**Prediction of Customer Ratings and Sentiment Analysis using Regression Techniques.**

**Akshay Venkataraman-** 300169348

**Table of contents.**

|  |  |  |
| --- | --- | --- |
| 1 | Goals of the project | Page3 |
| 2 | Description and exploratory analysis of the data | Page3 |
| 3 | Explanation of methodology | Page4 |
| 4 | Summary | Page5 |
| 5 | Appendix | Page 6 |

**Goals of the Project:**

Retail sales transactions that are made day to day in lives help us better understand what the customer actually needs and helps the owners to completely satisfy the customer needs. In this project I am applying sentiment analysis technique (affin scores) to understand the mood of the customer based on their reviews, also I am using Multiple linear regression technique to predict the ratings given by the customer and I am also using Logistic regression technique to estimate if a customer will recommend the store to others or not.

**Description and explanatory analysis of data.**

The data was collected from the following website- <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>

The dataset contains 23486 rows ,each row representing a individual customer and the dataset contains 11 columns where each column gives the details description about the reviews and ratings of the customer and about the type of clothes the customer purchased.

Since the number of rows were quite large, I have performed random sampling on the data and taken only 20% of the original data for predictions. Before performing any analysis I have made some data pre-processing to check for any missing values , after checking for the missing values I could find some rows with missing values , **Refer to figure 1.1 for the output**

All rows with missing values that contained text data were dropped from the dataset , and the columns with numerical variable were filled with median imputation technique. In the column name ratings, I have replaced all the zero’s with median imputation technique column wise and changed them with the median value. The type and description of the variable used for prediction are listed below

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
| Age | numerical | Age of the customer |
| Review Text | text | The detailed comment of the customer about the clothing experience. |
| Rating | numerical | The ratings are given regarding to the customer’s opinion on the product purchased. |
| Recommended IND | boolean | Contains Boolean values based on the customer decision to recommend the product to others |
| Department name | categorical | The department name is divided into 6 types of clothing sections |
| Division name | categorical | The division name is divided into 3 types of the department , the clothes belong to. |

**Explanation of the methodology:-**

**Sentiment analysis:**

Sentiment analysis was performed to briefly understand the mood of the customer while giving out the review, This will help us understand if the customer had a positive or a negative opinion about the product. I have used affin scores to compute the positivity in the review of the customer and I have also considered the emoticons in the review to deeply understand the mood of the customer, so if the affin score is greater than 1 , I have classified as ‘POSITIVE’ comment and if the affin score is negative then the comment will be classified as ‘NEGATIVE’ comment, by computing the nature of the review I have added a new column in the dataset called a “Sentiment” which I have later used this column in the regression analysis.

The output for the computed sentiment can be seen in **Figure2.1**

**K-means Clustering :**

K-means clustering technique is used to classify the data into k groups so that the data in each group has a similar property and lie closer to the centroid. To find the optimal number of clusters I have used elbow graph to find the optimal number of clusters, The elbow plot for this data can be seen in **Figure 2.2.** From the plot we can see that the 4 optimal clusters can be grouped and the following centroids for the centroids are obtained.

[-0.25085424,0.96536666]

[-0.25014491,-0.72571121]

[ 6.59064973,0.62055388]

[ 1.72375136,0.19721046]

**Multiple Linear Regression:**

For predicting the customer ratings ,I have used Rating column ( Y variable) and the corresponding X variables are:

* Age
* Recommended IND
* Sentiment
* Department name

So the multiple linear regression equation will be,

Y=Ѳ0+ Ѳ1X1+ ѲX2+ ѲX3+ ѲX4.

The values for the intercepts Ѳ, can be found in the **Figure 2.3** .I have also use another model with dropping department name and keeping the remaining X variables for prediction, The output for this model can be found in **Figure2.4**

Based on the two models , I have computed the cost function(RSS) value to compare the two models.

|  |  |  |
| --- | --- | --- |
| MODEL | RSS | R-Squared value |
| 1 | 13850.75 | 0.963 |
| 2 | 15178.59 | 0.959 |

From the above table we can see that the model 1 has lowest RSS and highest R- squared value, so Model 1 is good for prediction.

