Customer Satisfaction Prediction Project

Created by: Divya Krishnakumar

Type of project: Internship Project

Submitted to: Unified Mentor Pvt Ltd.

Date: 21/07/2025

Executive Summary

The Customer Satisfaction Prediction project aims to analyse and predict customer satisfaction resulting from resolution of various issues generated by different tech products over a certain period of time. The customer support dataset provided for the analysis and prediction of customer satisfaction contained the data about the tickets raised on the technical issues of the products purchased in the past. The dataset consisted of 8469 rows and 17 columns. Some of the columns were Ticket ID, Ticket Type, Product Purchased, Date of purchase, Customer Name, Customer Email, Customer Age, Customer Gender, Ticket Subject, Ticket Description, Ticket Status, First Response Time, Customer Satisfaction Rating. The Customer Satisfaction Rating ranges for closed tickets ranges from 1 to 5.

The methodology followed here involves the following steps such as Data Collection, Data wrangling, Feature Engineering, Data Analysis and Visualisation, Modelling, Evaluation. The required libraries are imported. The data is loaded to a data frame from the customer_support_tickets.csv file. The descriptive statistics is performed on data. The categorical columns are label encoded. Then the data set is checked for any null values. The null values in the categorical columns are replaced with constants and null values in the numerical columns are replaced with 0. The columns are renamed to relevant names. The datetime columns which were in object data type were converted to datetime types. The data is analysed with SQL and insights such as names and number of customers who are most satisfied with the resolution, number of different priority issues, number of priority issues raised from different channels and so on. The data is analysed and visualised with Python libraries and insights are displayed. The insights were customer segmentation by ticket types, ticket types of various priorities, the top 10 products that raises the issues most and so on. For each insight visualisation is displayed using pie chart, count plot, hist plot etc. Correlation analysis is done to find out correlation among features and the correlation among features is displayed using a bar plot.

Model is developed using RandomForestClassifier machine learning algorithm, stratifiedKFold cross validation is done and hyperparameters tuned using GridSearchCV. Predicted the unseen data and a classification report is generated which displayed the metrics precision, recall and f1 score, accuracy. The confusion matrix is plotted for multi class prediction.

Introduction

The aim of the project is to predict Customer Satisfaction using the given historical data. This involves analyse the cutomer_support_ticket dataset to find the factors that influence Customer Satisfaction and build a predictive model.

The key questions are:

- Which ticket type is most frequently raised?
- What are the top 10 common issues?
- What is the relationship between ticket type and priority?
- What is the duration between first response time and time to resolution?
- What is the channel that customers use most of the time to raise the issues?
- What is the gender that raises the ticket most?
- What is the age group that raises the ticket most?
- Which is the Customer Rating most rated by customers?

Methodology

- The data is collected from the site provided.
- The methodology consisted of the following steps - as Data Collection, Data wrangling, Feature Engineering, Data Analysis and Visualisation, Modelling, Evaluation.
- The collected data is cleaned, analysed, visualised and subjected to machine learning algorithms such as Random Forest Classifier, Logistic Regression and evaluated.
- Data Collection – Data is collected and loaded to a data frame. The columns of the data frame include Ticket ID, Ticket Type, Product Purchased, Date of purchase, Customer Name, Customer Email, Customer Age, Customer Gender, Ticket Subject, Ticket Description, Ticket Status, First Response Time, Customer Satisfaction Rating.

Data Wrangling

Data Wrangling – The columns are renamed to relevant names. The data set is checked for null values. The null values were there in the columns Resolution, First Response Time, Time To Resolution and Customer Satisfaction Rating columns. The null values were present in these columns because of resolution had not taken place and thus no Customer Satisfaction Rating or resolution has taken place but customer has not responded for these tickets which were not closed. To make the analysis and visualisations better the null values in the columns Resolution, First Response Time, Time To Resolution the null values were substituted with 'Unknown' and in the column Customer Satisfaction Rating the null values were substituted with 0.0 which shows the rating not available. The First Response Time, Time To Resolution columns were converted from object type to datetime type.

