Customer Review Amazon Products

1. Introduction

Customer feedback is increasingly accessible on the internet, covering a wide variety of products and services. These reviews complement the information provided by online stores, such as product descriptions, expert reviews, and personalized recommendations from automated systems. While researchers have highlighted the advantages of having customer reviews for online retailers, an aspect that has received limited exploration is what factors contribute to the usefulness of customer reviews. Such information is not only helpful for consumers in making a decision of a purchase, but also for online retailers to improve their products. Regarding the highly potential growth of reviewing products, we have determined to perform the analysis to address three questions:

- Can we build a classification model to distinguish negative and positive reviews?
- What are the characteristics of positive and negative reviews?
- Is there any correlation between the types of reviews and its helpfulness? (i.e. If a review is classified as positive, is it considered a helpful review?). What are the influencers of a helpful review?

We will use both parametric and non-parametric techniques for these goals. Our efforts can provide companies with valuable insights to enhance product features, boost sales, and assist customers in making informed buying decisions.

2. Literature Review

With the emergence of e-commerce platforms and a shift in customers preference towards online shopping, the online consumer reviews (OCR) system has become the most informative tool to help customers make wise decisions and improve companies' product development. A study by Liu et al. (2021), which delved into the topic of satisfaction measurements on Amazon online customer reviews, suggests that reviews length, the extremeness of reviews, and whether or not the reviewers are Vine members, are factors exerting significant influence on reviews helpfulness. This information will serve as a crucial foundation for our project, as we aim to analyze these factors in the context of Amazon product reviews. The result was derived under the implementation of a supervised machine learning technique (i.e. information gain model), hypothesis testing and application of OLS regression, followed by an establishment of customer satisfaction measurement using EMA to reflect the trend of satisfaction level over time. Such a framework can be applied to our own research to analyze and evaluate Amazon product reviews systematically. Another study by Baek et al. (2012) using sentiment analysis on data gathered from Amazon via web scraping has found that the content of review and review star rating are not always congruent. It will help our project by emphasizing the importance of looking beyond star ratings and

delving deeper into the actual content of reviews to gauge customer sentiments accurately. Besides, reviews are deemed to be most helpful when it is parallel with the majority average rating, or when they are more lengthy, with frequent use of negative words, and are written by high-ranked accounts. These factors can be incorporated into our analysis to better understand how review content influences consumer perception and decision-making.

3. Data, Data Sources and Characteristics

3.1. Data Gathering

This project employs a subset of the original dataset named "Customer Review Amazon Products" for modeling and prediction analysis. This random sample consists of 25000 observations and 22 variables. Each row provides the information related to a product on Amazon website, while the variables are features associated with each product. Appendix 1 is the data dictionary of all variables we considered in our data analysis (including original variables and newly created ones).

3.2. Data Cleaning and Preprocessing

- **Data summary**

To get an insight into the data content, we use function skim() in R to access summary statistics of each variable and spot potential missing values or outliers. The result (appendix 2) suggests that there is complete missing data of all observations in variables named "reviews.didPurchase", "reviews.id", "reviews.userCity", "reviews.userProvince". This means that these variables have no value at all. Besides, there are 19.5% of missing values in "name", 1.6% in "reviews.doRecommend", 1.5% in "reviews.numHelpful", and trivial amount in columns such as "asins", "reviews.date", "reviews.rating", "reviews.text", "reviews.title" and "reviews.username". Besides, variable "reviews.numHelpful" possesses a wide disparity between the minimum value (0) and the maximum value (814) with a mean and median of zero, indicating potential outliers that need further processing.

- **Data Visualization**

To understand the distribution of variables, we created a set of distribution plots for both numeric and categorical variables. First, we created a histogram on "reviews.numHelpful", which is shown to be right skewed (appendix 3.1 & 3.2). For categorical variables, frequency tables and bar charts on the proportion of each category are created. This step provides an insight into where we are having an imbalance dataset that could impact outcome. Appendix 3.3 indicates that class "helpful" is significantly less than "non-helpful". Given that our "helpful" is the important class we aim to predict, being a minority class

(10% out of all observations) requires oversampling to ensure it has equivalent probability in the prediction model. Imbalance among categories is also present in other variables such as “reviews.doRecommend”, “brand”, “manufacturer”, and “rating” (appendix 3.4, appendix 3.5, appendix 3.6, appendix 3.7). Next, a box plot (appendix 3.8) describing the relationship between the number of helpful reviews for each rating level is conducted. The box plot indicates that the higher the rating, the greater the dispersion of the number of helpful reviews, indicating greater number of outliers. The median of 3-star, 4-star, and 5-star review rating is close to the bottom of the box, meaning that products that have above 3 stars rating generally have little to no helpful reviews. The larger number of outliers specifically present in the 5-star rating products as compared to others suggests that we need to consider these variables carefully later as these might be important records. Lastly, the contingency table (appendix 3.8) excerts that 92% of the reviews that are perceived “helpful” also have the corresponding product recommended.

- **Treating missing values**

In this case, we focus on treating the missing values of variables that we will use in the analysis. Thus, “reviews.numHelpful”, “reviews.doRecommend”, “reviews.rating”, “categories”, “reviews.date”, “reviews.numdates” and “reviews.text” are considered the most important ones as they will be transformed and used to create new variables.

- “reviews.numHelpful” will be used to create a binary dependent variable, thus replacing their missing values by mean or median will create bias in the result. On top of that, the amount of missing values is very small, thus, we can remove all missing values without affecting results.
- For categorical variables like “reviews.doRecommend”, we will replace missing values with mode, which is considered appropriate as it represents the majority of our data. After removing all missing values, its proportion is 9.7% for “helpful” and 90.23% for “non-helpful” (appendix 3.9).

- **Adding new variables**

- Text.sentiment: To capture sentiment scores derived from “reviews.text”, using vader_df() function from the varder package. These scores represent the emotional tone (positive, negative, or neutral) of each review.
- Review Type: This variable is created from “text.sentiment” to see whether the review is positive or negative.
- isHelpful: To achieve the classification task between helpful and non-helpful reviews, we have created a new variable “isHelpful” from “reviews.numHelpful” that takes value 0 if the review received zero helpful vote, and takes value 1 if the review received at least 1 helpful vote. The

proportion of “helpful” and “non-helpful” in “reviews.Helpful” in the dataset including all missing values are 9.76% and 90.23% respectively.

