Predicting E-Commerce Reviews for Women's Clothing

Prepared by Rose Connolly May 2021

STU3002 Statistical Analysis III, Trinity College Dublin

Business success often depends on previous customer experience. Online reviews and ratings can determine the success of a purchase. Analysing reviews to identify customer attitude could be used to save costs. E.g. a business could use AI to identify and request additional feedback from dissatisfied reviewer or concentrate marketing towards positive reviewers.

This report describes the use of regression to predict whether a customer will recommend an item. Sentiment analysis - "Emotion AI" - will be used to analyse reviews. Potentially combined with other factors such as buyer age, a regression model will be used as a prediction tool.

Data available at: https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

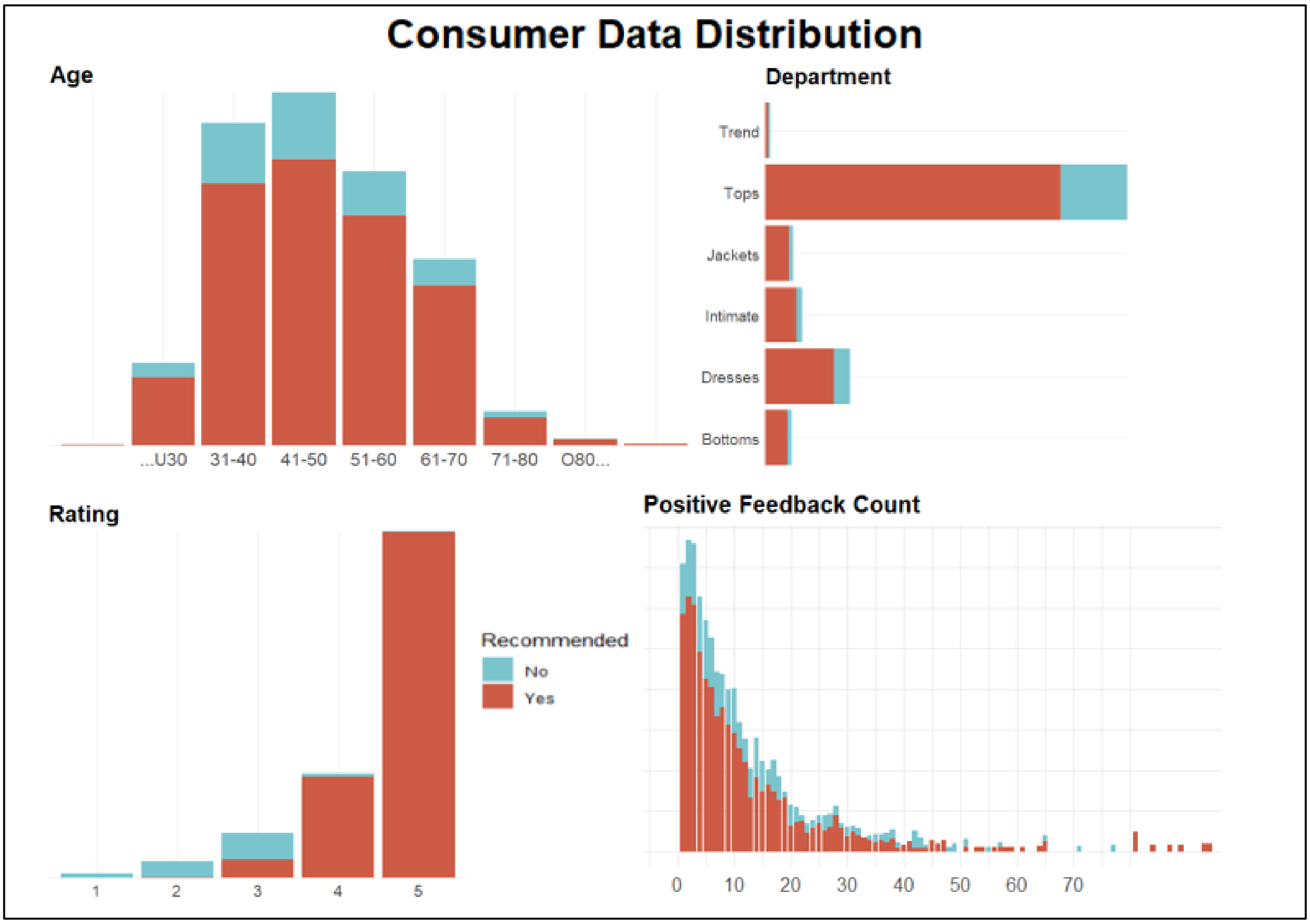
Introduction

Existing Literature

Sentiment analysis was chosen for this dataset as is commonly used for online web reviews in multiple studies.[1] [2] [3]The sentiment analysis lexicon used for this analysis was originally created from customer reviews and is generally used in 'opinion-mining' of reviews and social media (tweets, blogs etc.). The most common machine learning algorithm used with sentiment analysis is Naïve Bayes. Some common problems are the lack of sentiment analysis resources available for less widely used languages[4]. Sentiment analysis is commonly used in linear regression for uses such as predicting stock price or from opinions on social media[5] Therefore, using sentiment analysis in this regression is well justified and this research was consulted for guidance.

Dataset

The E-Commerce Reviews for Women Clothing dataset contains 23487 observations of 10 variables. The dataset records buyer, review, and clothing recommendation details. Variable data spread is roughly visualised below.



82% of buyers recommended the item they had purchased. Buyer age mimics a normal distribution with mean age 43. Most purchases came from the Tops department, and least from the Trend department. Most items were rated 5/5. Items generally had less than 10 existing positive reviews and the reviews follow an exponential distribution.

Methods

Data Cleaning

Columns Removed

- *Clothing ID*
- *Division Name (e.g. general):* Department (e.g. dresses) is a more insightful subset.

Missing Data

- The original data contained heading and spacer rows which were removed as they served aesthetic purposes only.
- The original data contained 'messy' rows where data included the wrong variable (e.g. age contained review text) and other missing variables. Filling variables with 'dummy data' (produced as average of similar rows) was considered. However, most messy rows had multiple missing variables so similar rows couldn't be identified. The missing variables tended to be essential to analysis (recommendation or buyer age) so estimating would be risky. 0.57% of data fell into this category. Removing them was not significant.
