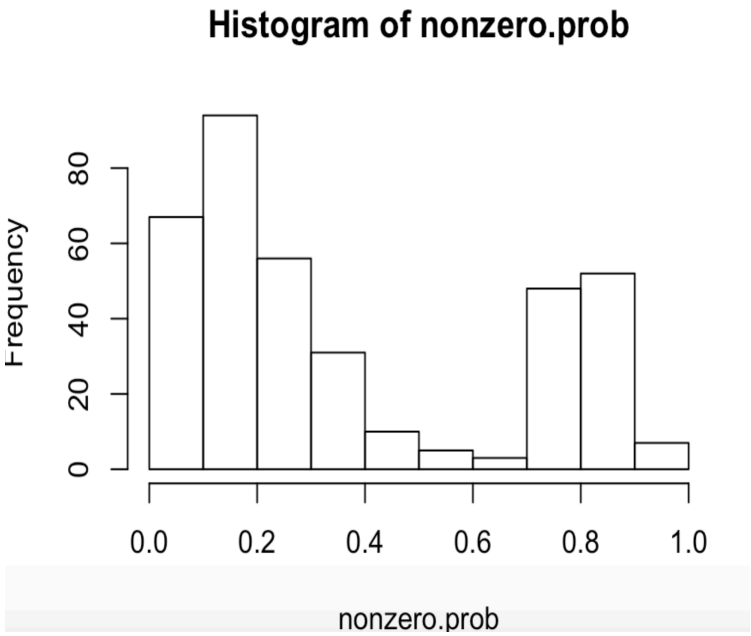

Figure 7.0.2 Histogram of data.prob

7.1 Test if the distribution of deal probability follows exponential distribution

In this section, random exponential distribution is generated based on the mean of deal probability. Then Goodness of Fit Test is conducted to test whether the deal probability distribution follows an exponential distribution generated based on mean of deal probability. H_0 is that the deal probability distribution follows an exponential distribution generated based on mean of deal probability. The alternative hypothesis H_1 will be that they do not have the same distribution. The result is X-squared = 7289.8, df = 19, p-value < 2.2e-16, so the null hypothesis should be rejected at significance level = 0.05. The deal probability distribution does not follow exponential distribution.

7.2 Test if the distribution of deal probability follows uniform distribution

In this section, random uniform distribution is generated based on the mean of deal probability. Then Goodness of Fit Test is conducted to test whether the deal probability distribution follows an uniform distribution generated based on mean of deal probability.

H_0 is that the deal probability distribution follows a uniform distribution generated based on mean of deal probability. The alternative hypothesis H_1 will be that they do not have the same distribution. The result is X-squared = 7289.8, df = 19, p-value < 2.2e-16, so the null hypothesis should be rejected at significance level = 0.05. The deal probability distribution does not follow uniform distribution.

7.3 Simulate similar distribution

In this section, two normal distributions is combined to simulate the distribution of non-zero deal probability. The simulated data is Figure 7.3.1

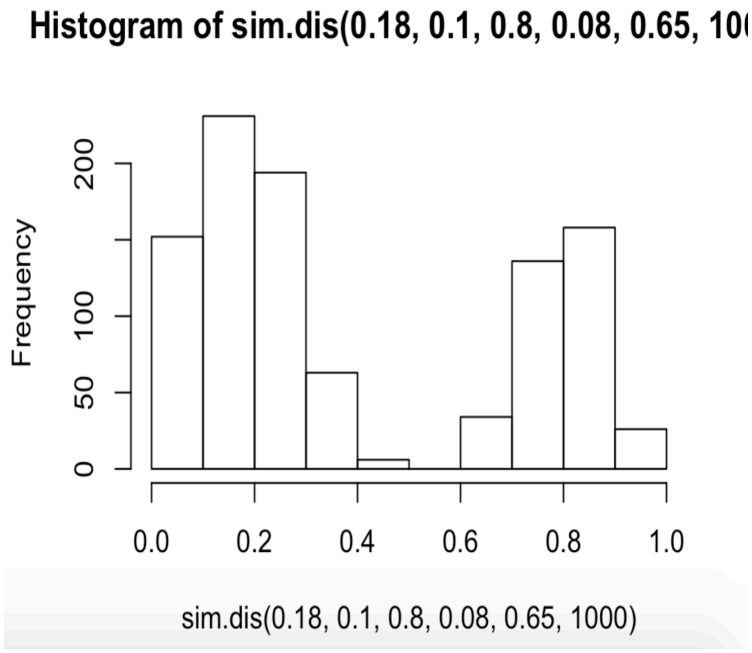

Figure 7.3.1 Histogram of simulated distribution

7.4 Bootstrap methods

In this section, three bootstrap methods are designed to investigate further on non- zero deal probability; more details can be found in code. They do not have huge impact on the regression model so one who interested in the investigation can view the code for more details.

