Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
A		# 2010755

XXX

Summary

Contents

1 Introduction

2 Assumption

There are some symbols appear in the model. We show them below:

3 Data preprocessing and Descriptive statistics

3.1 Data preprocessing

For data preprocessing on the data provided by Sunshine company, our team mainly considers two aspects. The first is the removal of useless data, and the second is the quantification of text data.

Table 1: Symbols in Chapter 3

Symbols	Description
i	Station variable
DS_i	Density of fish(mackerel or herring) at station i (kg/km^2)
H_i	Horizontal opening of trawl at station i (km)
TD_i	Distance of the trawl haul (km^2)
C_i	Catch at station i(kg)
λ_i	Longitude at station i $(^\circ W)$
ϕ_i	Latitude at station i $(^\circ N)$
SST_i	Sea Surface Temperature at station i $(^\circ C)$
z_i	Zooplankton's dry weight at station i (kg)
SSB_i	Spawning-stock biomass at station i
b_i	Number of biological species at station i
y	Year
j	Rectangular number
t	Time
$M_j(t)$	State of the j_{th} rectangle at time t
P_{M_w,M_k}	Transition probability for the State $M_w$changing into State M_k

Table 2: Symbols in Chapter 4

Symbols	Description
B	Backshift operator

Table 3: Symbols in Chapter 5 & 6

Symbols	Description
$CPUE_y$	Catch Per Unit Effort in year y

3.1.1 Removal of meaningless data

By observing the data given in the question, we find that there is a lot of irrelevant data. First, remove fields that are not relevant to the data analysis, such as "marketplace","customer_id","review_id",

"product_id","product_category" We remove extranical product evaluations in the dataset, such as the product evaluations for pillow driers in the baby pacifier product evaluation dataset. We also found that there were a lot of user comments that were meaningless, and we culled them out.

3.1.2 Data type conversion

The data marks "N" and "Y" are converted to "0" and "1", and the string data is converted to floating point numbers, which is convenient for the next statistical analysis and prediction.

4 The relationship between star rating and review text based on Logistic Regreesion

4.1 Quantification of review text based on TF-IDF Method

Now, we want to research the relationship between reviews and star rating. Generally speaking, the more positive the comment is, the higher the star rating will be, and the lower the star rating is, the more negative the comment will be.

On the one hand, we verify whether this conclusion is valid. On the other hand, we want to understand the features that people who give good reviews are more interested in and what features that people who give bad reviews complain about most. In order to achieve this purpose, we use TF-IDF technology to quantize review according to its importance in all reviews.

We concatenated the "review_headline" with the "review_body" and removed extraneous punctuation. Then we uniformly convert the letters to lowercase and use the spaCy method to convert the word into its root form, which is convenient for subsequent text analysis.

Then, deal with the data of star rating. We denote 'one star','two star' , 'three star' as 0 (negative), and others as 1 (positive). Thus we gain a series data of 0-1 from star rating data. After that, we get some words that have a great influence on consumers' attitude through logistic regression between the above 0-1 series and those quantified review.Thereupon, some product features consumers interested in most can be concluded, which Sunshine Company need to track once their three products are placed on sale in the online marketplace.

TF-IDF is a statistical method to reflect the how important a word is to a document in a collection or corpus.The following is TF-IDF formula:

$$tf - idf(t,d) = tf(t,d) \times idf(t) \tag{1}$$

$$tf(t,d) = \frac{n(t,d)}{\sum_k n_{k,d}} \tag{2}$$

$$idf(t,d) = log(\frac{n_d}{df(d,t)}) + 1 \tag{3}$$

$tf(t,d)$ is tf value, indicating the frequency fo term t in a text d.

$idf(t)$ is the calculation formula of idf value of term t; nd is the number of text in training set;

$df(d,t)$ is the total number of documents containing term t. The idf value is an improvement of word weight, not only considering the frequency of words in the text, but also the frequency of words in the general text.

To avoid 0 occurring in the denominator, the formula of idf we actually use is

$$idf(t) = log(\frac{1+n_d}{1+df(d,t)}) + 1 \tag{4}$$

4.2 Model Comparison

We labeled reviews with a score greater than 3 as 1 (positive), and those with a score less than 3 as 0 (negative).

We use the method of TF-IDF word vector to vectorize the text.

We first vectorize the text using the sklearn library's method of generating TF-IDF word vectors.

Using positive/negative tags, we use Bernoulli Naive Bayes classifier, Multinomial Naive Bayes classifier, Support Vector Machine(SVM) ,Stochastic gradient descent Classifier(SGD Classifier) and Logistic Regression model for text sentiment analysis. Then, we evaluated the fitting effects of the five models using four indexes, namely precision rate, recall rate, F1 measurement and ROC curve.