- Reviews.length: This variable is created from “reviews.text” to get more valuable insights into the depth of feedback, product understanding and how such factors relate to the reviews rating and perceived helpfulness.
- Day_of_week: This variable is extracted from “reviews.date” to see the day the review was posted.
- Reviews.numdates: We created these variables from “reviews.dateSeen” to get information about how many times the reviews were seen.

- **Dimension Reduction**

We exclude “reviews.didPurchase”, “reviews.id”, “reviews.userCity” and “reviews.userProvince” as 100% of them are missing values. Besides, we remove “x”, “id”, “asins” and “key” because they do not have potential contribution to the questions we aim to address. For grouping variables, there are 41 dummy variables in categories. We grouped these dummy variables into 6 groups: “Household Electronics”, “General Computer & Tablets”, “College Electronics”, “Office Electronics”, “E-readers”, and “Electronics Accessories”.

After being cleaned, the dataset is left with 9 variables and 24,563 observations.

4. Data Mining Technique

4.1. Define Tasks

Base on the business questions that we want to solve, we choose appropriate data mining tasks :

- First task: classifying negative and positive reviews and finding the best predictive model.
- Second task: Identifying the characteristics of positive and negative review groups.
- Third task: classifying significant factors associated with a helpful review. Identifying the relationship between the types of review and its helpfulness.

4.2. Technique Selection

For the first task, our target variable is “reviews type” as a category variable, so we use Logistic Regression and non-parametric techniques (Naive Bayes, Classification Tree, Random Forest) to refine the chosen model. In the second task, we employed Hierarchical Clustering for review segmentation. And in the third task, Logistic Regression is the best fit to classify response to reviews of products as our dependent variable (isHelpful) is categorical.

4.3. Treating imbalance in dataset (oversampling) & data partition

In fact, the records are imbalances, particularly there are 9.6% reviewers finding this review is helpful, so we determine to oversampling data. The training data includes 50% of reviewers finding this review being helpful and 50% of reviewers finding this review being non-helpful to ensure equal contribution of both groups on the predictive power of the model. The proportion of "helpful" and "non-helpful" in the validation data is similarly the same as original data with 9.6% of helpful and 90.4% of non-helpful (appendix 4.1). After partitioning, we have a training dataset with 2362 observations and 9 variables, and a validation dataset with 12,302 observations and 9 variables.

Similarly, we check the proportion of "positive review" and "negative review" in the data. "Positive review" accounts for 90.92% (appendix 4.2), we need to oversampling the train dataset for a better prediction. By increasing the instances of the minority class, we achieved a more equitable distribution (appendix 4.3), enhancing the model's ability to learn from both classes effectively.

4.4. Apply data mining technique

Regarding the first task of distinguishing between negative and positive reviews, four models—Logistic Regression, Naive Bayes, Classification Tree, and Random Forest—were employed and evaluated based on their performance metrics. The first technique (Logistic Regression) is parametric, which helps with choosing the best model that includes the most significant predictors by evaluating estimated coefficients. The rest are non-parametric, which are more efficient in prediction power and have greater capability of fitting large data.

From the application of Logistic Regression and Stepwise Selection method, we have found the best model for classifying review types. The model is written as:

> *P(Review Type | Product Categories, Reviews Rating, Time Seen, Review Length, Day of Week, doRecommend, isHelpful).*

The dependent variable in this case is Review Type and the selected independent variables are Product Categories, Reviews Rating, Time Seen, Review Length, Day of Week, doRecommend and isHelpful. Next, such a model is employed and tested on the validation set by all techniques. Among these, Random Forest emerged as the most suitable choice based on its higher accuracy of 74.85% and specificity of 78.85% (appendix 4.4). The Random Forest algorithm demonstrates its efficacy in accurately categorizing reviews, leading us to select it as our model of choice for this classification task. Based on the result of Random Forest (appendix 4.5), we have "day_of_week" and "product category" as important predictors.

For the second task of segmenting positive and negative reviews based on their optimism level, Hierarchical Clustering was the chosen method as it helps with exploring the granularity of each segment. First, we converted categorical variables into dummies, and normalized continuous variables to ensure equal contribution of all features. Next, we computed the Euclidean distance between groups and used the Elbow Curve plot to determine the optimal number of 4 clusters (appendix 4.6). Thereafter, these clusters were evaluated through the centroid values (appendix 4.7). Based on the result, we found that the positive reviews are mostly reviews from Household Electronics Product, are longer reviews, and have higher number of views by other online customers. On the other hand, negative reviews are mostly reviews from College Electronics Products, have moderate length and lower number of views by other online customers. Other features such as Day of Week and Rating do not exert distinctive characteristics between the two groups, which can be explained by the inherently dominant of one class over the others in these variables.

Lastly, to identify significant factors of a helpful review, we employed logistic regressions and used the estimated coefficients and p-values to answer the corresponding research question. The chosen model is written as:

P(isHelpful | Product Categories, Time Seen, Review Length, Day of Week, doRecommend).

The dependent variable is isHelpful and the independent variables are Product Categories, Time Seen, Review Length, Day of Week and doRecommend. These independent variables are all significant factors in determining whether a review is helpful or not. Our findings revealed no correlation between the sentiment—whether positive or negative—of a review and its perceived helpfulness among users. Instead, factors such as review length, recommendation status, and the product category wield significant influence over how users perceive and rate the helpfulness of reviews (with p-values of <2e-16, 1.00e-05 and 2.55e-11 respectively) and reviews rating is the least important factor (appendix 4.8). These insights emphasize the necessity of considering multiple dimensions beyond sentiment analysis when assessing the usefulness of reviews in guiding consumer decisions.

