Customer Segmentation and Recommendation

Using Age and Reviews

Gourav Verma

DSC-550 Machine Learning

Bellevue University, NE

Summer-2020

Abstract

Most of the marketing companies are able to segment the customers in order to create personal, timely and pertinent content, but how many companies know how their audience feel before they spend their scrutinised marketing budgets trying to get them to spend more? People will forget what you said, people will forget what you did, but people will never forget how you made them feel –this statement holds extremely true in the world of retail as well. By analysing these sentiments accurately and analysing the things that upsets the customer, a retailer can focus more on what will make a difference. We can leverage machine learning technology as an opportunity to find it. In this project women-clothing review dataset from Kaggle is used which has many variables like age of the customer, clothing review text, ratings provided by customers, unique clothing ids, and department. To generate sentiment scores from the review text given by women on clothing items, the AFINN lexicon library was used. Later K-Mean clustering algorithm was applied to the sentiment score to generate customer segments based on their age. Review ratings by the customer and unique clothing item ids and age parameters from the dataset were used to build a recommendation model. Using the model an unknown user-rating can be identified, moreover, this model can be leveraged to recommend a clothing item to a woman based on her age. For building the recommendation model surprise package from the scikit group is used and post performing cross-validation BaselineOnly algorithm was chosen for the task. Code base for the project can be found at https://github.com/GARV3007/Data-Mining-using-Python/tree/master/Term%20Project

Keywords: K-Mean, AFINN, elbow method, clustering, scikit-surprise, recommendation model, BaselineOnly.

Table of Contents

Data

This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions. Because this is real commercial data, it has been anonymized, and references to the company in the review text and body have been replaced with "retailer". The dataset is found on kaggle.com. The bullet list below explains the features as given by kaggle.

Features

Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.

Age: Positive Integer variable of the reviewer's age.

Title: String variable for the title of the review.

Review Text: String variable for the review body.

Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 worst, to 5 best.

Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.

Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.

Division Name: Categorical name of the product high level division.

Department Name: Categorical name of the product department name.

Class Name: Categorical name of the product class name.

Customer Segmentation

This research aims to gain insight into the sentiments of the women customers using their review comments on clothing items. This section further discuss the research method, the sampling method and the data analysis procedure. A company can get voice of customers in many ways (not survey), like recordings from customer services, product reviews, ratings etc. Customer segmentation can be used by retailers to target individuals to improve their service, track customer sentiment over time, determine if a particular customer segment feel more strongly about your product, track how a change in product or service affects how customers feel.

Sentiment Analysis

Sentiment analysis refers to assigning a metric to a piece of text that details how positive or negative said text is. Sentiment analysis is a method of machine learning that allows bots to go through customer reviews and feedback to determine whether the comments are positive, negative, or neutral. Sentiment analysis can be helpful in identifying negative comments that you can then have someone personally address (and improve your customer service) as well as identifying customers who have provided high praise. Such customers can then be turned into brand ambassadors. The sentiments expressed by the customer have vital data hidden in them. With the help of sentiment analysis, a retailer can classify whether a customer is satisfied, happy, or irate by the product or the services provided by the retailer. Sentiment analysis categorizes the feedbacks on the basis of the mood of the customer. This allows the retailer to improve the marketing and sales strategies which as a result leads to a better customer retention rate and a higher profit margin. If your brand is getting a lot of criticism online, machine intelligence will help you identify that in real-time so that you can take appropriate actions and solve the issue

before it becomes a major crisis. We will apply Afinn lexicon to get the sentiment score of the women review texts in the database.

AFINN Lexicon

AFINN is an English wordlist-based approach for sentiment analysis developed by Finn Årup Nielsen. Words scores range from minus five (negative) to plus five (positive). The current version of the lexicon is AFINN-en-165.txt and it contains over 3,300+ words with a polarity score associated with each word.

K-Means Clustering

K-Means clustering is a type of unsupervised machine learning that uses data that is not assigned to a specific category or group, called unlabeled data. The algorithm works assigns the data points values based on features that are inputted. K-Means clustering finds the most significant number of clusters and then determines the grouping of clusters based on the distance between the data points based on inputted attributes. As the clusters are formed directly from the data itself rather than business rules or filtering, the results correspond to the direct truth of the data. That means that marketers can segment their customers based on examining the data alone rather than on preconceived notions. Attributes that have not been traditionally considered might suddenly appear highly relevant. Customer segments are dynamic. Due to its direct nature, K-Means clustering picks up on changes in clustering over time. Depending on data availability, K-Means clustering can deliver real-time or near-time customer segmentation.