- Some review titles and reviews were missing, and these rows were removed. Estimating text is difficult and as they are essential to the analysis estimating would be unwise. 17% (quite a significant percentage) of the remaining data fell into this category. The final dataset consists of 17086 observations, still however a substantial size.

Identifying Sentiments

The sentiment of a review can be obtained in R using one of 3 sentiment lexicons:

- **AFIN (selected!)** – assigns a mathematical score to strings based on their sentiment. A pro is that this numerical score is useful as a regression variable which is why it was selected.
- **NRC** – classifies strings as indications of emotions e.g. disgust or joy, as well as classified as being overall positive or negative. A challenge is words are duplicated in the result which complicates analysis. E.g. 'wonderful' is classified as surprise, joy, and positivity - instead of one emotion it most exhibits. NRC meant including several extra variables (joy_score, disgust_score, disappointment_score etc.). It was disregarded for complexity as well as being outperformed by AFIN in model accuracy.
- **BING** – classifies reviews as just negative or positive (without emotions provided by NRC). Regression using a BING factor variable was considered. BING and AFIN both classify a review as positive or negative. AFIN scores give a more detailed measure of *how positive* or *how negative* a review is. Afin score-based regression was also more accurate.

Pitfalls Identified

Data-specific Stop Words not identified

- Sentiment analysis analyses keywords. Words such as 'like' and 'as' are stop words and are not analysed for sentiment. This is automatically done in sentiment packages in R.
- However, words can have multiple meanings. Sentiment analysis considers "top" a positive adjective (e.g. top notch) whereas in the context of clothing, top is just an item and is considered a stop word. Words such as 'fall' (meaning implied was the way a top fell while being worn) or buckle (an accessory, not a response to pressure!) caused similar issues. These data specific stop words were manually removed to improve accuracy.
- Words removed were misidentified as positive or negative meaning but really had no sentiment. This challenge was only caused due to the data nature. In most cases sentiment analysis accurately filters true stop words.

Negation and Sarcasm

- Sarcasm poses issues; "This dress was clearly made very well! The shape completely changed in the wash!" might be misclassified as positive.
- Another pitfall of sentiment analysis is negation. "I doubt anyone could find a fault with this dress" might receive a negative sentiment score. 'Fault,' and 'doubt' would be perceived as negative and other stop words which ultimately change the meaning to positive are ignored. The library sentiment (which provides the selected Afin lexicon) deals with negation, so negation issues were technically automatically solved.
- However, negation and sarcasm were not always accurately identified. Some reviews were misclassified. The only identified solution was to manually alter incorrect scores. This was impossible due to the data size. Hence, the scores were not fully accurate.