From the decile chart (appendix 4.10), our model can correctly classify a 2.9 times higher number of helpful reviews in the first 10% of reviews as compared to random selection. The second bar chart shows that the second 10% is giving us close to a 1.5 times in terms of “helpful” compared to random selection. We can see the numbers more clearly in the cumulative lift chart (appendix 4.10). The chart shows that, using the classification model, the top 10% reviews (1200 cases) with the highest propensity ranking by the model contain 31% of the total helpful reviews, as compared to only 9% of the total helpful reviews if

we pick 1200 cases randomly. Adding 10% more cases (2400 cases), we can capture 35% of the total helpful reviews using the model, as compared to 17% by random guessing. Overall, the model performs 2 times better as compared to the random baseline.

5. Empirical results

In this paper, we have tried different models on reviews types classification, assessed the factors affecting the helpfulness of reviews and explored the relationship between reviews type and the perceived helpfulness. We have been able to build an effective classification model for review types, and Random Forest is the best algorithm for serving such a task. The same attempt was conducted for review helpfulness, in which we wanted to classify helpful versus unhelpful reviews. Different methods from logistic regression to Naive Bayes, Classification Tree, K-NN, and Random Forest was employed, yet none of them successfully produced a good classification model as the sensitivity (% of unhelpful class correctly predicted) was always significantly higher than specificity (% of helpful class correctly predicted). One explanation could be due to the lack of information on the total vote of each review. Particularly, there is only information regarding the number of helpful votes in the original dataset, therefore, we created the variable “isHelpful” based on the logic that a review is considered helpful if it has at least 1 helpful vote. This conversion can create bias as it assigns equal weight to reviews that received one helpful vote and reviews that received many helpful votes, when in fact the weight should be different based on the proportion of helpful vote to total vote. This might to some extent affect our results in finding the relationship between review type and review helpfulness. As such, it is suggested that more information regarding overall vote should be added to the dataset for a better prediction. Besides, the product categories in the dataset are dominated by electronics of all kinds, which may reduce the model's ability to classify reviews type of other categories that are not concerned. For future analysis on the same topic, merging data from other sources is highly recommended for deeper and more valuable insights to be used for e-commerce businesses.

6. Conclusion and Recommendation

Our analysis suggests that the combination of logistic regression (parametric) and Random Forest (non-parametric) is highly effective and enables high accuracy for review type classification. Besides, the segmentation of positive and negative reviews based on characteristics such as product categories, review length, and day of the week when the review is posted provides actionable insights. From the odds values (appendix 4.9), positive reviews are mostly posted on Saturday and Sunday, and more likely to be reviews

from E-readers products. Marketers can utilize such information to create appropriate strategies to increase positive reviews, such as launching promotions and marketing campaigns targeted for weekends (i.e. creating customer engagement initiatives in the weekend through Q&A sessions, webinars, and social media lives events where customers can learn more about E-reader and asking question. Positive interactions during these activities can contribute to more positive reviews).

Besides, Amazon can implement a review voting system tied to a contributor badge or donation program that presents an opportunity to incentivize and engage voters. Offering exclusive benefits at different contribution levels, even with just one vote, creates an inclusive space where each vote contributes to charitable causes.

Positive reviews serve as an importance for enhancing business prospects. By actively promoting products associated with positive feedback, businesses can boost sales and increase customer satisfaction. Expressing gratitude to customers who provide positive feedback through communications via email or social media makes stronger customer relationships. Additionally, enhancing visibility by featuring most-rated positive reviews to attract potential customers. Meanwhile, negative reviews are pivotal for improvement. Encouraging dissatisfied customers to reconsider their opinions by offering exclusive discounts or extended warranties demonstrates a commitment to rectification and customer-centricity. Continuously monitor negative reviews for emerging or recurring issues & possibly spot out spam reviews. Further, enhancing customer support channels, such as introducing live chat options for immediate issue resolution, becomes instrumental in addressing and resolving common concerns swiftly.

References

Liu, Y., Wan, Y., Shen, X., Ye, Z., & Wen, J. (2021). Product customer satisfaction measurement based on multiple online consumer review features. *Information*, *12*(6), 234. https://doi.org/10.3390/info12060234

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of Online Consumer Reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, *17*(2), 99–126. https://doi.org/10.2753/jec1086-4415170204

Appendix

Appendix 1. Data Dictionary using for analysis

Variable	Data Type	Description
isHelpful	Binary	Whether or not the review is helpful
Rating	Categorical	A 1 to 5 star value for the review
Timeseen	Numerical	The number of time the review was seen
Day of week	Categorical	The day of week the review was posted
Length	Numerical	The length of the full text of review
Product Categories	Categorical	A list of product category
DoRecommend	Categorical	Whether or not the reviewer recommends the product
Text Sentiment	Numerical	Score of a review (negative score = negative reviews, zero score = neutral reviews, positive score = positive reviews)
Review Type	Categorical	Whether the review is positive or negative