Data Preparation.

To begin with, the offline csv file was downloaded from the keggle.com and using python pandas package the file was read into the dataframe. For K-Mean unrequired variables were dropped and Age, Review text, Rating, and positive feedback counts were only kept. Rows

with null values were dropped from the dataframe. Before applying AFINN score calculation Review texts were converted into lower case. Afinn scores for each review text were stored in the sent_score column. A positive score means positive sentiment and negative score means negative sentiment. To get the distribution on the positive side (>= 0) lowest negative score values were added into each score. The distribution plot drawn using the seaborn package showed skewed distribution for Age and sent_score variable (Figure 1). The Log values of the variables showed normal distribution (Figure 2).

Application.

As the K-mean clustering requires, normally distributed variables, log values were used to get the K value by the elbow method (Figure 3). From the graph, I selected the K value as 6. K-Mean from the sklearn package was applied to the data. (Table 2) shows the count and percentage for each cluster values. The cluster graph was plotted using seaborn lmplot (Figure 4) for age and sent_score.

Outcome.

These clusters can be interpreted based on their age and sentiment scores. A marketing company can leverage this information to focus on a selected group and customize their strategy to have more satisfied customers.

0 - Middle Aged customers with Negative reviews

1 - Middle Aged customers with Fairly Positive reviews

2 - Older customers with positive reviews

3 - Young customers with fairly positive reviews

4 - Older customers with fairly positive reviews

5 - Fairly young customers with positive reviews

a) The ranters (unhappy customers, Cluster 0) - In most businesses, this segment of customers will be handled by customer service team, but, don't try to sell to them. Instead, we can ask them what we could change to make their experience better or ask them for their ideas to engage them.
b) The 'on the fence' bunch (Cluster 1, 3 & 4) - These are the trickiest group as it's hard to get a true understanding into how they feel and whether or not the marketing efforts will be receptive or not. A good approach with this group could be to get ravers to engage with them.
c) The ravers (happy customers, Cluster 2, 5) - We know they are happy, so let them tell the world. Give them some sort of incentive to promote your services on social media (small is fine and it will help drive repeat or additional purchases too). Their comment (positive word of mouth) next to an exclusive offer for their friends and family is ideal. This type of activity is often referred to as advocacy marketing.

Recommendation

Any organization looking to do business on the internet is interested in what its customers have to say. We want to know if we can predict rating based on the content of the review using recommendation model. We also want to see if we can predict how likely a customer will recommend products to their friends. A recommender system, or a recommendation system, can be thought of as a subclass of information filtering system that seeks to predict the best "rating" or "preference" a user would give to an item which is typically obtained by optimizing for objectives like total clicks, total revenue, and overall sales. The basic principle of recommendations is that there are significant dependencies between user- and item-centric

activities. I used collaborative filtering with the inbuilt recommender algorithm in the surprise package and for the evaluation cross-validation was done with RMSE and MAE measurements.

Scikit-Surprise

Surprise (Simple Python Recommendation System Engine) is a recommendation system library, which is one of the scikit series. Simple and easy to use, while supporting a variety of recommendation algorithms (basic algorithm, collaborative filtering, matrix decomposition, etc.). In Collaborative filtering, the model learns from the users past behaviour, user's decision, and preference to predict items the user might have an interest in. Scikit-Surprise is an easy-to-use Python scikit for recommender systems.

Application.

For the analysis Age, Clothing ID, and Rating variables were chosen. Selected 25% of random records as a test set. Using the cross-validation method from surprise package evaluated RMSE and MAE scores on 5 folds for SVD, KNNBasic, KNNWithMeans, KNNWithZscore, BaselineOnly, and NormalPredictor algorithms (Table 3).

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. Root mean squared error (RMSE) is a quadratic scoring rule that also measures the average magnitude of the error. Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better.

All algorithms showed similar RMSE and MAE values. I selected the BaselineOnly algorithm for building the model, because, a baseline prediction algorithm provides a set of predictions that you can evaluate as you would any predictions for your problems, such as classification accuracy or RMSE. BaselineOnly model was build using alternate least squares (ALS) method. The regularization parameter for items was kept as '5' and that for users was kept as '12'. The number of iterations of the ALS procedure was chosen as '5'. Model cross-validation is shown in (Table 4). A prediction model was made by fitting the trainset on the model along with the test using the preselected testset. (Table 5) and (Table 6) shows the best and the worst predictions respectively.