The data frame thus obtained was as follows:-

	Ticket_ID	Customer_Name	Email	Age	Gender	Product_Purchased	DateOfPurchase	Ticket_Type	Ticket_Subject	Description	Customer_Satisfaction_Rating
0	1	Marisa Obrien	carrollallison@example.com	32	Other	GoPro Hero	2021-03-22	Technical issue	Product setup	I'm having an issue with the {product_purchase...	0.0
1	2	Jessica Rios	clarkeashley@example.com	42	Female	LG Smart TV	2021-05-22	Technical issue	Peripheral compatibility	I'm having an issue with the {product_purchase...	0.0
2	3	Christopher Robbins	gonzalestracy@example.com	48	Other	Dell XPS	2020-07-14	Technical issue	Network problem	I'm facing a problem with my {product_purchase...	3.0
3	4	Christina Dillon	bradleyolson@example.org	27	Female	Microsoft Office	2020-11-13	Billing inquiry	Account access	I'm having an issue with the {product_purchase...	3.0
4	5	Alexander Carroll	bradleymark@example.com	67	Female	Autodesk AutoCAD	2020-02-04	Billing inquiry	Data loss	I'm having an issue with the {product_purchase...	1.0

Feature Engineering

- The categorical features are label encoded to numerical to enable them for machine learning.
- From FirstResponseDate and FirstResponseTime Date and Time are extracted to two separate columns. From ResolutionTime Data and Time are extracted to two separate columns.
- A new feature Time_Elapsed_to_respond_seconds is created which gives the time elapsed to resolve the issue in seconds

Data Analysis with SQL Results

Ticket_ID and Customer names of most satisfied customers

```
%sql select Ticket_ID,Customer_Name from Customer_Satisfaction where Customer_Satisfaction_Rating=5.0
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Ticket_ID	Customer_Name
20	Jeffrey Robertson
29	Christine Wang
34	Timothy Lyons
59	Kimberly Mack
67	John Robertson
78	Alfred Ortiz
90	Lisa Hill
96	Linda Campbell
99	Nichole Huang
117	Sabrina Weber

Data Analysis with SQL Results

Number of customers who are most satisfied with the resolution

```
%sql select count(Ticket_ID) from Customer_Satisfaction where Customer_Satisfaction_Rating=5.0
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

count(Ticket_ID)
544

Number of different priority of issues

```
%sql select Priority,count(Priority) as 'Priority_Count' from Customer_Satisfaction group by Priority
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Priority	Priority_Count
Critical	2129
High	2085
Low	2063
Medium	2192

Data Analysis with SQL Results

Number of priority issues raised from channel Phone

```
%sql select Priority,count(Priority) as 'Priority_Count' from Customer_Satisfaction group by Priority having channel='Phone'
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Priority	Priority_Count
High	2085

The ticket_id, customer name,age of customer whose age is maximum

```
%sql select Ticket_ID,Customer_Name,Ticket_Subject,Age from Customer_Satisfaction where Age=(select max(Age) from Customer_Satisf
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Ticket_ID	Customer_Name	Ticket_Subject	Age
191	Regina Castillo	Product setup	70
221	Jacqueline Weaver	Installation support	70
323	Cindy Hale	Battery life	70
351	Ryan Murillo	Product compatibility	70
485	Carrie Wise	Data loss	70
498	Darrell Cook	Software bug	70
571	Levi Valencia	Refund request	70
623	Juan Hayes	Payment issue	70
713	Dawn Jones	Product recommendation	70
722	Beth Watson	Product setup	70

Data Analysis with SQL Results

The channel type of customers whose age is minimum