**Logistic Regression:**

Since thevariable recommended Ind is a Boolean value, logistic regression is used the predict the Y variable using the same X variable in the linear regression model, only the class name is added instead of the department name.

Y={recommend , not recommend}

X={age, Rating, sentiment, class name}

f ˆ(Y) = g(θ T X)

The out put summary of the logistic regression can be found in **Figure 2.5.** I have also used another model by dropping class name to predict the Y variable , the results of this model can be found in the **Figure 2.6**, The comparison of both the models can be seen in the below table

|  |  |  |
| --- | --- | --- |
| MODEL | Log- likelihood | R-Squared value |
| 1 | -4232 | 0.5492 |
| 2 | -4729 | 0.4964 |

Based on the table we can see the R- squared and likelihood vale to be higher in model 1 , therefore model 1 is preferred for prediction.

**Summary:**

After classifying a new variable called sentiment based on the customer reviews , I have used the variable for prediction. After performing Logistic regression to estimate the variable rating we can see that Model1 is preferred for estimating .And after performing logistic regression to predict the Boolean variable called recommendation, we can see that the model gives a better prediction results.

**Appendix:**

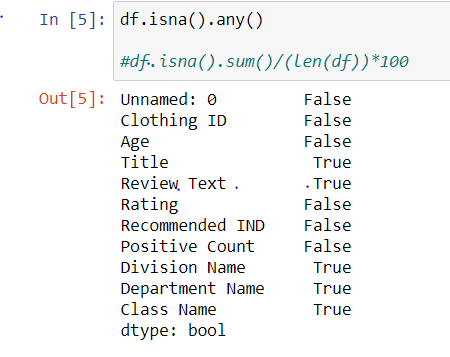


Figure 1.1

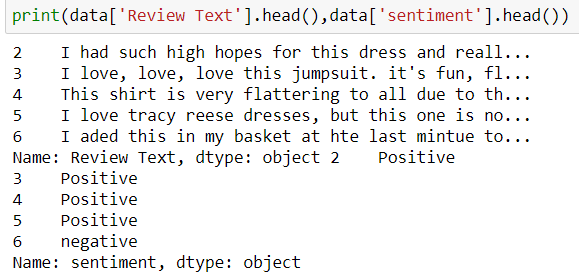


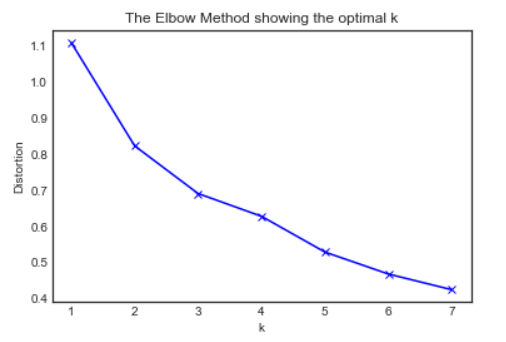
Figure 2.1 

Figure2.2

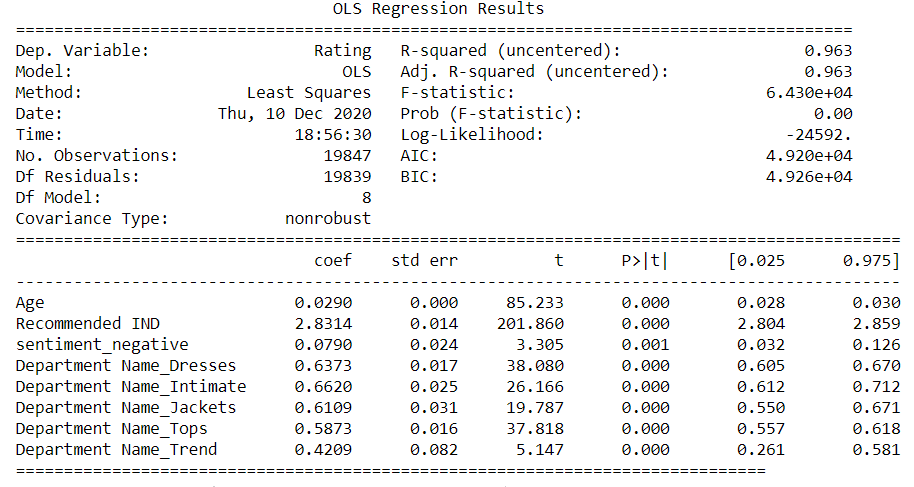


Figure 2.3

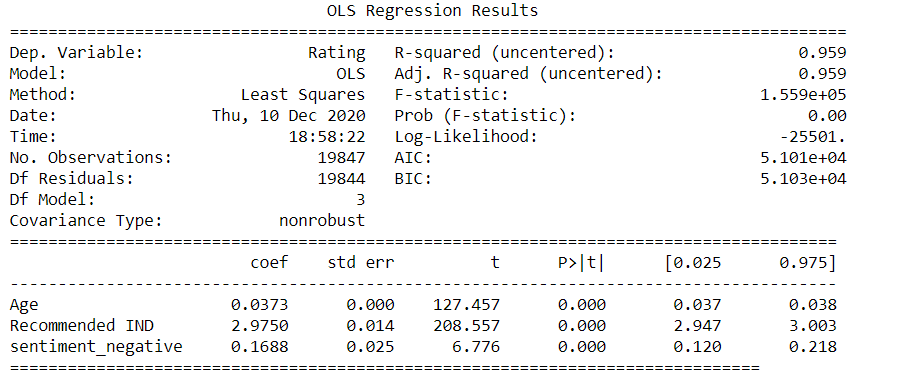


Figure 2.4

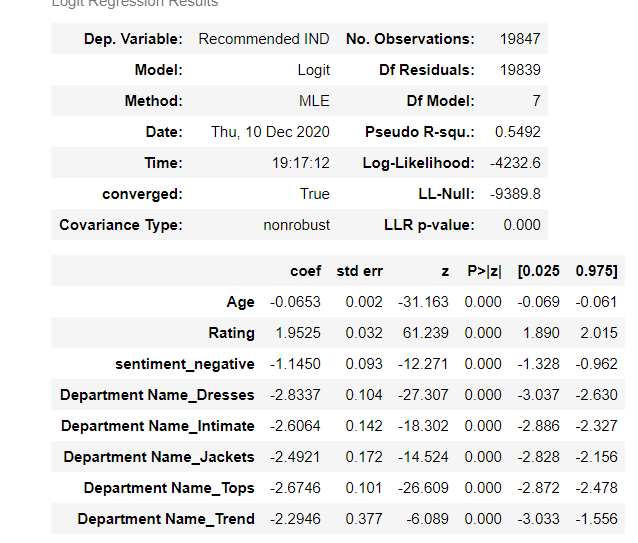


Figure 2.5

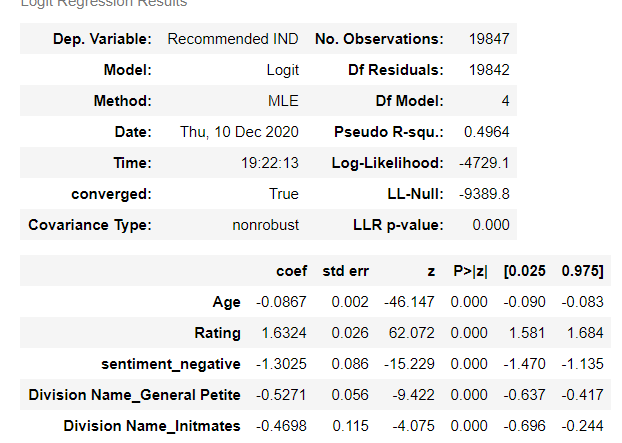


Figure 2.6