By comparison, the logistic regression model is the best.

* Precision rate: As can be seen from the following table, the accuracy rate of Logistic Regression model is as high as 89%.

The precision rate of each classifier

Classifier	Logistic Regression	SGD Classifier	SVM	Bernoulli BN	Multinomial BN
Precision rate	0.89276	0.89186	0.87873	0.84027	0.83077

* Recall,f1-score measurements and metrics It can be seen from the Figure 1 that the recall rates of SGD Classifier, logistic regression model and support vector machine are almost the same, while the F1 score of logistic regression model is higher, indicating that the logistic regression model has better performance.

* ROC curve: AUC (Area Under Curve) is defined as the Area Under the ROC Curve. In many cases, the ROC curve cannot clearly indicate which classifier has better effect, so we use AUC value as the evaluation criterion of the model: classifier with larger AUC has better effect.The result is shown in Figure 2 .

4.3 Logistic Regression

Logistic regression is a classification model,whose principle is as follows.

LogisticRegression:

	precision	recall	f1-score	support
positive	0.75	0.57	0.65	382
negative	0.91	0.96	0.94	1828
accuracy			0.89	2210
macro avg	0.83	0.77	0.79	2210
weighted avg	0.89	0.89	0.89	2210

SGDClassifier:

	precision	recall	f1-score	support
positive	0.76	0.56	0.64	382
negative	0.91	0.96	0.94	1828
accuracy			0.89	2210
macro avg	0.83	0.76	0.79	2210
weighted avg	0.89	0.89	0.89	2210

Figure 1: Classifiers comparision with recall rates and f1-score

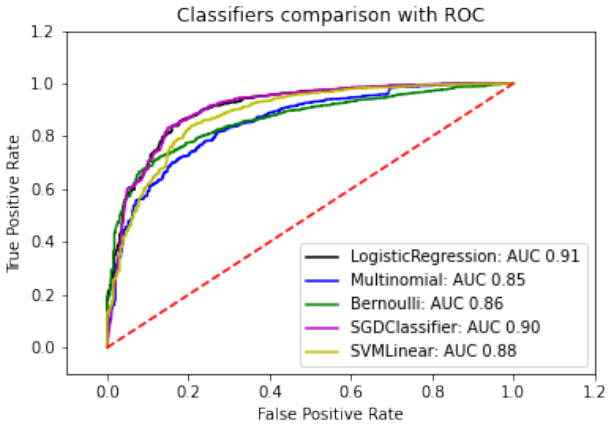

Figure 2: Classifiers comparision with ROC

For a given training data set,$T = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

Among them,$x_i \in R^n$, $y_i \in 0, 1$Assume$z = -(wx + b)$.Then we call the following probability distribution logistic regression model:

$$P(Y = 0|x) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{wx+b}} \tag{5}$$

$$P(Y = 1|x) = 1 - \frac{1}{1 + e^{-z}} = \frac{e^{-z}}{1 + e^{-z}} = \frac{e^{wx+b}}{1 + e^{wx+b}} \tag{6}$$

In the equation above ,we call the ω the weight coefficient.To facilitate the representation of multiple variables,let's introduce the weight coefficent vector $W = (w1, w2, ..., wn, b)'$ .The sample vector$X = (x1, x2, ..., xn, 1)'$.Now,the matrix of logistic regression is represented as follows.

$$P(Y = 0|x) = \frac{1}{1 + e^{WX}} \tag{7}$$

$$P(Y = 1|x) = \frac{e^{WX}}{1 + e^{WX}} \tag{8}$$

In this model, we need to estimate the weight coefficient vector$W$.The method of estimation is maximum likelihood estimation(MLE). We assume:

$$P(Y = 1|x) = \pi(x) \tag{9}$$

$$P(Y = 0|x) = 1 - \pi(x) \tag{10}$$

The logarithmic likelihood function

$$L(W) = \sum_{i=1}^{n} (y_i log(\pi(x_i))) + (1 - y_i) log(1 - \pi(x_i)) \tag{11}$$

$$L(W) = \sum_{i=1}^{n}(y_i log(\frac{\pi(x_i)}{1-\pi(x_i)})) + log(1-\pi(x_i)) \tag{12}$$

$$L(W) = \sum_{i=1}^{n}(y_i(w_i x_i - log(1+e^{w_i * x_i}))) \tag{13}$$

Let $\frac{\partial L}{\partial w_i} = 0$, We can get the maximum of the likelihood function.The vector $\hat{W}$ that we get from that is the maximum likelihood estimation of the weight parameter.According to statistical study, maximum likelihood estimation has good statistical properties.So we now have a logistic regression estimation model as follows:

$$P(Y=0|x) = \frac{1}{1+e^{\hat{W}X}} \tag{14}$$

$$P(Y=1|x) = \frac{e^{\hat{W}X}}{1+e^{\hat{W}X}} \tag{15}$$

4.4 The result of Regression

Previously, we have introduced the TF IDF algorithm to deal with the word frequency of text. We applied this algorithm to the comments of three products and took the data obtained as the independent variable of logistic regression. For example, in the hair dryer product, we extract words such as " quiet,spark' in user comments as independent variables, and the value of independent variables is obtained by the TF IDF algorithm above.As for the treatment of dependent variables ,we denote 'one star','two star' , 'three star' as 0, and others as 1.