```
%sql select Ticket_ID,Customer_Name,Channel,Age from Customer_Satisfaction where Age=(select min(Age) from Customer_Satisfactio
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Ticket_ID	Customer_Name	Channel	Age
16	Elizabeth Foley	Social media	18
77	Matthew Scott	Phone	18
97	Charles Simpson	Social media	18
102	Danielle Rogers	Phone	18
118	Glenda Lopez	Phone	18
194	Tiffany Wilson	Chat	18
217	Angela Thompson	Chat	18
236	Ruth Fritz	Email	18
314	Joe Collins	Phone	18
318	Tamara Olson	Chat	18

The number of Channel type whose age is minimum

```
%sql select Channel,count(Channel) from Customer_Satisfaction where Age=(select min(Age) from Customer_Satisfaction)
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Channel	count(Channel)
Social media	163

Data Analysis with SQL Results

Ticket subjects of different products purchased

```
%sql select Product_Purchased,Ticket_Subject from Customer_Satisfaction group by Ticket_Subject
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

Product_Purchased	Ticket_Subject
Microsoft Office	Account access
Philips Hue Lights	Battery life
Lenovo ThinkPad	Cancellation request
Autodesk AutoCAD	Data loss
Xbox	Delivery problem
Amazon Kindle	Display issue
Nintendo Switch Pro Controller	Hardware issue
Fitbit Versa Smartwatch	Installation support
Dell XPS	Network problem
Microsoft Office	Payment issue
LG Smart TV	Peripheral compatibility
GoPro Action Camera	Product compatibility
GoPro Action Camera	Product recommendation
GoPro Hero	Product setup
Microsoft Surface	Refund request

Data Analysis with SQL Results

Ticket subjects whose proirity is critical or high

37]:
```
%sql select Ticket_Subject,Priority from Customer_Satisfaction group by Ticket_Subject  having Priority in (select Priority fro
```

```
 * sqlite:///CustomerSatisfactionDB
Done.
```

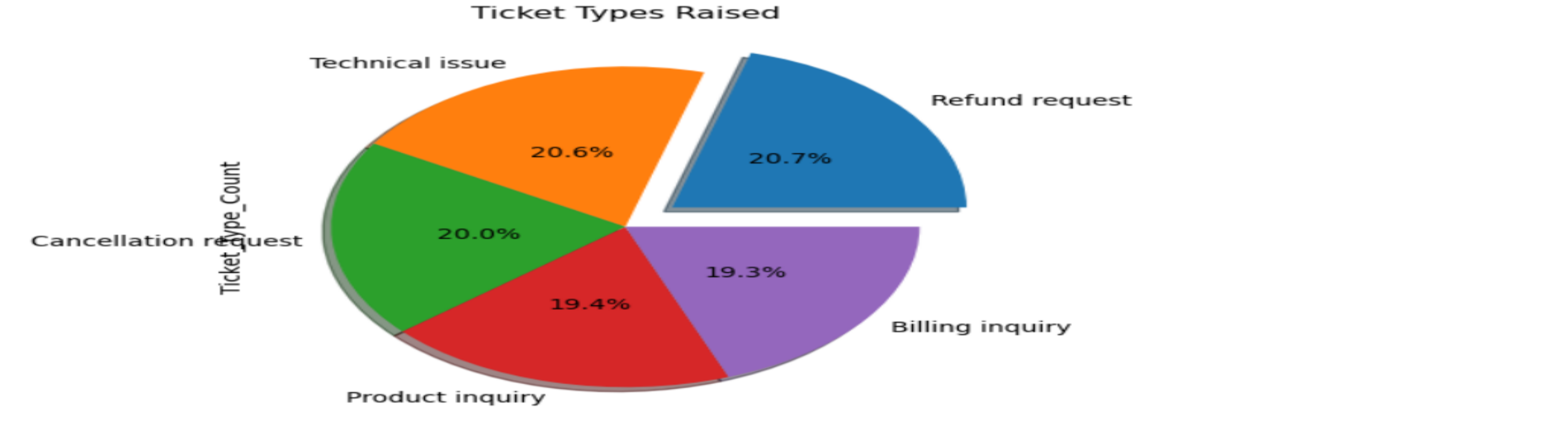
37]:

Ticket_Subject	Priority
Battery life	Critical
Display issue	Critical
Peripheral compatibility	Critical
Product setup	Critical
Refund request	Critical
Product recommendation	High
Software bug	High

Exploratory Data Analysis(EDA) and visualisation results

Which ticket type is most frequently raised?-Customer Segmentation by ticket types

Ticket_Type	Ticket_Type_Count
Refund request	1752
Technical issue	1747
Cancellation request	1695
Product inquiry	1641
Billing inquiry	1634



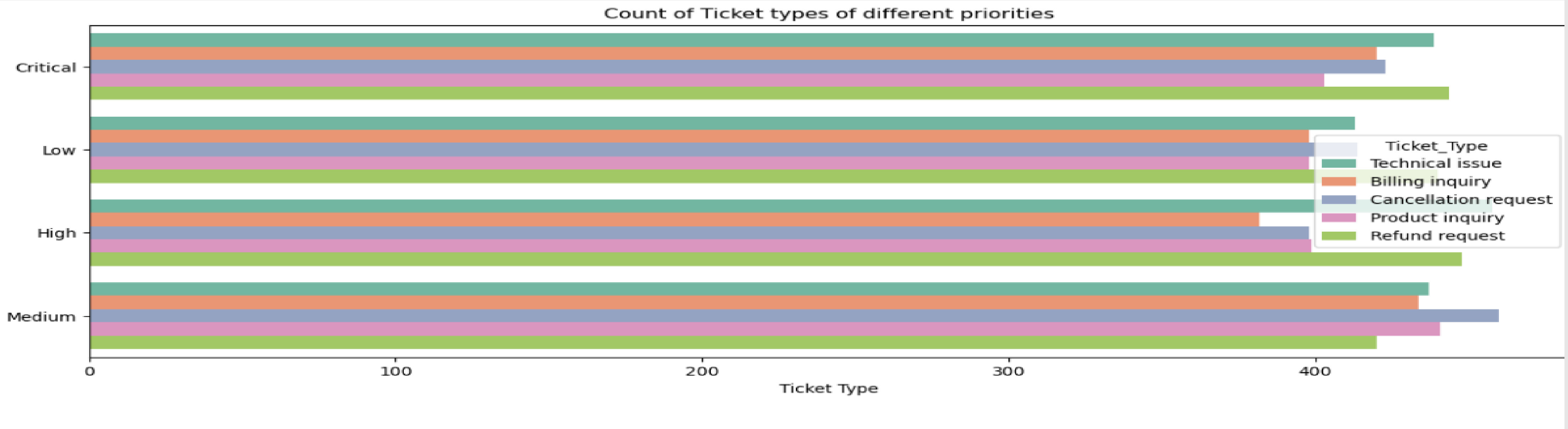
Exploratory Data Analysis(EDA) and visualisation results

Relationship between ticket type and priority

Ticket_Type	Priority	Priority_Count
Cancellation request	Medium	460
Technical issue	High	458
Refund request	High	448
	Critical	444
Product inquiry	Medium	441
Refund request	Low	440
Technical issue	Critical	439
	Medium	437
Billing inquiry	Medium	434
Cancellation request	Critical	423
Refund request	Medium	420
Billing inquiry	Critical	420
Cancellation request	Low	414
Technical issue	Low	413
Product inquiry	Critical	403
	High	399
Cancellation request	High	398
Billing inquiry	Low	398
Product inquiry	Low	398
Billing inquiry	High	382

Exploratory Data Analysis(EDA) and visualisation results

A count plot is plotted to visualize count of ticket types of different priorities