8. Logistic Regression

9. Numerical Regression

10. Extra Section

10.1 The Try-Out Process: Tukey's Honest Significant Difference

As Section 6 mentioned, the bootstrap methods used have a problem. The study tried another method to solve this problem, although it fails, we feel like it is worth noting here as part of the try-out process.

The method we chose is Tukey's Honest Significant Difference. The motivation is that the study wants to evaluate the different sections inside each variable's range and see which section within each variable has more influence on the target deal probability. The starting point is that, since a large proportion of distributions of ad features is very skewed and concentrates in the small values, the concentrated area may have a large impact on the target and the rest may be outliers. Tukey's Honest Significant Difference is applicable to test the pairwise difference in distribution means among multiple groups.

H_0: pairwise means are equivalent vs. H_1: otherwise

However, it has three assumptions:

1.The samples among groups are independent

2.The distributions of each group is close to normal

3.Homoscedasticity

These assumptions are important and can not be cured by large sample size. The investigation first tried to regard of these assumptions and test, but the confidence intervals are quite wide and all fail to reject the null (shown in Figure 10.1.1). This is reasonable since the distribution of deal probability for the whole selected data is very positive skewed, and the groups are obviously not independent because they have to sum to a fixed number. These results also contradict to the results gained using Test of

Independence and bootstrap. This also helps emphasize the significance of the assumptions.

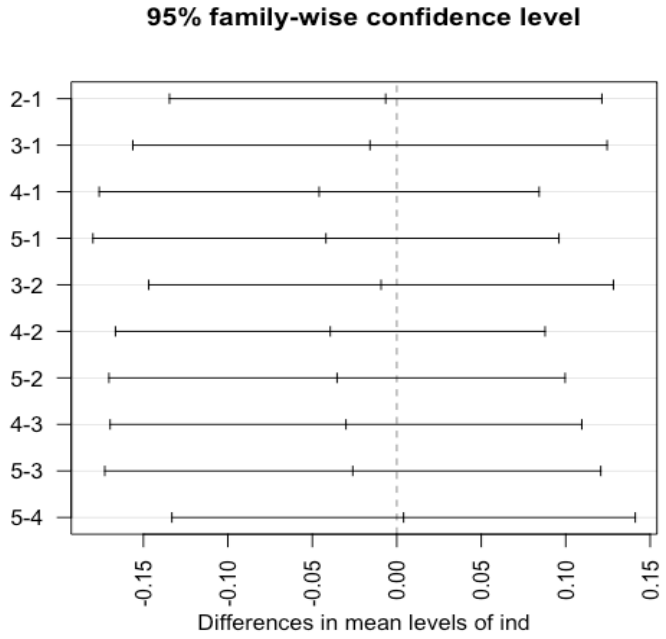

Figure 10.1.1 95% family-wise confidence intervals given be Tukey's Honest Significant Difference (X = price, using all deal possibility)

To overcome this problem and try to satisfy the assumptions, the study get rid of the data with deal probability equal 0 and split the rest of deal distribution into two groups based on the 0.5 cutoff. The handling of data is based on the distribution of deal probability that it has a mode in 0 and two modes almost symmetric around 0.5 after getting rid of all 0's. The distribution of deal probability for each group after this handling of data is more close to normal, although it still seems apart from normal. The independence and homoscedasticity are still hard to satisfied. However, the investigation sees a big improvement in results (see Figure 10.1.2).