Final Data

Wordcount and AFIN scores are computed columns. The review text and title were removed as they are reflected in their score. Variable type was recoded as necessary.

Variable	Description	Type
Age	Age of buyer	Numeric
Recommended_IND	1 (recommended) or 0 (not recommended).	Factor
Positive_Feedback_Count	Dataset only differentiated between types of clothing e.g. Knits. This variable is count of specific garment's existing positive reviews e.g.: Beige woollen knit (not in dataset).	Numeric
Department_Name	5 departments such as 'dresses' or 'trend'.	Factor
Class_Name	20 clothing categories in departments such as 'Layering' or 'Sleep'	Factor
afin_score_review	(Afin) Calculated sentiment score of review content.	Numeric
Afin_score_title	(Afin) Calculated sentiment score of review title	Numeric
Wordcount	Count of words in review content	Numeric

Note: Ratings (1-5) column was dropped.

It is assumed that when predicting a customer recommendation, the rating will be given at the same time. Therefore, we can't use rating as a predictor variable for recommendation. If this is not the case, and the rating can be included, the final model accuracy increases from 85% to approx. 97%. (The higher you rate an item the more likely you are to recommend it.)

Evaluation Methods

With this data, it is not essential to minimise false positives (as it would be in identifying cancerous tumours, for example). Accuracy is used to measure the model success (the percentage of correctly classified reviews (TP + TN)/(TP + TN + FP + FN)). Accuracy is used instead of precision, recall etc. as the cost of a misclassifying a clothing review is not detrimental [6] [7].

Variable Selection

Lasso, Ridge, and elastic net were considered to develop a model. The produced coefficients are available in appendix 1. A train/test/validation split of 60:20:20 was used. A split of 70:15:15 gave similar accuracy results. It was assumed several variables would be used as predictor variables. Manual variable selection would be too complex using a generalised linear regression model (GLM).

LASSO has the highest accuracy measure (77%) and was deemed the best variable weight indicator. LASSO disregarded variables class and department names.

They had little to no correlation with recommendation. If included, the model would involve 23 extra factor variables adding huge complexity without an accuracy increase.

LASSO weighted the variables;

- afin score for review title
- afin score for review content
- number of positive reviews
- wordcount of review

Number_postive_reviews should be disregarded, despite being given a weight by the LASSO algorithm. Its coefficient of -0.013 is;

a) Extremely close to 0
b) Negative – in the context of the data, number of positive reviews should have a positive weight when predicting whether a user recommends an item.
c) Only improved accuracy by 4%, which may only have been a coincidence.

Word count was also disregarded as its coefficient was -0.004. Ridge and elastic net models of regression also gave wordcount a weight with absolute value less than 0.01. Again, the accuracy only improved very marginally by including this variable.

Final predictor variables: review title sentiment score and review content sentiment score

Fitting the model

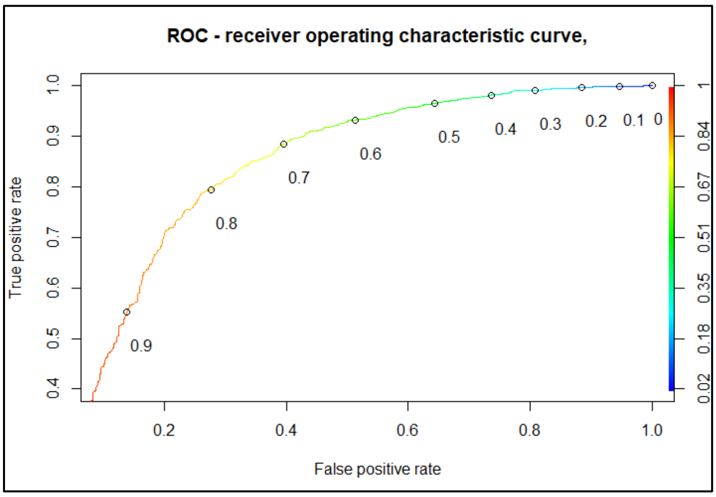
Using the variables identified using LASSO, a GLM was fitted. The GLM model improved the accuracy over LASSO by approximately 10%. If more than two predictor variables were identified LASSO may have been preferable.