Exploratory Data Analysis(EDA) and visualisation results

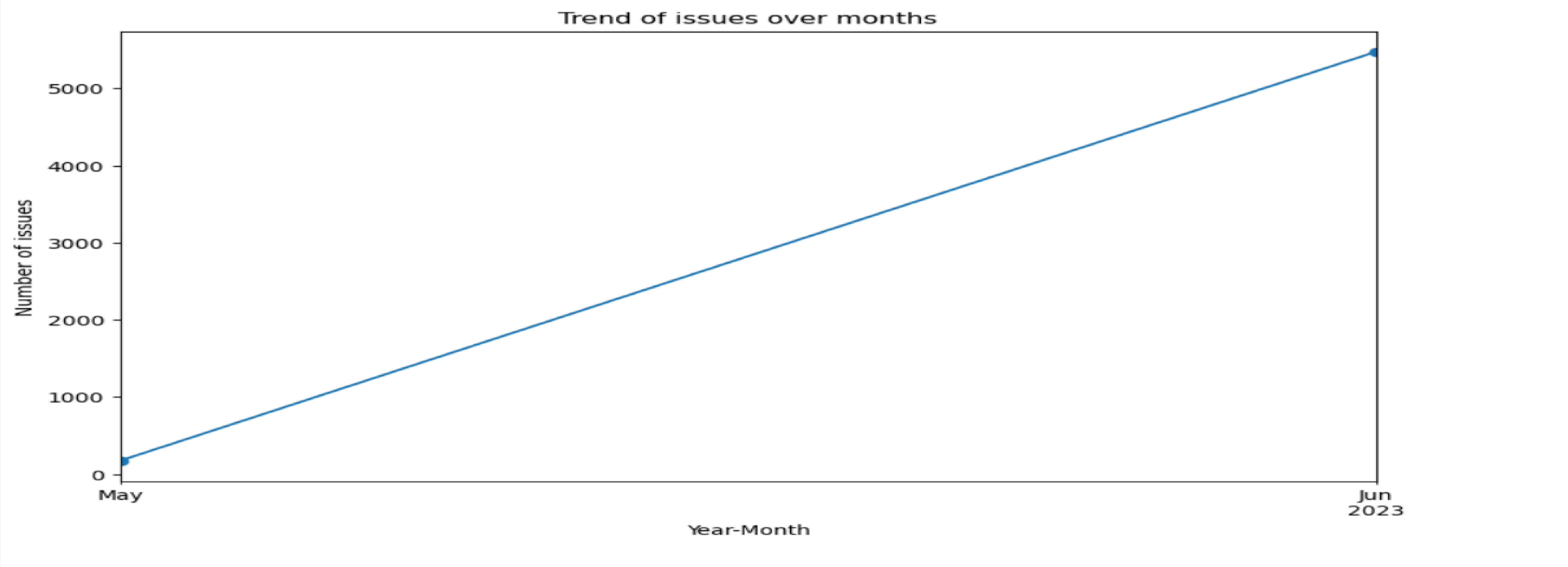
Top 10 common issues

```
df_grpcomissue=df['Ticket_Subject'].value_counts()
df_grpcomissue
```

```
Ticket_Subject
Refund request              576
Software bug                574
Product compatibility       567
Delivery problem            561
Hardware issue              547
Battery life                542
Network problem             539
Installation support        530
Product setup               529
Payment issue               526
Product recommendation      517
Account access              509
Peripheral compatibility    496
Data loss                   491
Cancellation request        487
Display issue               478
Name: count, dtype: int64
```