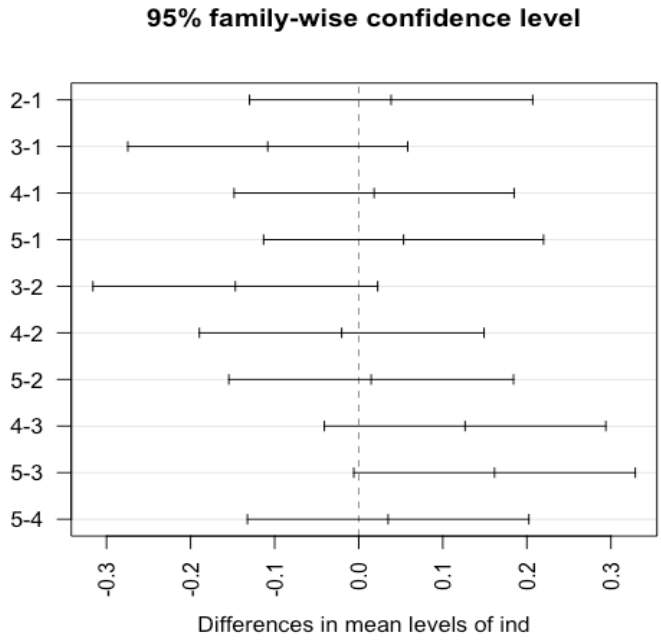

Figure 10.1.2 95% family-wise confidence intervals given be Tukey's Honest Significant Difference (X = price, using only 0<deal probability<0.5)

Although the study tries this method for many other variables such as month day, week day, etc, since the assumptions are not satisfied and the results are valid or convincing, the investigation will not include those results here.

Hopefully, someone have better ideas that may help resolve the problems for both bootstrap and Tukey's methods, so that the different influences of each subsection in one variable can be tested and evaluated for future uses.

10.2 The Idea for multiple testing

We plan to use Benjamini-Yekutieli procedure to control the False Discovery Rate through adjusting p-values of all the hypothesis tests mentioned in the investigation (FDR). Unfortunately, we do not have enough time. Hopefully, we can proceed this in the future.

11. Theory

Hypothesis Tests

Test of independence (To test whether two variables are independent)*[5]:*

Suppose that n observations are taken on a sample space partitioned by the events $A_1, A_2, \cdots A_r$ and also by events $B_1, B_2, \cdots B_c$. Let $p_i = P(A_i)$, $q_j = P(B_j)$, and $p_{ij} = P(A_i \cap B_j)$ for $i = 1,2, \cdots, r$; $j = 1,2, \cdots, c$. Let X_{ij} denote the number of observations belonging to the intersection $A_i \cap B_j$. To test H_0: the A_i's are independent of the Bj's, calculate the test statistic

$$d_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(k_{ij} - n\widehat{p_i}\widehat{q_j})}{n\widehat{p_i}\widehat{q_j}}$$

where k_{ij} is the number of observations in the sample that belong to $A_i \cap B_j$, $i = 1,2, \cdots, r$; $j = 1,2, \cdots, c$ and $\widehat{p_i}$, $\widehat{q_j}$ are the maximum-likelihood estimates for p_i and q_j , respectively. The null hypothesis should be rejected at the α level of significance if

$$d_2 \geq \chi^2_{1-\alpha,(r-1)(c-1)}$$

(Analogous to the condition stipulated for all other goodness-of-fit tests, it will be assumed that $n\widehat{p_i}\widehat{q_j} \geq 5$ for all *i* and *j*.)

Goodness-of-Fit test [5]

Suppose that a random sample of *n* observations is taken from $f_Y(y)$ $[or\ p_X(k)]$, a *pdf* having *s* unknown parameters. Let $r_1, r_2, \cdots, r_t$ be a set of mutually exclusive ranges (or outcomes) associated with each of the n observations. Let $\hat{p}_i$= estimated probability of r_i, $i = 1, 2, \cdots, t$ (as calculated from $f_Y(y)$ $[or\ p_X(k)]$, after the *pdf*s' unknown parameters have been replaced by their maximum likelihood estimates). Let X_i denote the number of times that

$r_i$occurs, , $i = 1, 2, \cdots, t$. Then

the random variable $D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$ has approximately a χ^2 distribution with $t - 1 - s$ degrees of freedom. For the approximation to be fully adequate, the r_i's should be defined so that $n\hat{p}_i \geq 5$ for all i.

to test $H_0: f_Y(y) = f_0(y)$ $[or\ H_0: p_X(y) = p_0(y)]$ at the α level of significance, calculate $d_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_{i_0})^2}{n\hat{p}_{i_0}}$

Where $k_1, k_2, \cdots, k_t$ are the observed frequencies of $r_1, r_2, \cdots, r_t$ respectively, and $n\hat{p}_{1_0}, n\hat{p}_{2_0}, \cdots, n\hat{p}_{t_0}$ are the corresponding estimated expected frequencies based on the null hypothesis. If $d_1 \geq \chi^2_{1-\alpha, t-1-s}$, H_0 should be rejected. (The r_i's should be defined so that $n\hat{p}_i \geq 5$ for all i.)