The training sets of the obtained independent variables and dependent variables are carried out logistic regression.We used test sets to verify the accuracy of the model. The overall regression results of the three products and their accuracy are shown in Table 4.

Table 4: Logistic regression accuracy results

	features	train records	test records	Model Accuracy
hairdryer	156971	8602	2868	0.88479
microwave	51778	1211	404	0.87227
pacifier	251940	14204	4735	0.89276

According to the above table, the accuracy of logistic regression estimation of the three products is above 80%, so it can be considered that the results of logistic regression estimation are statistically satisfactory. Next, we will show the coefficient estimation obtained by logistic regression and its specific meaning.Since there are too many independent variables in logistic regression to show them all, we only show some representative variables in the following table.

The specific meaning of the coefficient of the variable obtained by logistic regression is the influence of the word frequency of the variable appearing in the comment on the star rating.If the variable is a more positive term, the coefficient is positive and larger,if the variable is a more negative term, the coefficient is negative and smaller.According to this principle, combined with the results obtained in the above table, we can find the following rules. People generally like the

heat setting in the hair dryer, and some hair dryers has the advantages of fast, light, small sound and so on. Among them, heat setting is the most attractive.But some brands of hair dryer and there will be sparks, too hot, too heavy, too loud and other shortcomings.We can see that customers are very concerned about the temperature and lightness of the hair dryer. Customers generally concern about the advantage of microwave oven,which is simple, occupying small space , clean and the low price.But the whirlpool microwave got a lot of bad reviews, and users hated the constant need for repairs and worried about its quality. In the eyes of customers, the pacifier is a perfect gift, soft but tough, and very cute.However, some customers think that the price of pacifier is too high, some of the quality of the pacifier is very poor.Even some customers hate pink pacifiers.

5 Strengths and Weaknesses

5.1 Strengths

5.2 Weaknesses

6 A Letter

MEMORANDUM
TO: Hook Line and Sinker
FROM:Team #2010755

References

[1] Olafsdottir, A. H., Utne, K. R., Jacobsen, J. A., Jansen, T., Óskarsson, G. J., Nøttestad, L., ... & Slotte, A. (2019). Geographical expansion of Northeast Atlantic mackerel (Scomber scombrus) in the Nordic Seas from 2007 to 2016 was primarily driven by stock size and constrained by low temperatures. Deep Sea Research Part II: Topical Studies in Oceanography, 159, 152-168.

[2] Wu shengnan, Chen xinjun, & liu zhonan. (2019). Prediction model of Japanese mackerel resource abundance in the northwest Pacific based on GAM. Acta oceanologica sinica, 41(8), 36-42.

[3] http://ecosystemdata.ices.dk/Map/index.aspx?Action=AddLayer&TAXA=6799&YEAR=2017&Grid=-1&Color=random&Type=Count

[4] Nøttestad, L., Anthonypillai, V., Tangen, Ø., Høines, A., Utne, K. R., Oskarsson, G. J., ... & Jansen, T. (2016). Cruise report from the International Ecosystem Summer Survey in the Nordic Seas (IESSNS) with M/V M. Ytterstad', M/V 'Vendla', M/V 'Tróndur ı Gøtu', M/V 'Finnur Fríði'and R/V 'Arni Friðriksson, 1-31.

[5] Zhang yunquan, zhu yaohui, li cunlu, feng renjie, & ma lu. (2015). Implementation of generalized additive model in R software. China health statistics, 32(6), 1073-1075.

[6] Li dewei, zhang long, wang Yang, & zhu wenbin. (2015). Analysis of the relationship between CPUE and environmental factors in Argentine sliders based on GAM. Fisheries modernization, (2015 04), 56-61.

[7] XiaoXin Han(2009). Research on the Contribution Rates of Three Industries in China Based on Markov Chain. Cooperative Economy & Science (15), 24-25.

[8] Chang, X., Gao, M., Wang, Y., & Hou, X. (2012). Seasonal autoregressive integrated moving average model for precipitation time series. Journal of Mathematics & Statistics, 8(4).

[9] Akaike H. (1987) Factor Analysis and AIC. In: Parzen E., Tanabe K., Kitagawa G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY

Appendices

Appendix A First appendix