Exploratory Data Analysis(EDA) and visualisation results

Trends of issues over the years



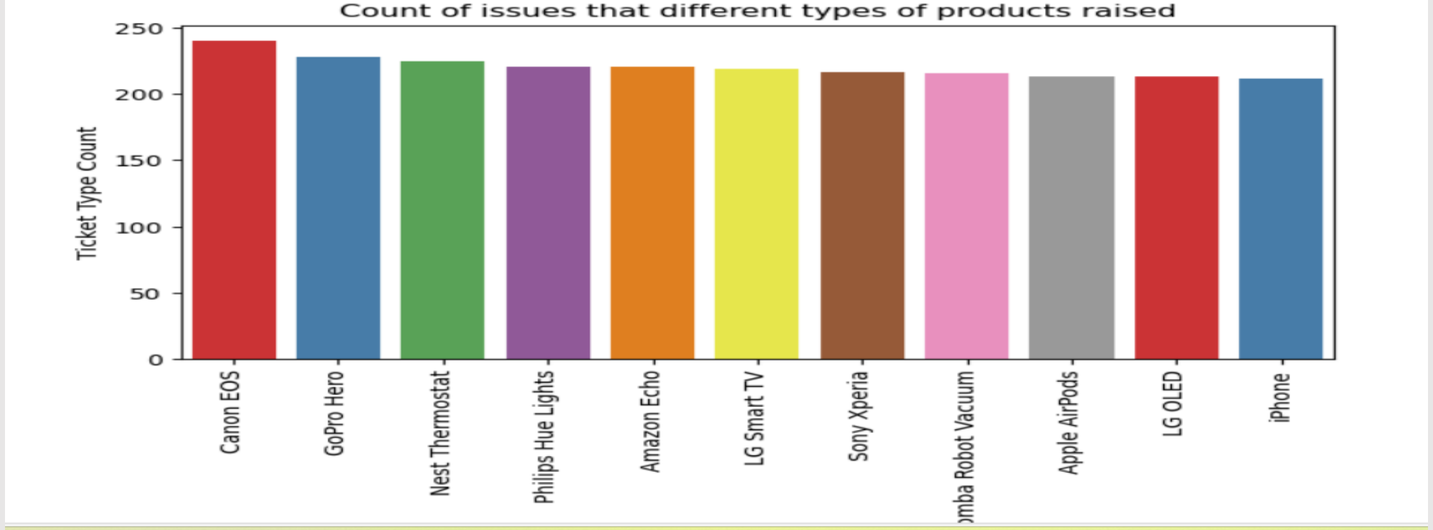
Exploratory Data Analysis(EDA) and visualisation results

Which 10 products raises issues most?

Product_Purchased	Ticket_Type_Count
Canon EOS	240
GoPro Hero	228
Nest Thermostat	225
Philips Hue Lights	221
Amazon Echo	221
LG Smart TV	219
Sony Xperia	217
Roomba Robot Vacuum	216
Apple AirPods	213
LG OLED	213

Exploratory Data Analysis(EDA) and visualisation results

A bar plot is plotted to visualize the count of issues that different types of products raised.