Significance test

$H_0: b = 0$ $(X\ is\ not\ linearly\ associated\ with\ Y\ in Y = a + bX + \varepsilon)$

$H_1: otherwise$

Compute test statistic: $T = \frac{\hat{b}}{SE(\hat{b})} \sim t_{n-2}$

Decision rule: Reject $H_0\ if\ |T_{obs}| > T_{\alpha/2, n-2}$

Or compute p-value $= 2P(T > T_{obs})$, reject when p-value $< \alpha$

Numerical summaries:

Sample mean: for a list of numbers $x_1, x_2, \cdots x_n$, *sample mean* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of the data. The sample mean is frequently used as the numerical summary of the location of the data

Graphical summaries:

Histogram: A representation of distribution of numerical data where normally the frequency is on the Y-axis and the variable to be investigated is on the X-axis. The histogram will display the frequency of each possible value (interval) of the variable and thus resemble the distribution of that data.

Box plot: A method of graphically depicting groups of numerical data through their quartiles. It will include median, 1st quartile, 3rd quartile, maximum and minimum of the group of data.

Regression: $Y = a + bX + \varepsilon$ *(simple linear form)*

Correlation: Computed by $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[XY]-E[X]E[Y]}{\sigma_X \sigma_Y}$.

$Corr(X,Y)$ describes the strength of the linear association between two variables X and Y, $Corr(X,Y) = 0$ indicates no linear association between X and Y.

Scatterplot: A type of mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. It's useful to make a scatterplot first before fitting data with regression model since it's easier to observe whether there might be a linear relationship between two variables directly from the plot.
Example of scatterplot:

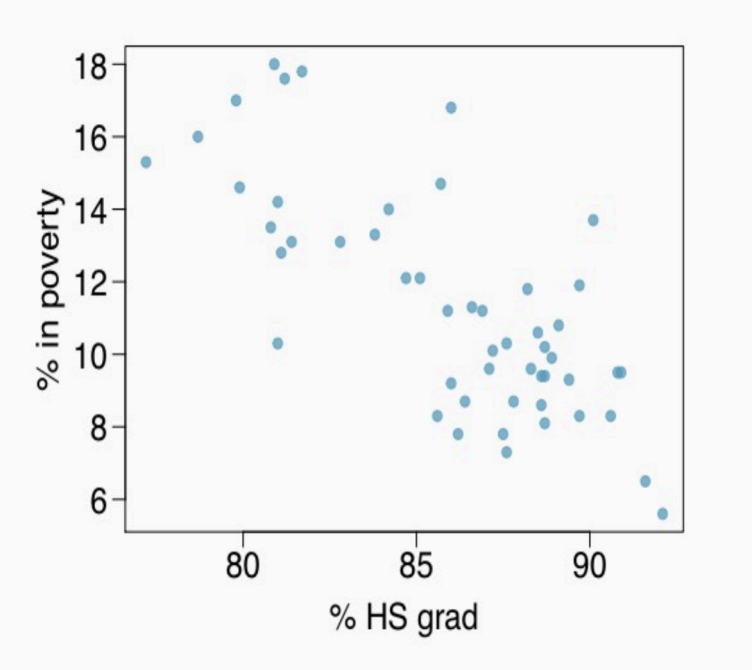

Figure 5.1.1 Scatterplot of % in poverty versus % high school graduates

Response variable: The variable changing corresponding to explanatory variable, that is Y in $Y = a + bX + \varepsilon$

Explanatory variable: The independent variable, that is X in $Y = a + bX + \varepsilon$

Residuals*:* Leftovers from the model fit: $Data = Fit + Residual$, that is ε in the regression model $Y = a + bX + \varepsilon$. It measures the difference between the observed (y_i) and predicted $\hat{y}_i$, that is $e_i = y_i - \hat{y}_i$.

$\boldsymbol{R^2}$***:*** Measures the percentage of variability in the response variable is explained by the model. Computed by $R = \text{empirical correlation coefficient} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i \cdot Y_i) - \bar{X}\bar{Y}}{s_X s_Y}$

Best-fit line (OLS): A line that minimize the sum of squared residuals – ***Least Squares***. That is, $Y = \hat{a} + \hat{b}X$ where $(\hat{a}, \hat{b}) = \underset{a,\,b}{argmin} \sum_{i=1}^{n}(Y_i - a - bX_i)^2$.