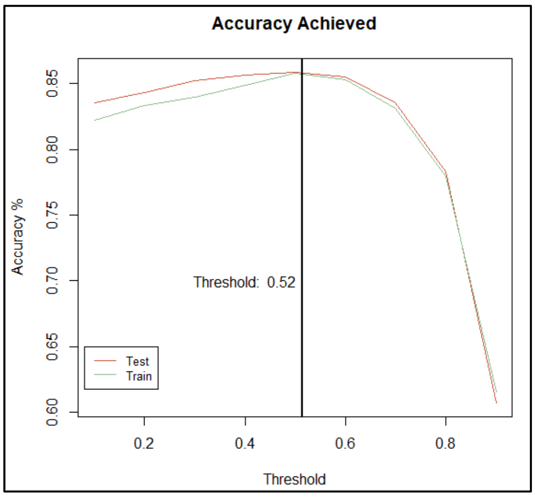
Including variables disregarded by LASSO selection (such as department name) was tested in the GLM, to ensure that variables were still optimal using GLM. The AIC was lowest, and accuracy highest for the 2 (LASSO-identified) variables in our GLM model.

Choosing a threshold

As mentioned, misclassifying reviews is not detrimental. From the ROC curve, the model is prone to giving numbers of False Positives for many thresholds. Thresholds between 0.5 and 0.7 balance the trade-off of maximising True Positives and minimising False Positives. It is difficult to identify the most effective threshold from the ROC curve alone.

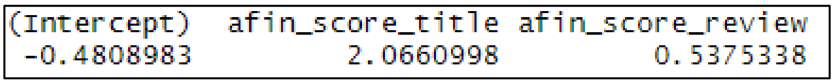
As accuracy is the chosen performance metric, the threshold that maximises accuracy for both the training and testing dataset is 0.52. This concludes a probability higher than 0.52 meant the buyer will recommend the product and vice versa.





Results

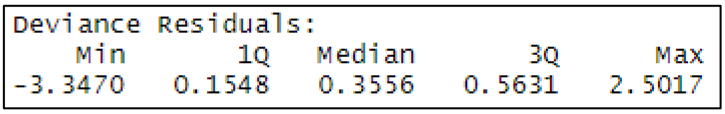
GLM was initially disregarded for the complex variable selection process involved with 10 variables. However, LASSO was used and only 2 variables were adequate predictors. GLM was used as it gave the highest accuracy with these variables. A threshold of 0.52 was selected to maximise accuracy. The final GLM had the coefficients;



```
(Intercept)  afin_score_title afin_score_review
 -0.4808983         2.0660998         0.5375338
```

The review title had a very large effect on the outcome variable but also had a larger standard error.

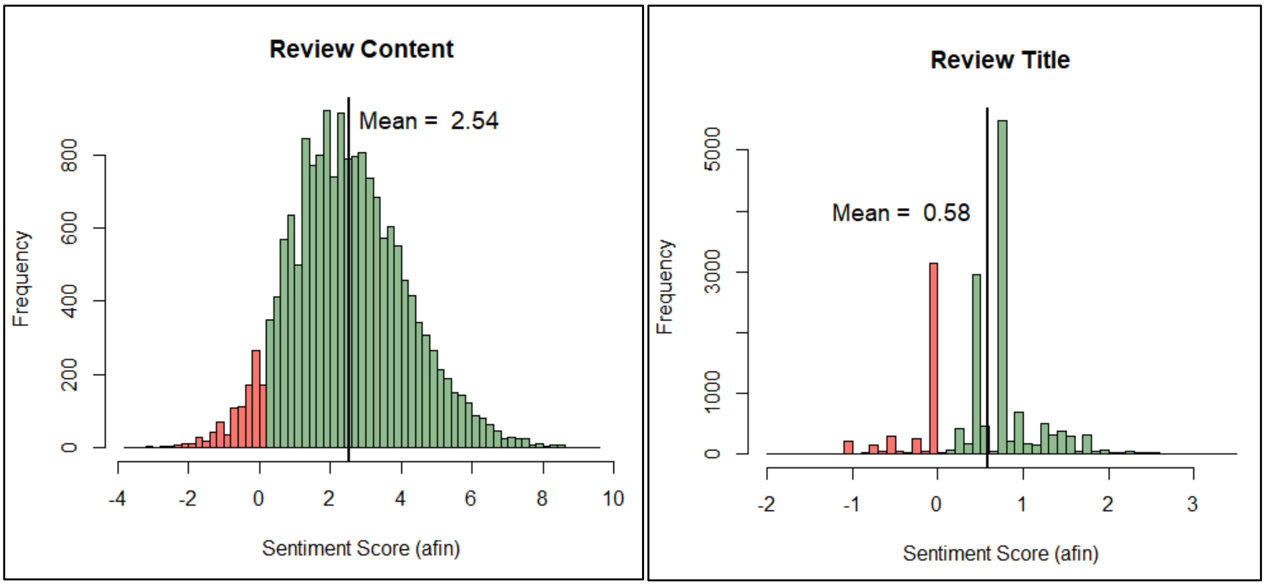
Accuracy achieved was 0.8528379. The deviance residuals from the fitted model are displayed below. The model residuals were plotted but had no indication of an underlying pattern or distribution that might be included to improve the model. The model is slightly skewed with median 0.3556 (perfect model has median 0).