Conclusion

Analysis of women's clothing review dataset shows that we can segment the customer base into different clusters that can be utilized to understand the customer sentiments better. Moreover, a marketing agency can effectively alter its strategy based on different segments to upgrade customer satisfaction. We applied K-Mean clustering to assign each customer into different clusters based on their sentiment score generated by their clothing review comments. After plotting their sentiments vs age and using color coding based on clusters we clearly see different customer segments. Broadly we can divide the customers into three groups, ranters (cluster 0), on the fence bunch (clusters 1, 3, and 4), ravers (clusters 2 and 5). Additional analysis has been made on the dataset to create a recommendation model. The model was build using clothing item numbers and different customer ratings against it. We can use this model to identify unknown ratings by customers for a clothing item. Also, this can be leveraged to predict customer response in advance before launching a new product or changing any marketing strategy.

Table 1 Sample data from file

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

Table 2 Cluster wise count and Percentage

	clusters	count	percentage
0	0	2622	11.58
1	1	5564	24.57
2	2	3880	17.14
3	3	3111	13.74
4	4	4007	17.70
5	5	3457	15.27

Table 3 Cross-validation results

Algorithms	Test_RMSE	Test_MAE	Fit_time	Test_time
BaselineOnly	1.098524	0.883113	0.025738	0.023345
KNNWithMeans	1.121399	0.891298	0.153384	0.908731
KNNWithZScore	1.124861	0.892554	0.146855	0.947234
KNNBasic	1.127862	0.886328	0.147432	0.955465
SVD	1.132511	0.903668	0.974561	0.021864
NormalPredictor	1.427513	1.083934	0.029032	0.027179

Table 4 Cross-validation results for BaselineOnly model

Test_RMSE	1.12082879	1.09012101	1.09077031	1.09164502	1.10357604
Fit_time	0.02892708	0.03789973	0.02992057	0.02293848	0.01392793
Test_time	0.02009034	0.06956839	0.03490567	0.02396869	0.02696514

Table 5 Best Predictions

	Age	Cid	rui	est	details	Uct	Ict	err
781	28	864	4.0	4.000459	{'was_impossible': False}	332	117	0.000459
5160	48	871	4.0	4.001413	{'was_impossible': False}	454	40	0.001413
3895	50	941	4.0	3.998346	{'was_impossible': False}	305	63	0.001654
950	42	857	4.0	4.003275	{'was_impossible': False}	484	52	0.003275
4801	51	1089	4.0	4.003652	{'was_impossible': False}	283	66	0.003652
5582	41	900	4.0	4.004853	{'was_impossible': False}	533	11	0.004853
249	46	861	4.0	3.994305	{'was_impossible': False}	527	185	0.005695
1925	40	155	4.0	3.994138	{'was_impossible': False}	460	8	0.005862
573	55	827	4.0	4.005927	{'was_impossible': False}	249	6	0.005927
2345	46	1052	4.0	4.006146	{'was_impossible': False}	527	24	0.006146

Table 6 Worst Predictions

	Age	Cid	rui	est	details	Uct	Ict	err
5088	39	719	1.0	4.465238	{'was_impossible': False}	969	8	3.465238
4886	39	850	1.0	4.466607	{'was_impossible': False}	969	234	3.466607
4608	63	872	1.0	4.477956	{'was_impossible': False}	182	414	3.477956
2922	41	768	1.0	4.496127	{'was_impossible': False}	533	1	3.496127
4565	54	907	1.0	4.504113	{'was_impossible': False}	304	103	3.504113
5399	65	940	1.0	4.523195	{'was_impossible': False}	161	94	3.523195
2470	21	824	1.0	4.527146	{'was_impossible': False}	76	72	3.527146
2773	41	1048	1.0	4.562624	{'was_impossible': False}	533	7	3.562624
5360	39	939	1.0	4.634485	{'was_impossible': False}	969	62	3.634485
4052	32	964	1.0	4.698071	{'was_impossible': False}	456	57	3.698071