(1) The ***slope*** of regression, which measures the linear relation between *X* and *Y*, can be computed as $\hat{b} = \frac{s_X}{s_Y} R$, where R is the estimator of $Corr(X, Y), R = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i Y_i) - \bar{X}\bar{Y}}{s_X s_Y}$. (2) The ***intercept***, where the regression line intersects with the y-axis, can be computed by $\hat{a} = \bar{Y} - \hat{b}\bar{X}$.

(3) Conditions for the least square line

<1> Linearity: the relation of *X* and *Y* should be nearly linear

<2> Nearly normal residuals: if ε_i is normally distributed, OLS estimator is a MLE

<3> Constant variability: the true ε_i should not depend on *i*, there is no pattern of the scatter of residuals

All three conditions are satisfied in this study so one can fit OLS regression model to the data.

Prediction: As long as the regression $Y = \hat{a} + \hat{b}X$ is fitted to the data set, for given *x*, one can predict corresponding *y* with $\hat{y} = \hat{a} + \hat{b}x$.

Prediction interval: An estimate of an interval in which future observations will fall, with certain probability, given what has already been observed. It resembles the confidence interval for the value of response variable at given value of explanatory variable in this case study. The formula for prediction interval in this case study is

$$PI_Y(L, R)$$

$$L = \left(\hat{a} + \hat{b}X\right) - Z_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{m} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

$$R = \left(\hat{a} + \hat{b}X\right) + Z_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{m} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Theorem 1: $Var\left(\hat{Y} - Y\right) = \sigma^2\left(\frac{1}{n} + 1 + \frac{(X-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)$

Proof:

We have $a = \bar{Y} - b\bar{X} - \bar{\varepsilon}$ and $\hat{a} = \bar{Y} - \hat{b}\bar{X}$

$$Var(\hat{Y} - Y) = Var(\hat{a} - a + (\hat{b} - b)X - \varepsilon)$$
$$= Var(\hat{a} - a + (\hat{b} - b)X) + \sigma^2$$
$$= Var\left(\bar{\varepsilon} + (\hat{b} - b)(X - \bar{X})\right) + \sigma^2$$
$$= \frac{\sigma^2}{n} + \sigma^2 + (X - \bar{X})^2 Var(\hat{b} - b)$$

We have $Y_i = a + bX_i + \varepsilon_i$ and $\bar{Y} = a + b\bar{X} + \bar{\varepsilon}$

$$Y_i - \bar{Y} = b(X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon})$$

$$\hat{b} = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n {X_i}^2 - (\sum_{i=1}^n X_i)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})[b(X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon})]}{\sum_{i=1}^n (X_i - \bar{X})^2} = b + \frac{\sum_{i=1}^n (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

where $\sum_{i=1}^n (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon}) = \sum_{i=1}^n (X_i - \bar{X})\varepsilon_i - \bar{\varepsilon}(\sum_{i=1}^n X_i - n\bar{X})$

$$= \sum_{i=1}^n (X_i - \bar{X})\varepsilon_i$$

so $\hat{b} - b = \frac{\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$

$$Var(\hat{b} - b) = \frac{1}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} Var(\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i)$$

$\varepsilon_i, \varepsilon_j, X_i, X_j$ independent if $i \neq j, Var(\varepsilon_i) = \sigma^2$

$$Cov((X_i - \bar{X})\varepsilon_i, (X_j - \bar{X})\varepsilon_j)$$
$$= E[((X_i - \bar{X})\varepsilon_i - E[(X_i - \bar{X})\varepsilon_i]) \cdot ((X_j - \bar{X})\varepsilon_j - E[(X_j - \bar{X})\varepsilon_j])]$$
$$= E[(X_i - \bar{X})\varepsilon_i \cdot (X_j - \bar{X})\varepsilon_j] = 0$$

$$Var(\hat{b} - b) = \frac{1}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \sum_{i=1}^n (X_i - \bar{X})^2 Var(\varepsilon_i) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

so $Var(\hat{Y} - Y) = \frac{\sigma^2}{n} + \sigma^2 + (X - \bar{X})^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$

$$= Var(\varepsilon) + Var(\bar{\varepsilon}) + Var((\hat{b} - b)(X - \bar{X}))$$
$$= \sigma^2 \left(\frac{1}{n} + 1 + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

Theorem:

Residual sum of squares (RSS)

$$= \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \widehat{Y}_{\iota})^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota})^2 + k\sum_{i=1}^{m}(\widehat{Y}_{\iota} - \overline{Y}_{\iota})^2$$