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3470   0.1548   0.3556   0.5631   2.5017
```

The model had 0.96 sensitivity and 0.39 specificity.
In general, high sensitivity tests have low specificity and indicate the model produces a high rate of false positives. This was evident in the ROC curve.

Sentiment Distribution

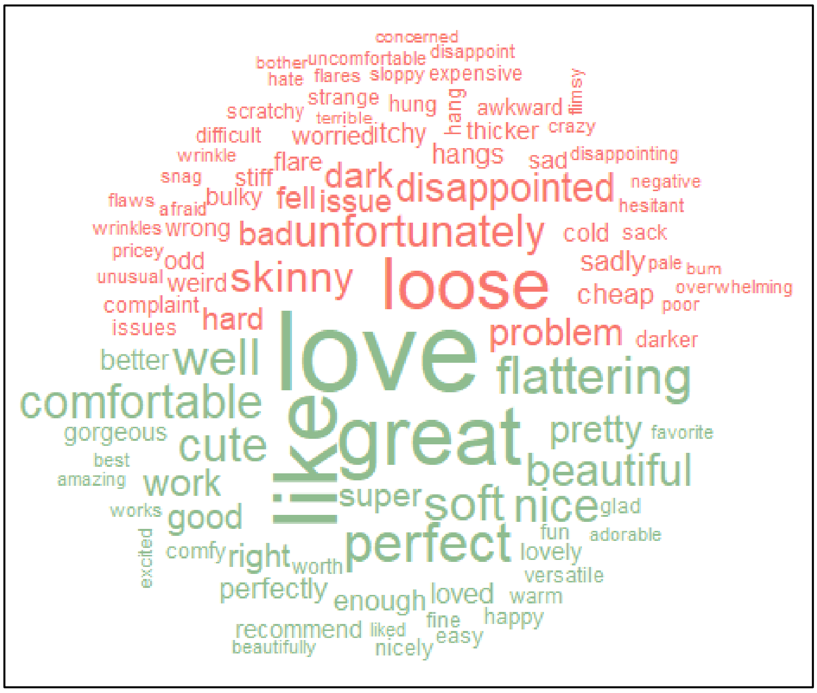


The content score mimics a normal distribution with mean 2.5 and standard deviation 1.6. The title score does not appear to follow a distribution. It was unexpected that the distributions did not match, as the title and content sentiment should be similar. The content score mean was higher, perhaps buyers are less expressive in their title.

For content and title, positive scores are more common and numbers trail as scores become more positive. Negative title and content scores are less common. Most negative reviews are just under 0. The title score has an obvious spike just under 0 - *negative reviews are only slightly negative, while positive reviews are very positive.*

Review	***Recommended?***	***Average Afin Sentiment Score***
Title	No	*0.11*
Content	No	*1.39*
Title	Yes	*0.68*
Content	Yes	*2.79*

As shown in the table, mean title and content scores for buyers who did not recommend their product are 0.11 and 0.68 respectively. Afin scores may overestimate as they would have been predicted to be under 0 (negative).



The word cloud displays the most common words that contributed to a review's sentiment. Words such as 'love', 'like', and 'great' were most common in positive (green) reviews. 'Disappointed', 'unfortunately' and "loose" contributed mostly to negative (red) reviews. As there as larger sized positive words, *positive reviews had more words in common*. Negative reviews had a wider range of vocabulary.

Discussion

85% accuracy was achieved. If repeated with another dataset, this should improve. Variables such as age might display correlation with recommendation and be considered in the final regression model. It was unexpected that the number of positive reviews an item has would have no substantial effect on item recommendation. Likewise, for age and item bought. Prior to exploratory analysis, it was predicted that certain items (e.g. hard to fit items such as jeans) would be less likely to be recommended.

It is possible that the variables were produced data. The reviews did however appear to be genuine. The sentiment scores alone provided a decent prediction accuracy. This is promising for repeated analysis with more appropriate data.