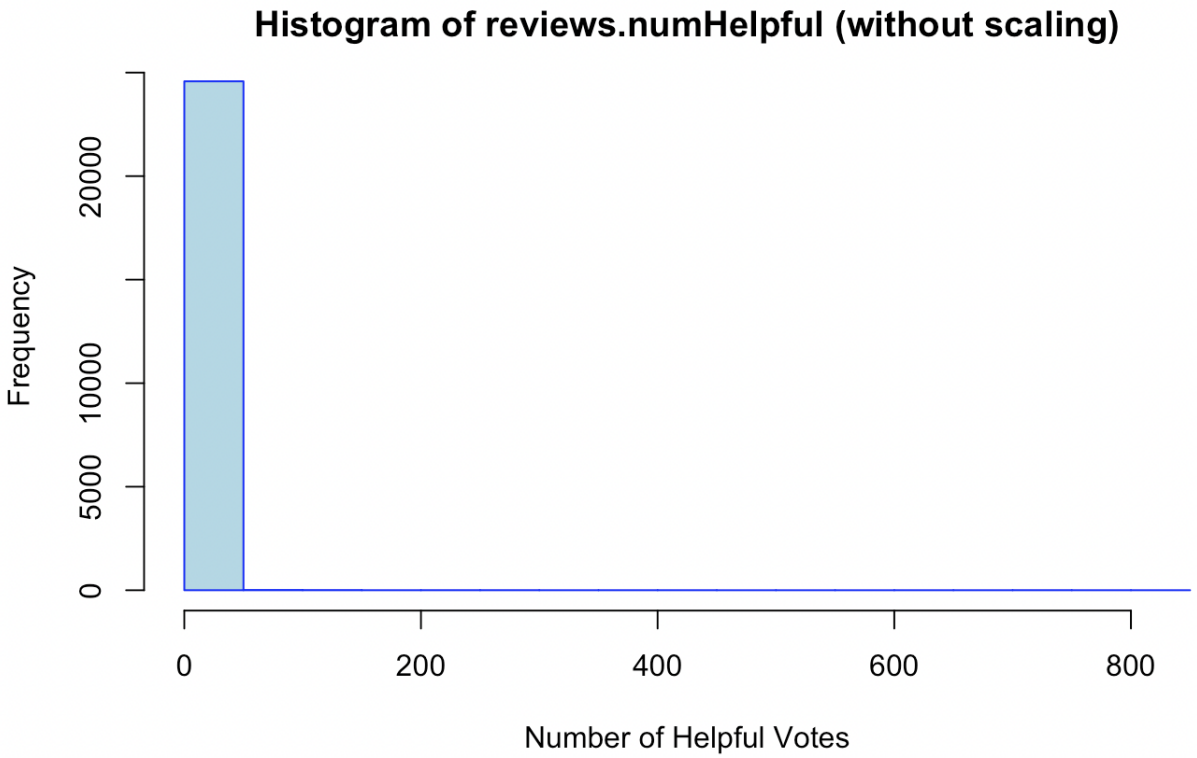
Appendix 2. Missing Value

Number of missing values

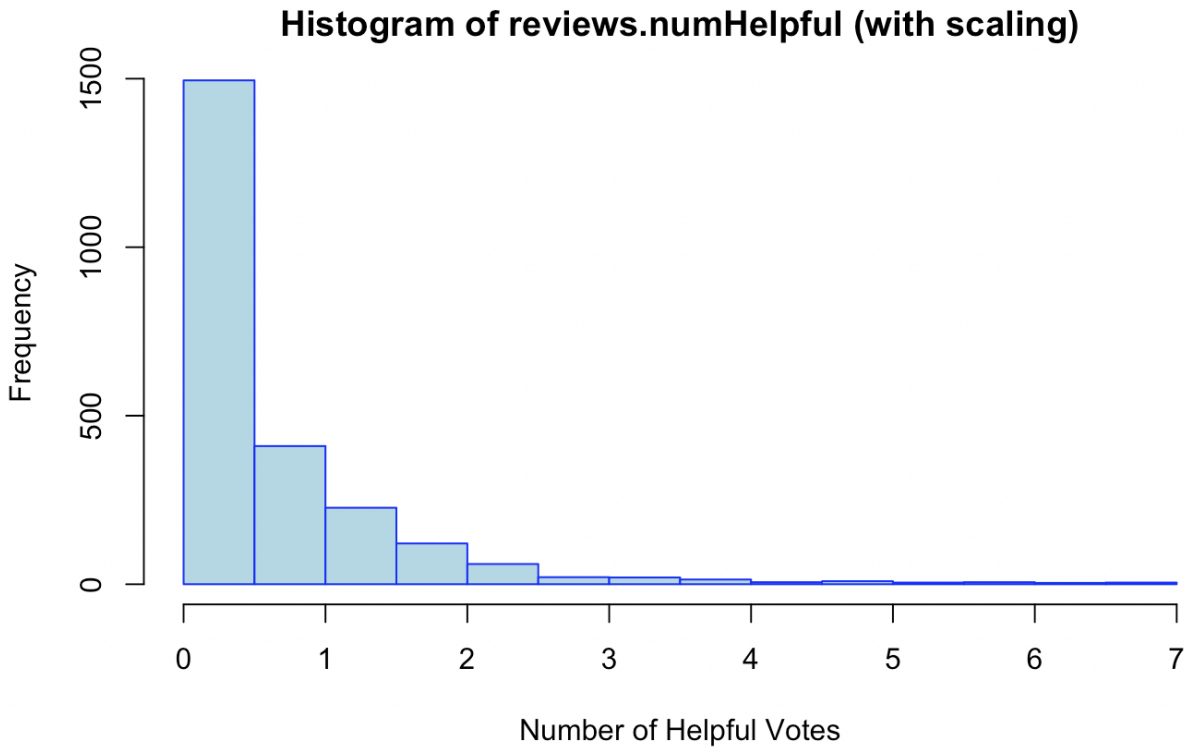
	miss.val
X	0
id	0
name	4875
asins	2
brand	0
categories	0
keys	0
manufacturer	0
reviews.date	31
reviews.dateAdded	7633
reviews.dateSeen	0
reviews.didPurchase	25000
reviews.doRecommend	424
reviews.id	25000
reviews.numHelpful	383
reviews.rating	24
reviews.sourceURLs	0
reviews.text	1
reviews.title	3
reviews.userCity	25000
reviews.userProvince	25000
reviews.username	2