Proof:

$\sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \widehat{Y}_{\iota})^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota} + \overline{Y}_{\iota} - \widehat{Y}_{\iota})^2$

$= \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota})^2 - 2\sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota})(\widehat{Y}_{\iota} - \overline{Y}_{\iota}) + \sum_{i=1}^{m}\sum_{j=1}^{k}(\widehat{Y}_{\iota} - \overline{Y}_{\iota})^2$

Since $\sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota})(\widehat{Y}_{\iota} - \overline{Y}_{\iota})$

$= \sum_{i=1}^{m}\sum_{j=1}^{k} Y_{ij}\cdot\widehat{Y}_{\iota} - \sum_{i=1}^{m}\sum_{j=1}^{k} Y_{ij}\cdot\overline{Y}_{\iota} - \sum_{i=1}^{m}\sum_{j=1}^{k}\overline{Y}_{\iota}\cdot\widehat{Y}_{\iota} + \sum_{i=1}^{m}\sum_{j=1}^{k}\overline{Y}_{\iota}^{\,2}$

$= \sum_{i=1}^{m}\widehat{Y}_{\iota}\sum_{j=1}^{k} Y_{ij} - \sum_{i=1}^{m}\overline{Y}_{\iota}\sum_{j=1}^{k} Y_{ij} - k\sum_{i=1}^{m}\overline{Y}_{\iota}\cdot\widehat{Y}_{\iota} + k\sum_{i=1}^{m}\overline{Y}_{\iota}^{\,2}$

$= k\sum_{i=1}^{m}\widehat{Y}_{\iota}\cdot\overline{Y}_{\iota} - k\sum_{i=1}^{m}\overline{Y}_{\iota}^{\,2} - k\sum_{i=1}^{m}\overline{Y}_{\iota}\cdot\widehat{Y}_{\iota} + k\sum_{i=1}^{m}\overline{Y}_{\iota}^{\,2}$

$= 0$

We have $\sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \widehat{Y}_{\iota})^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota})^2 + \sum_{i=1}^{m}\sum_{j=1}^{k}(\widehat{Y}_{\iota} - \overline{Y}_{\iota})^2$

$= \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{\iota})^2 + k\sum_{i=1}^{m}(\widehat{Y}_{\iota} - \overline{Y}_{\iota})^2$

Logistic regression

$$P(Y = 1|X = x) = \frac{\exp(\beta^T X)}{1+\exp(\beta^T X)} = \frac{1}{1+\exp(-\beta^T X)}$$

$$P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = \frac{1}{1+\exp(\beta^T X)}$$

$$\log\left(\frac{P(Y = 1|X)}{P(Y = 0|X)}\right) = \beta^T X$$

Since $logit(X) = \log\left(\frac{X}{1-1}\right)$, $expit(X) = logit^{-1}(X) = \frac{\exp(X)}{1+\exp(X)}$

$$logit\big(P(Y = 1|X)\big) = \log\left(\frac{P(Y = 1|X)}{P(Y = 0|X)}\right) = \beta^T X$$

$$e^{\log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)} = e^{\beta^T X}$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta^T X}$$

$$P(Y=1|X)(1+e^{\beta^T X}) = e^{\beta^T X}$$

$$P(Y=1|X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$$

$$P(Y=0|X) = \frac{1}{1+e^{\beta^T X}}$$

$E(Y|X) = P(Y=1|X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$ like Bernoulli distribution

$$Var(Y|X) = P(Y=1|X)\big(1-P(Y=1|X)\big) = \frac{e^{\beta^T X}}{(1+e^{\beta^T X})^2} = \frac{1}{2+e^{\beta^T X}+e^{-\beta^T X}}$$

Logistic regression and case control studies

Set a binary maker $M \in \{0,1\}$ for each group of (X, Y)

$$P(M=m|Y) = \frac{\exp\left((\alpha+\beta Y)m\right)}{1+\exp\left(\alpha+\beta Y\right)},\ m \in \{0,1\}$$

$$\log\left(\frac{P(M=m|Y=1)}{P(M=m|Y=0)} \cdot \frac{P(Y=1)}{P(Y=0)}\right) = \log\left(\frac{P(Y=1|M=m)}{P(Y=0|M=m)}\right)$$

$$= \log\left(\frac{\frac{\exp\left((\alpha+\beta Y)m\right)}{1+\exp\left(\alpha+\beta Y\right)}}{\frac{\exp\left((\alpha+\beta)m\right)}{1+\exp\left(\alpha\right)}} \cdot \frac{P(Y=1)}{P(Y=0)}\right) = \beta m + \log\left(\frac{1+\exp\left(\alpha\right)}{1+\exp\left(\alpha+\beta\right)} \cdot \frac{P(Y=1)}{P(Y=0)}\right)$$