Review title sentiment had a large coefficient (2.07). The title was a more accurate predictor for buyer recommendation. This is possibly because the title is a more concise measure of attitude towards products. Negation or sarcasm is more likely to appear in review content, throwing off sentiment analysis. The title is usually straight to the point. Therefore, sentiment analysis is more effective on the title.

The ROC curve and Confusion Matrix identified the model is prone to giving False Positives.

Challenges

A major challenge was the lack of correlation of variables with recommendations. Initially, the analysis was going to combine multiple variables with the sentiment of a review for a regression model.

The dataset was quite large (17086 with data cleaning). Running functions on this dataset in R was time-consuming and occasionally not supported on the machine used. Code was run with smaller proportions of the data to get a sense of how models were performing. Techniques and functions were refined on the sub-dataset. Results were obtained from the full dataset.

Conclusion

We must consider that people leaving reviews will likely be either extremely happy or extremely disappointed. Analysis of reviews will fail to analyse customers in the middle 'satisfaction bracket'.

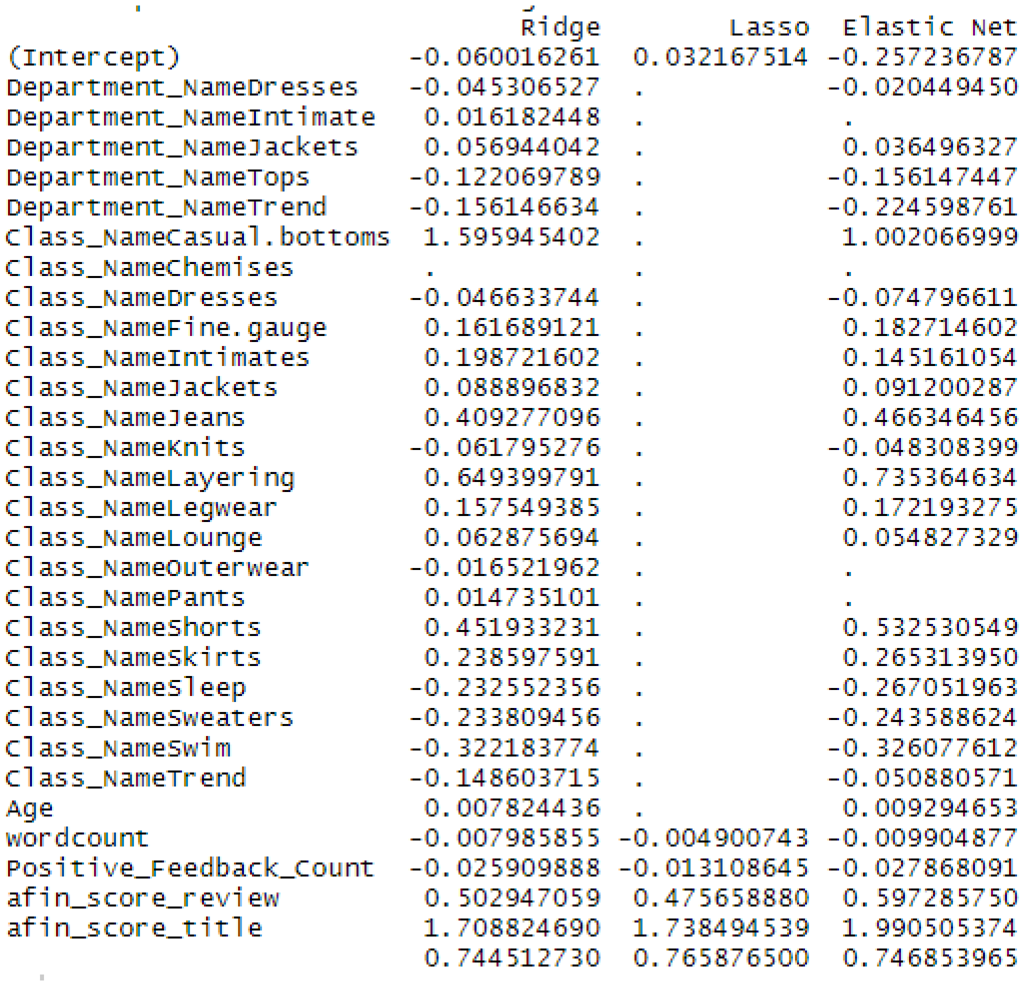
Online reviewers might also be of a certain age bracket or demographic. It was predicted that buyer recommendations would be heavily influenced by age and item type. According to our model, this was not the case. However, this is not conclusive as some of these columns may not genuine data, but produced data combined with online reviews.

Future Recommended Research

- This dataset included only females. This analysis could be repeated with males and/or a different product type such as electronics or home goods.
- As mentioned, the age of buyers and the proportion of different items purchased seemed to almost perfectly mimic distribution. Some data appeared to be produced and variables that should have been correlated with recommendation were not. However, the reviews seemed sound enough and ended up the only accurate predictor variables. Future research could involve combining review sentiment with buyer demography and item purchase. This could be done by sourcing another dataset.
- Further research could analyse how the title sentiment performs as a predictor without the content sentiment. This would be useful for very large datasets, as the titles were quicker to analyse, being shorter than content. Sentiment analysis of the title is also more accurate. A title coefficient score of 2.07 (versus 0.54 for content) indicates it could potentially stand alone as an accurate prediction.

Appendices

Appendix 1 – Coefficients for Ridge, Lasso and Elastic Net model