Appendix 3. Data Exploration

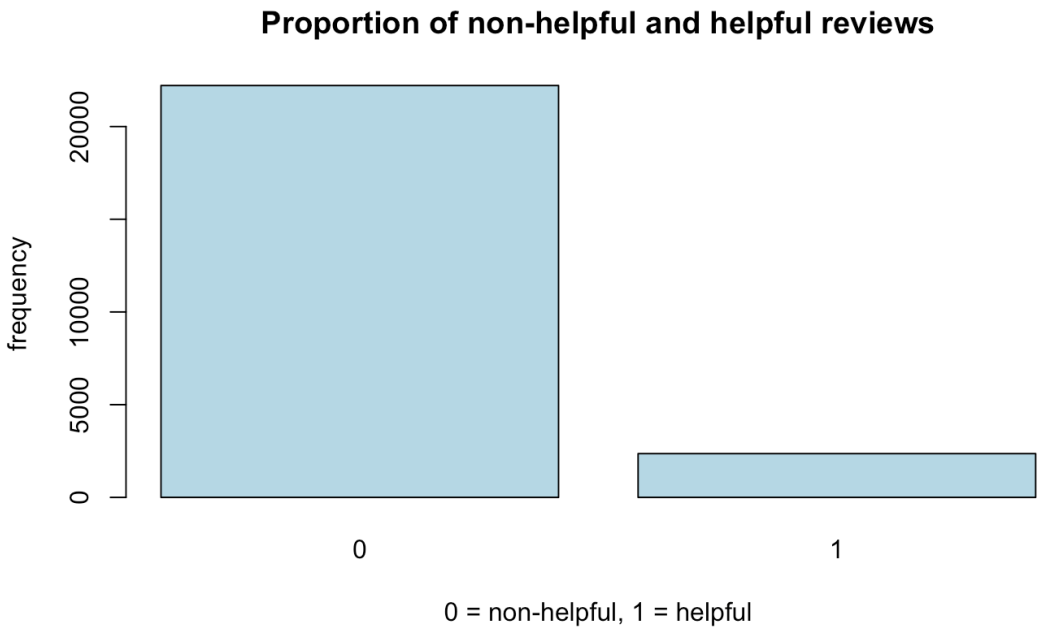
Appendix 3.1. Histogram of the number of helpful votes



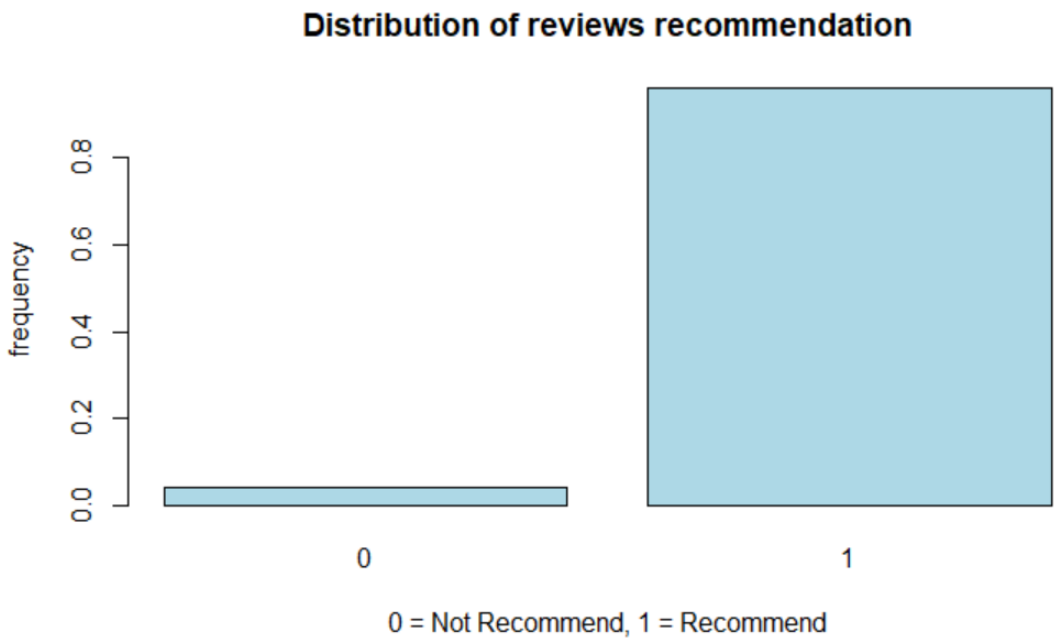
Appendix 3.2. Histogram of the number of helpful votes (with scaling)



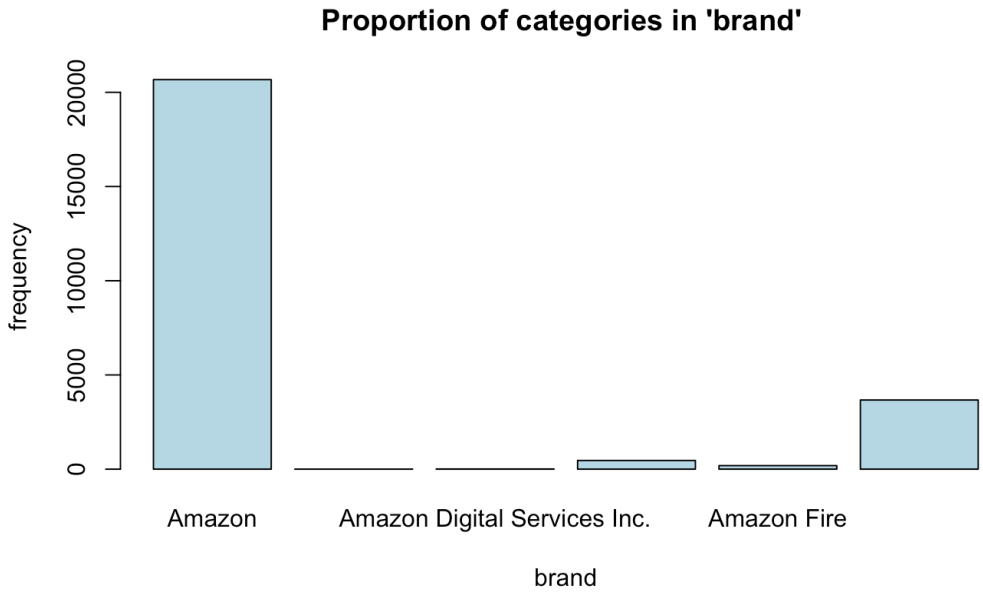
Appendix 3.3. The distribution of the number of helpful votes