$= \beta m + \theta$ (θ is constant, so for different M, we have different intercept)

Since $P(Y_i = y_i) = \left(\frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}\right)^{y_i}$

$$L(\beta|Y,X) = log\prod_{i=1}^{n}(\frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}})^{Y_i}$$

$$= \sum_{Y_i=1}\beta^T X_i - \sum_i \log(1+\exp(\beta^T X_i))$$

$$\frac{\partial L}{\partial \beta} = \sum_{Y_i=1} X_i - \sum_i \frac{\exp(\beta^T X_i)}{1+\exp(\beta^T X_i)} X_i$$

$$\beta^T X_i = \beta_0 X_{i0} + \cdots + \beta_p X_{ip}\ , \frac{\partial(\beta^T X_i)}{\partial \beta} = X_i$$

Hessian Matrix: $H(\beta|Y,X) = \frac{\partial^2 L}{\partial\beta\beta^T} L(\beta|Y,X) = -\sum_i \frac{\exp(\beta^T X_i)}{(1+\exp(\beta^T X_i))^2} X_i\, X_i{}^T$

where $\frac{\exp(\beta^T X_i)}{(1+\exp(\beta^T X_i))^2} > 0$

For non-zero vector $\vec{a} \in \mathbb{R}^{P+1}, a = \begin{bmatrix} a_0 \\ \vdots \\ a_p \end{bmatrix}$

$$\vec{a}^T X_i = [a_0 \quad \cdots \quad a_p]\begin{bmatrix} X_{i0} \\ \vdots \\ X_{ip} \end{bmatrix} = a_0 X_{i0} + \cdots + a_p X_{ip}$$

$$\overrightarrow{X_\iota}^T \vec{a} = [X_{i0} \quad \cdots \quad X_{ip}]\begin{bmatrix} a_0 \\ \vdots \\ a_p \end{bmatrix} = a_0 X_{i0} + \cdots + a_p X_{ip}$$

$$\sum_i \vec{a}^T X_i \overrightarrow{X_\iota}^T \vec{a} > 0$$

So we get idea that $\vec{a}^T H(\beta|Y,X)\vec{a} < 0$ for any non-zero vector $\vec{a}$

$H(\beta|Y,X)$ is negative definite

Proposition: for continuously differentiable function f: $A \to \mathbb{R}$ if and only if $f(X+Z) \le f(X) + +\nabla f(X)\cdot Z$ for all $X \in A$ and $Z \in \mathbb{R}^N$

Proof:

for given $\alpha \in (0,1)$

$$f(\alpha(X+Z) + (1-\alpha)X) \ge \alpha f(X+Z) + (1-\alpha)f(X)$$

$$f(X) + \frac{f(X+\alpha Z) - f(X)}{\alpha} \ge f(X+Z)$$

for $X, Z \in \mathbb{R}, \alpha \in (0,1)$

$f(X + \alpha Z) \approx f(X) + \nabla f(X) \cdot (\alpha Z) + \frac{\alpha^2}{2} Z \cdot D^2 f(X + \beta Z)^Z$

(Second-order Taylor Expansion)

So $\frac{\alpha^2}{2} Z \cdot D^2 f(X + \beta Z)^Z < 0$ when α and β small

Thus, strict concave $\rightarrow$ strict negative definite, vice versa.

So $L(\beta|Y, X)$ is strict concave function since $H(\beta|Y, X)$ strictly negative definite.

Thus, $L(\beta|Y, X)$ has unique maximum value (MLE). The logistic regression has unique MLE estimator ($\hat{\beta}$) for coefficient of X.

$I(\beta) = -(EH(\beta|Y, X)|X)^{-1}$ since $H(\beta|Y, X) = -\sum_i \frac{\exp(\beta^T X_i)}{(1+\exp(\beta^T X_i))^2} X_i {X_i}^T$

not depend on X (here considered as random value), $I(\beta) = -H(\beta|Y, X)^{-1}$

$\hat{\beta}$ is MLE, consistent, unbiased.