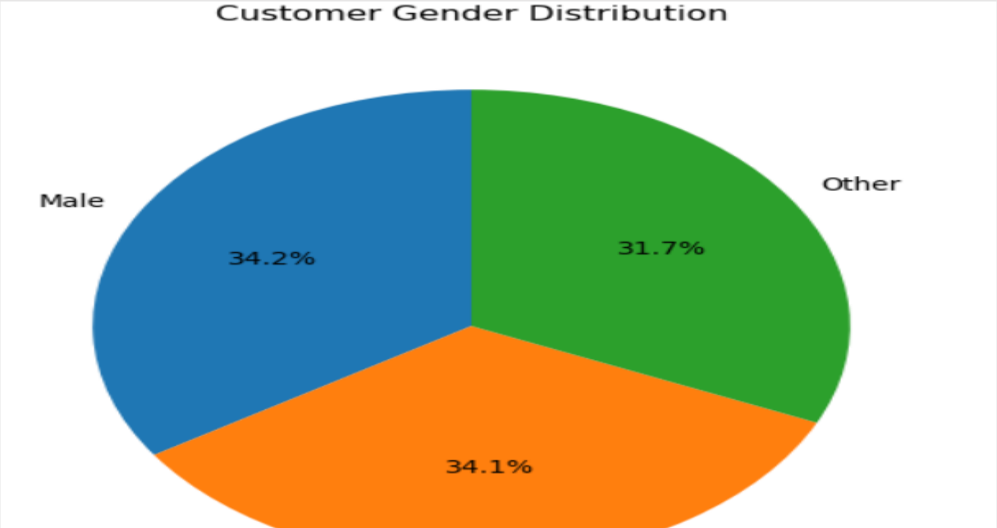


Exploratory Data Analysis(EDA) and visualisation results

The gender that raises ticket most

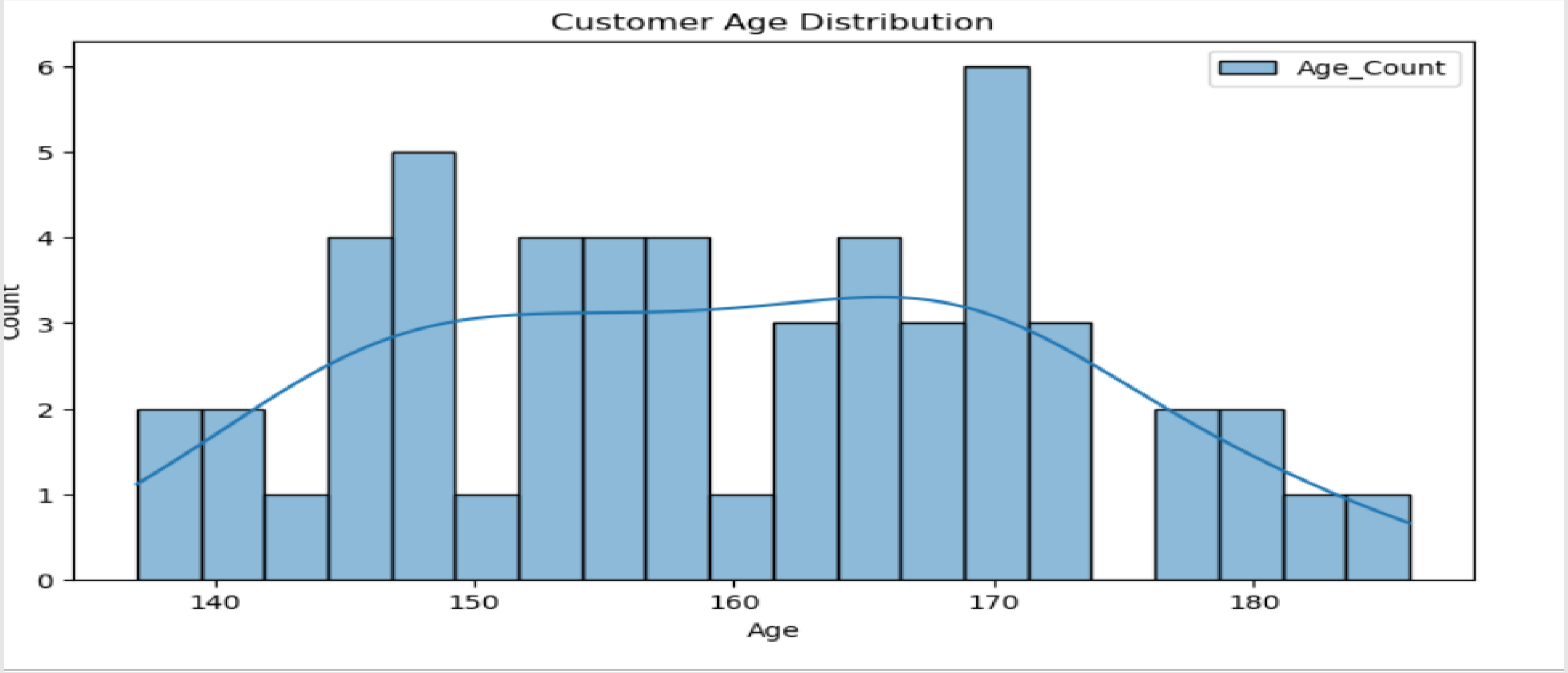
```
Gender
Female     2887
Male       2896
Other      2686
Name: Gender, dtype: int64
```

A pie chart is plotted to depict gender distribution.