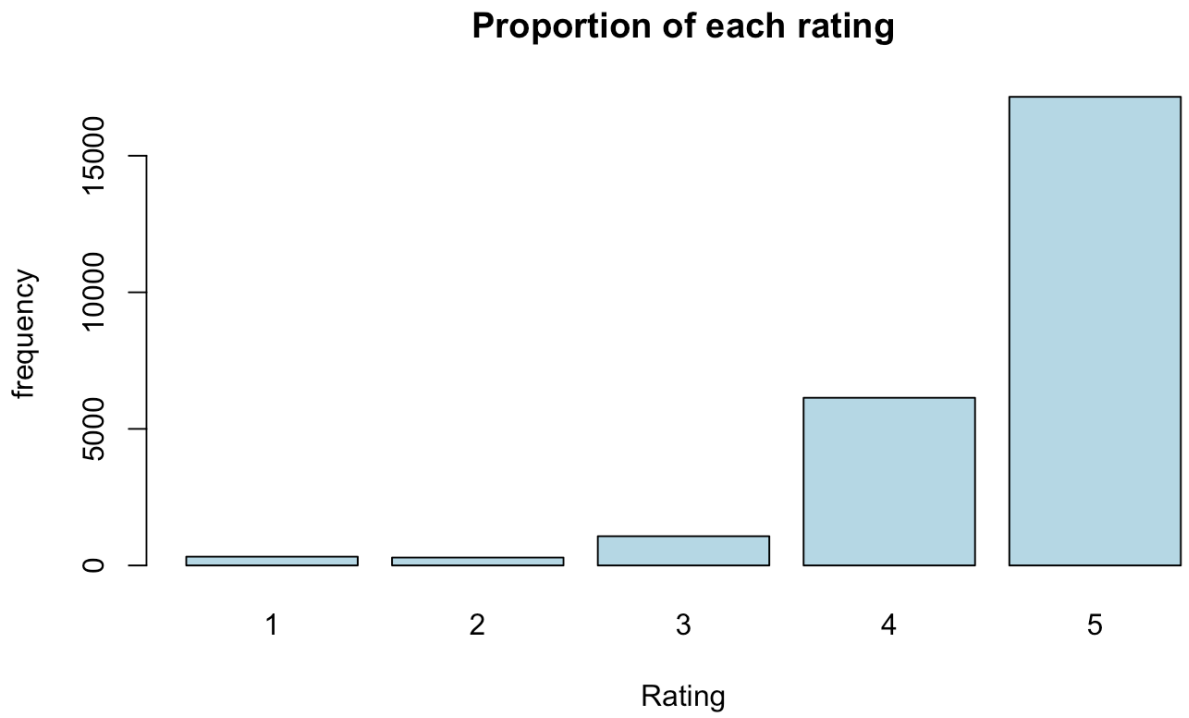
Appendix 3.4. The distribution of the reviews recommendation



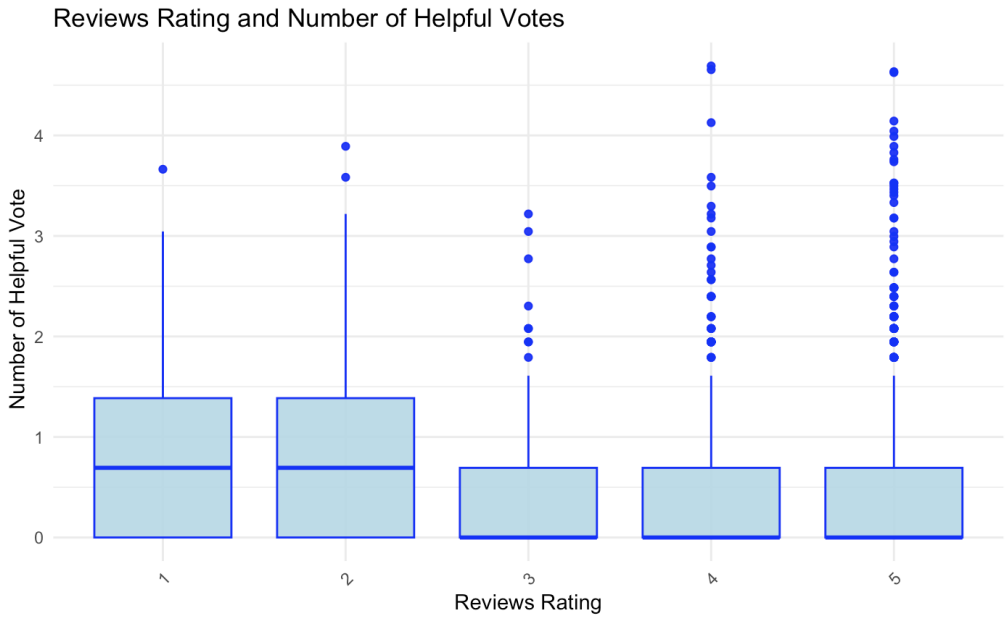
Appendix 3.5. The distribution of brand



Appendix 3.6. The distribution of rating



Appendix 3.7. Box plot between Rating and The number of helpful votes



Appendix 3.8. Contingency table

df$reviews.doRecommend	df$reviews.helpfullness 0	1	Total
FALSE	824	197	1021
	0.807	0.193	0.042
TRUE	21389	2164	23553
	0.908	0.092	0.958
Total	22213	2361	24574

Appendix 3.9. The proportion of "helpful" and "non-helpful" without missing

Table: The proportion of helpful and non-helpful reviews without missing values

reviews.numHelpful	count	percent
0	22214	0.9023845
1	2403	0.0976155

Data Mining Techniques

Appendix 4.1. The proportion of "isHelpful" before and after oversampling

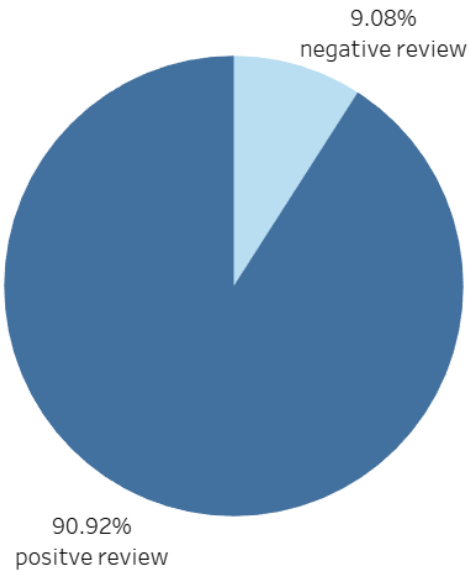
Table: Checking training set after oversampling

isHelpful	count
0	1194
1	1194

Table: Checking validation after oversampling

isHelpful	count	percent
0	11088	0.9027846
1	1194	0.0972154

Appendix 4.2. The proportion of "positive review" & "negative review"