```
                                Ridge        Lasso  Elastic Net
(Intercept)               -0.060016261  0.032167514 -0.257236787
Department_NameDresses    -0.045306527  .           -0.020449450
Department_NameIntimate    0.016182448  .            .
Department_NameJackets     0.056944042  .            0.036496327
Department_NameTops       -0.122069789  .           -0.156147447
Department_NameTrend      -0.156146634  .           -0.224598761
Class_NameCasual.bottoms   1.595945402  .            1.002066999
Class_NameChemises         .            .            .
Class_NameDresses         -0.046633744  .           -0.074796611
Class_NameFine.gauge       0.161689121  .            0.182714602
Class_NameIntimates        0.198721602  .            0.145161054
Class_NameJackets          0.088896832  .            0.091200287
Class_NameJeans            0.409277096  .            0.466346456
Class_NameKnits           -0.061795276  .           -0.048308399
Class_NameLayering         0.649399791  .            0.735364634
Class_NameLegwear          0.157549385  .            0.172193275
Class_NameLounge           0.062875694  .            0.054827329
Class_NameOuterwear       -0.016521962  .            .
Class_NamePants            0.014735101  .            .
Class_NameShorts           0.451933231  .            0.532530549
Class_NameSkirts           0.238597591  .            0.265313950
Class_NameSleep           -0.232552356  .           -0.267051963
Class_NameSweaters        -0.233809456  .           -0.243588624
Class_NameSwim            -0.322183774  .           -0.326077612
Class_NameTrend           -0.148603715  .           -0.050880571
Age                        0.007824436  .            0.009294653
wordcount                 -0.007985855 -0.004900743 -0.009904877
Positive_Feedback_Count   -0.025909888 -0.013108645 -0.027868091
afin_score_review          0.502947059  0.475658880  0.597285750
afin_score_title           1.708824690  1.738494539  1.990505374
                           0.744512730  0.765876500  0.746853965
```

References

[1] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9730&rep=rep1&type=pdf

[2] https://www.cs.uic.edu/~liub/publications/www05-p536.pdf

[3] https://www.tidytextmining.com/sentiment.html

[4] https://www.sciencedirect.com/science/article/pii/S2090447914000550#t0005

[5] https://ieeexplore.ieee.org/abstract/document/7415179

[6] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/#:~:text=Yes%2C%20accuracy%20is%20a%20great,false%20negatives%20are%20almost%20same.&text=Precision%20%2D%20Precision%20is%20the%20ratio,the%20total%20predicted%20positive%20observations.

[7] https://towardsdatascience