12.Conclusion

[Section 4: The structure of selected data]

A large part of the data distribution is skewed. All distribution of 11 of them are all discussed in the investigation part. More details can seen in that part with figures. This part is just an overview of the data structure.

[Section 5: Test of independence]

By test of independence, one can see that the variables—including price, user type, image width, image height, image dimension, image whiteness, image/no image, parent category, description word count and title word count—are not independent of the deal probability. So they have higher probability to become candidates of explanatory variables in the regression model. Since the data is skewed and the variables might have affect on each others, the test of independence is not enough to eliminate those variables for which the null hypothesis is not rejected from the regression model in which the study

is interested. Further elimination process according to p-value will be applied in a following section to eliminate less relavent variables.

[Section 6: Bootstrap]

Bootstrap is applied to further test the independence because the data from sample are mostly skewed and Bootstrap can capture the skewness. After applying bootstrap method on three varaibles—With/without image, Description word count and Title word count, it can be concluded from the test that all three variables have effect on the deal probability. Tests in this section all have the same result as the test of independence used in section 5, indicating the outcomes and conclusion in section 5 are to some extent accurate. Besides, As for the problem of the testing method in this part, investiagtions are conducted to solve the problem and it will be further discussed in "extra credit" section.

[Section 7: Deal probability distribution analysis]

The deal probability distribution does not follow exponenetial distribution or uniform distribution. The non- zero deal probability has a distribution similar to the combination of two normal distribution with different mean.

[Section 8: Logistic Regression]

After deleting all less relevant variable, the final model of the logistic regression is shown in Figure 12.8.1. Thus, the logistic regression model shows the most significant variables are price, with or without image, capital letter count for description, digit count for title, and is or is not company user. Among all of these, only intercept, price, and digit count for title have a positive influence in deal probability. All the others have negative influence. Among all the influences, the

```
Call:
glm(formula = mydata$with.deal ~ mydata$price + mydata$with.image +
    newdata$data.des.CapsCount + newdata$data.ti.DigitCount +
    mydata$is.company, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0929  -0.8179  -0.8026   1.1246   1.9171

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.593e-01  2.892e-01   0.897   0.3698
mydata$price               1.185e-07  6.123e-08   1.935   0.0530 .
mydata$with.image         -1.183e+00  2.908e-01  -4.067 4.77e-05 ***
newdata$data.des.CapsCount -4.408e-02  2.452e-02  -1.798   0.0721 .
newdata$data.ti.DigitCount  5.245e-01  8.319e-02   6.304 2.89e-10 ***
mydata$is.company          -3.442e-01  1.677e-01  -2.053   0.0401 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1218.0  on 933  degrees of freedom
Residual deviance: 1121.6  on 928  degrees of freedom
  (66 observations deleted due to missingness)
AIC: 1133.6

Number of Fisher Scoring iterations: 4
```

Figure 12.8.1 R outputs for final Logistic Model after deleting variables based on p values

The investigation further tests the accuracy of the final logistic model with the same data. In the prediction of the logistic regression given X, if the predicted probability that Y = 1 is larger than or equal to 0.5, the Y for corresponding X is predicted to be one. Otherwise, Y is predicted to be zero. In the test for accuracy, the investigation input all the data into the model and get only 108 out of 363 are correctly to be 1, which means the model is not ideal.

[Section 9: Numerical Regression]

After the logistic regression model, a linear model is conducted for non- zero deal probability. In this model, only digit count and word count for title will have an observable impact. More specifically, word count negative impact, and digit has a positive impact. However, the Model R-square is to some extent small so it is not an ideal model.

The distributions of single variable give the investigation a blue print of further testing and also show some problems. The test of independence and bootstrap give some reasonable statistically significant variables that may influence the targeted deal probability. Although the regression part does not give an ideal and convincing model, the ads posters for Avito can still learn some thing.

13.Reference

[1] *"Avito."* https://en.wikipedia.org/wiki/Avito.ru

[2] https://www.kaggle.com/c/avito-demand-prediction

[3] https://www.kaggle.com/gunnvant/russian-to-english-translate-with-progress-bar

[4] https://www.kaggle.com/shivamb/ideas-for-image-features-and-image-quality

[5] Larsen, Richard J., and Morris L. Marx. An Introduction to Mathematical Statistics and Its Applications. Pearson, 2017.

[6]*"The History of E-Commerce, Online Shopping Evaluation, and Buyers Behavior."* https://www.altushost.com/the-history-of-e-commerce-online-shopping-evolution-and-buyers-behaviour/

s