Appendix 4.3. The proportion of "positive review" & "negative review" after oversampling

```
over.rt <- ovun.sample(reviews.type ~ ., data = train.df2, method = "over", p = 0.5)$data

# Checking the number of observations in each class in the oversampled data
table(over.rt$reviews.type)
```

```
##
##     1     0
## 13389 13278
```

Appendix 4.4. Techniques Evaluation for "review type"

Evaluation	**Logistic**	**Naive Bayes**	**Classification**	**Random Forest**

	Regression		**Tree**	
Accuracy	0.092	0.7029	0.6273	0.75
Specificity	0.0231	0.72248	0.6303	0.7902

Appendix 4.4.1. Confusion Matrix of Logistic Regression

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  697 8740
##          1  184  207
##
##                Accuracy : 0.092
##                  95% CI : (0.0863, 0.0979)
##     No Information Rate : 0.9104
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0345
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.79115
##             Specificity : 0.02314
##          Pos Pred Value : 0.07386
##          Neg Pred Value : 0.52941
##              Prevalence : 0.08964
##          Detection Rate : 0.07092
##    Detection Prevalence : 0.96022
##       Balanced Accuracy : 0.40714
##
##        'Positive' Class : 0
##
```

Appendix 4.4.2. Confusion Matrix of Naive Bayes

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  444 2483
##          1  437 6464
##
##                Accuracy : 0.7029
##                  95% CI : (0.6937, 0.7119)
##     No Information Rate : 0.9104
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1106
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.50397
##             Specificity : 0.72248
##          Pos Pred Value : 0.15169
##          Neg Pred Value : 0.93668
##              Prevalence : 0.08964
##          Detection Rate : 0.04518
##    Detection Prevalence : 0.29782
##       Balanced Accuracy : 0.61322
##
##        'Positive' Class : 0
##
```

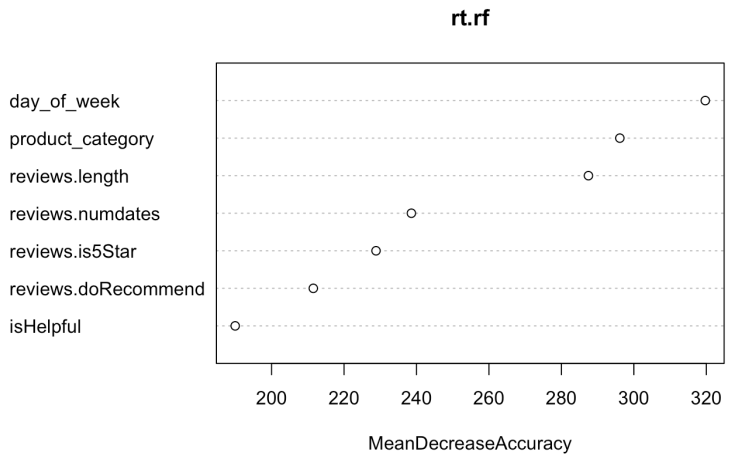
Appendix 4.4.3. Confusion Matrix of Classification Tree

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  524 3306
##          1  357 5641
##
##                Accuracy : 0.6273
##                  95% CI : (0.6176, 0.6369)
##     No Information Rate : 0.9104
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0898
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.59478
##             Specificity : 0.63049
##          Pos Pred Value : 0.13681
##          Neg Pred Value : 0.94048
##              Prevalence : 0.08964
##          Detection Rate : 0.05332
##    Detection Prevalence : 0.38970
##       Balanced Accuracy : 0.61263
##
##        'Positive' Class : 0
##
```

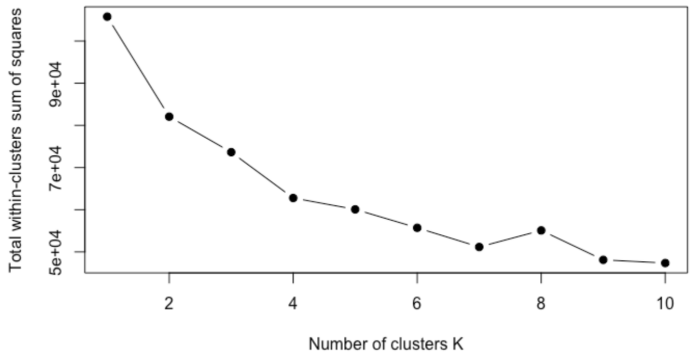
Appendix 4.4.4. Confusion Matrix of Random Tree

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  301 1877
##          1  580 7070
##
##                Accuracy : 0.75
##                  95% CI : (0.7413, 0.7585)
##     No Information Rate : 0.9104
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0793
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.34166
##             Specificity : 0.79021
##          Pos Pred Value : 0.13820
##          Neg Pred Value : 0.92418
##              Prevalence : 0.08964
##          Detection Rate : 0.03063
##    Detection Prevalence : 0.22161
##       Balanced Accuracy : 0.56593
##
##        'Positive' Class : 0
##
```

Appendix 4.5. Plot of important predictors of chosen model



Appendix 4.6. Elbow method for finding the optimal number of clusters



Appendix 4.7. Centroids value of cluster analysis

	reviews.doRecommend_NO	reviews.doRecommend_YES	reviews.is5Star_NO	reviews.is5Star_YES	day_of_week_Friday	day_of_week_Monday
1	0.039648367	0.9603516	0.2850736	0.71492644	0.3616792	0.1327592
2	0.009630267	0.9903697	0.1112640	0.88873603	0.3836629	0.1098882
3	0.002130209	0.9978698	0.0000000	1.00000000	0.3368393	0.1561709
4	0.128263938	0.8717361	0.9652435	0.03475653	0.3189838	0.1554340

	day_of_week_Saturday	day_of_week_Sunday	day_of_week_Thursday	day_of_week_Tuesday	day_of_week_Wednesday
1	0.1824543	0.06512379	0.06117689	0.06404736	0.1327592
2	0.1840069	0.07102322	0.05967326	0.05434222	0.1374033
3	0.1673545	0.06177606	0.06590334	0.06856610	0.1433897
4	0.1725476	0.07233592	0.06563162	0.07374735	0.1413197