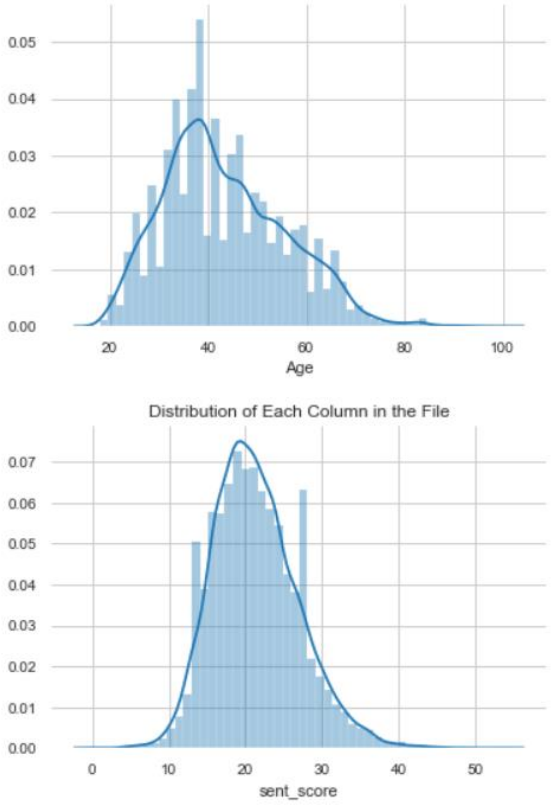

Figure 1 Distribution plot for Age and sent_score

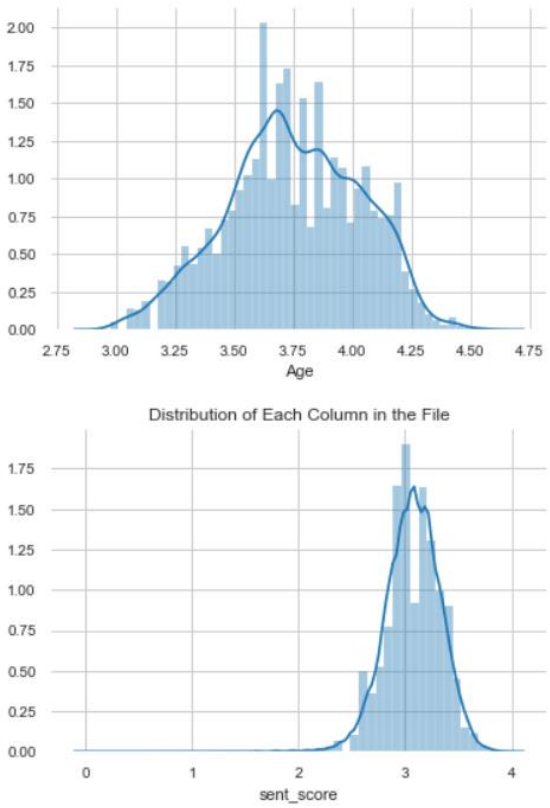

Figure 2 Distribution plot for log values of Age and sent_score

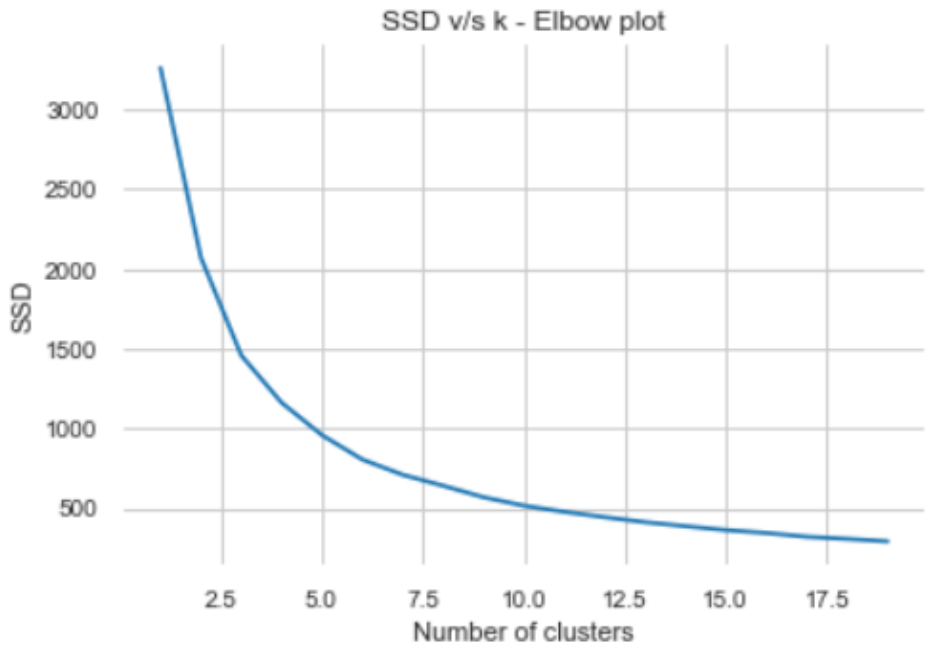

Figure 3 Elbow plot

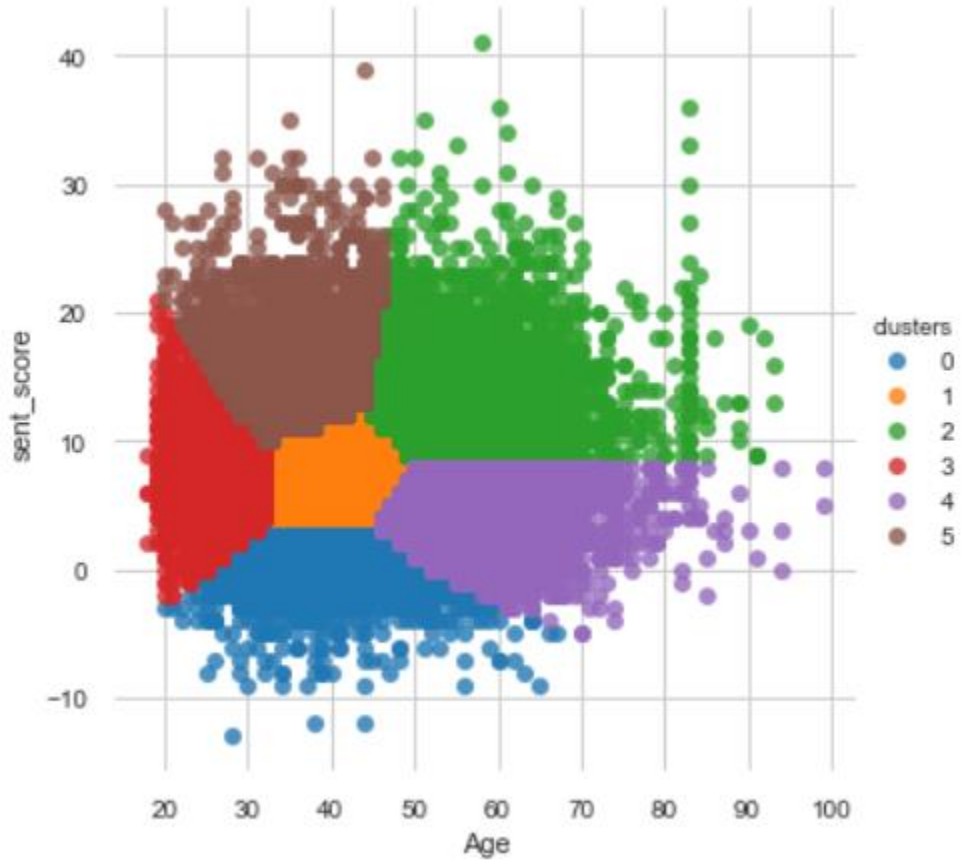

Figure 4 Cluster Plot

References

Brooks, N. (2018, February 3). *Women's E-Commerce Clothing Reviews*. Kaggle. https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

Nielsen, F. Å. (2011). A new ANEW: *Evaluation* of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Nielsen, F. Å. (n.d.). fnielsen*/afinn*. GitHub. https://github.com/fnielsen/afinn

Riaz, S., Fatima, M., Kamran, M., & Nisar, M. W. (2019). Opinion mining on large scale data using sentiment analysis and k-means clustering. *Cluster Computing*, 22(3), 7149-7164.

Korovkinas, K., Danėnas, P., & Garšva, G. (2019). SVM and k-means hybrid method for textual data sentiment analysis. *Baltic Journal of Modern Computing*, 7(1), 47-60.

PyCon 2018. (2018, May 13). *Daniel Pyrathon - A practical guide to Singular Value Decomposition in Python - PyCon 2018*. YouTube. https://www.youtube.com/watch?v=d7iIb_XVkZs&feature=emb_title

Hug, N. (n.d.). *Welcome to Surprise' documentation! — Surprise 1 documentation*. Surprise. https://surprise.readthedocs.io/en/stable/

Li, X., Zhao, H., Wang, Z., & Yu, Z. (2020, May). Research on Movie Rating Prediction Algorithms. In 2020 *5th IEEE International Conference on Big Data Analytics (ICBDA)* (pp. 121-125). IEEE.

Yoon, S., Parsons, F., Sundquist, K., Julian, J., Schwartz, J. E., Burg, M. M., ... & Diaz, K. M. (2017). Comparison of different algorithms for sentiment analysis: Psychological stress notes. *Studies in health technology and informatics*, 245, 1292.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1), 79-82.