	product_category_College	Electronics	product_category_E-readers	product_category_Electronics	Accessories
1	0.0204520990		0.023681378	0.0053821313	
2	0.6044711952		0.011177988	0.0018916595	
3	0.0001331381		0.003062175	0.0000000000	
4	0.0388143966		0.007939308	0.0008821454	

	product_category_General	Computer and Tablets	product_category_Household	Electronics	product_category_Office	Electronics	isHelpful_NO
1	0.123071403		0.75905992	0.068353068		0.9284177	
2	0.005503009		0.07205503	0.304901118		0.9349957	
3	0.943283185		0.04366929	0.009852217		0.8945547	
4	0.830275229		0.03934368	0.082745236		0.8551517	

	isHelpful_YES	reviews.type_Negative	reviews.type_Positive	reviews.length	reviews.numdates
1	0.07158235	0.07517043	0.9248296	1.0328476	3.3853945
2	0.06500430	0.09492691	0.9050731	0.8360164	1.7379487
3	0.10544535	0.05751564	0.9424844	0.7412410	0.9511946
4	0.14484827	0.14661256	0.8533874	1.2479634	1.0559829

	Cluster	Optimistic Level	Recommend or Not?	Day of Week	% of Rating Below 5	Helpful Review or Not?	Product Category	Review Length	Time Seen
Positive	2	High	Recommend	Friday	Low	Non-helpful	Household Electronics	Moderate	Highest View
	4	Medium-High	Recommend	Friday	High	Non-helpful	General Computer and	Long	Moderate

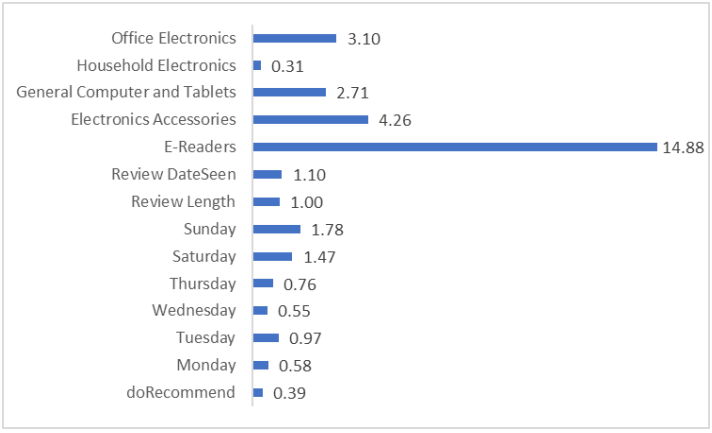
							Tablets		
Negative	1	Medium-Low	Recommend	Friday	Low	Non-helpful	General Computer and Tablets	Moderate	Lowest View
	3	Low	Recommend	Friday	Low	Non-helpful	College Electronics	Moderate	Moderate

Appendix 4.8. The results of logistic regression model using STEPWISE elimination

```
##
## Call:
## glm(formula = isHelpful ~ reviews.doRecommend + day_of_week +
##     reviews.length + reviews.numdates + product_category, family = binomial(link = "logit"),
##     data = train.df)
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                             -0.683934   0.285886  -2.392
## reviews.doRecommend1                    -0.930642   0.210710  -4.417
## day_of_weekMonday                       -0.538561   0.146575  -3.674
## day_of_weekSaturday                      0.385593   0.126334   3.052
## day_of_weekSunday                        0.574231   0.194210   2.957
## day_of_weekThursday                     -0.279827   0.193993  -1.442
## day_of_weekTuesday                      -0.027632   0.178918  -0.154
## day_of_weekWednesday                    -0.595099   0.156582  -3.801
## reviews.length                           0.003504   0.000327  10.717
## reviews.numdates                         0.093703   0.035921   2.609
## product_categoryE-readers                2.700121   0.506524   5.331
## product_categoryElectronics Accessories  1.448425   1.169122   1.239
## product_categoryGeneral Computer and Tablets  0.997677   0.149566   6.670
## product_categoryHousehold Electronics   -1.157006   0.217607  -5.317
## product_categoryOffice Electronics       1.132631   0.176147   6.430
```

```
##                                          Pr(>|z|)
## (Intercept)                              0.016742 *
## reviews.doRecommend1                     1.00e-05 ***
## day_of_weekMonday                        0.000238 ***
## day_of_weekSaturday                      0.002272 **
## day_of_weekSunday                        0.003109 **
## day_of_weekThursday                      0.149174
## day_of_weekTuesday                       0.877262
## day_of_weekWednesday                     0.000144 ***
## reviews.length                            < 2e-16 ***
## reviews.numdates                         0.009092 **
## product_categoryE-readers                9.78e-08 ***
## product_categoryElectronics Accessories  0.215383
## product_categoryGeneral Computer and Tablets 2.55e-11 ***
## product_categoryHousehold Electronics    1.06e-07 ***
## product_categoryOffice Electronics       1.28e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3310.5  on 2387  degrees of freedom
## Residual deviance: 2841.2  on 2373  degrees of freedom
## AIC: 2871.2
##
## Number of Fisher Scoring iterations: 5
```

Appendix 4.9. Log Odds of Feature Levels



Appendix 4.10. Lift Chart & Decile-Wise Lift Chart

Figure: Lift chart of helpfulness reviews' propensity

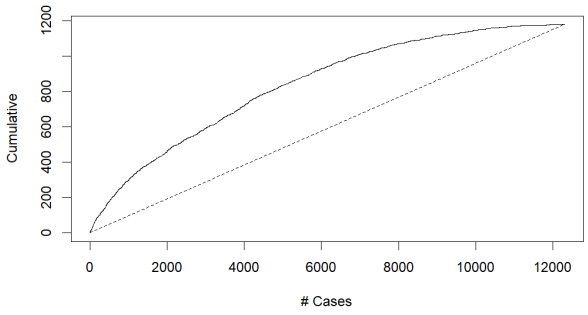


Figure: Decile-wise lift chart