Exploratory Data Analysis(EDA) and visualisation results

The age group that raises the ticket most



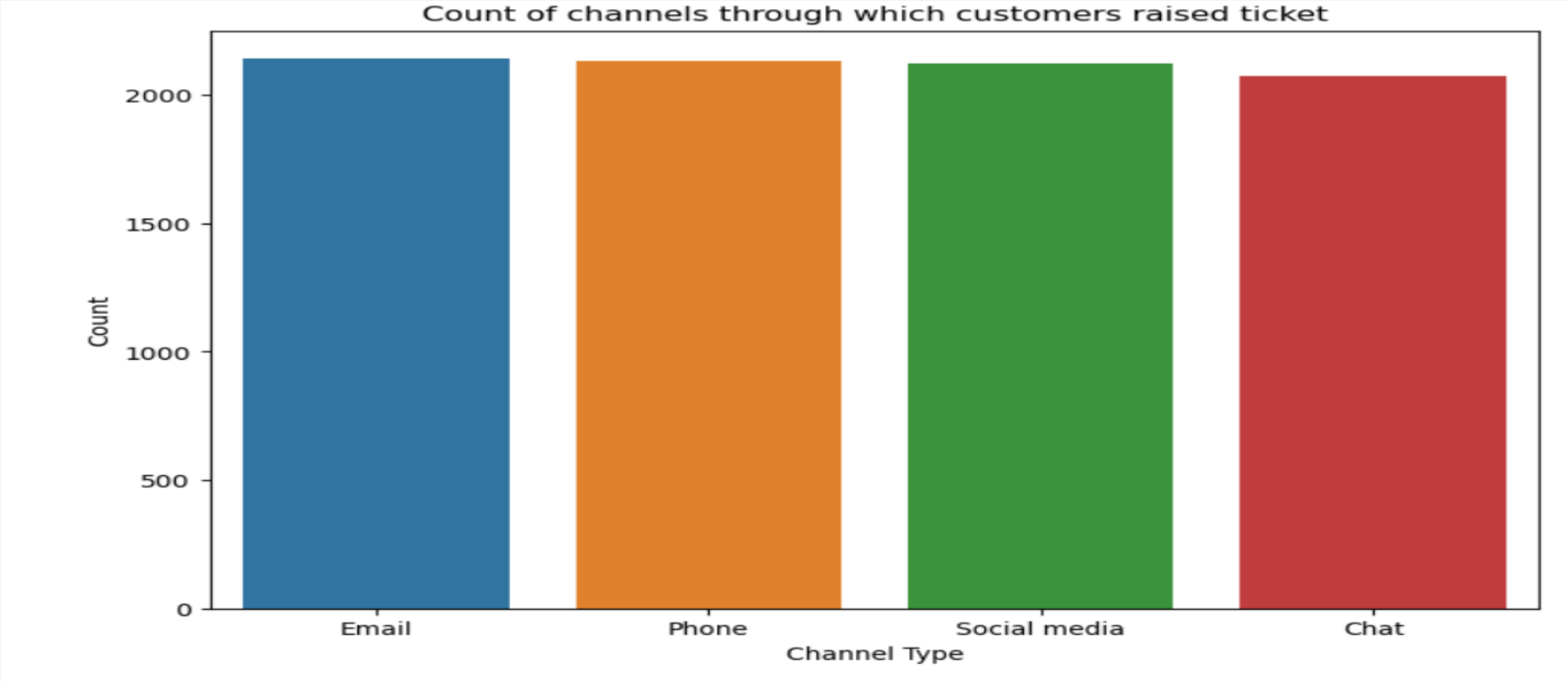
Exploratory Data Analysis(EDA) and visualisation results

The channel that customers use to raise the ticket most of the time

Channel	Channel_Count
Email	2143
Phone	2132
Social media	2121
Chat	2073

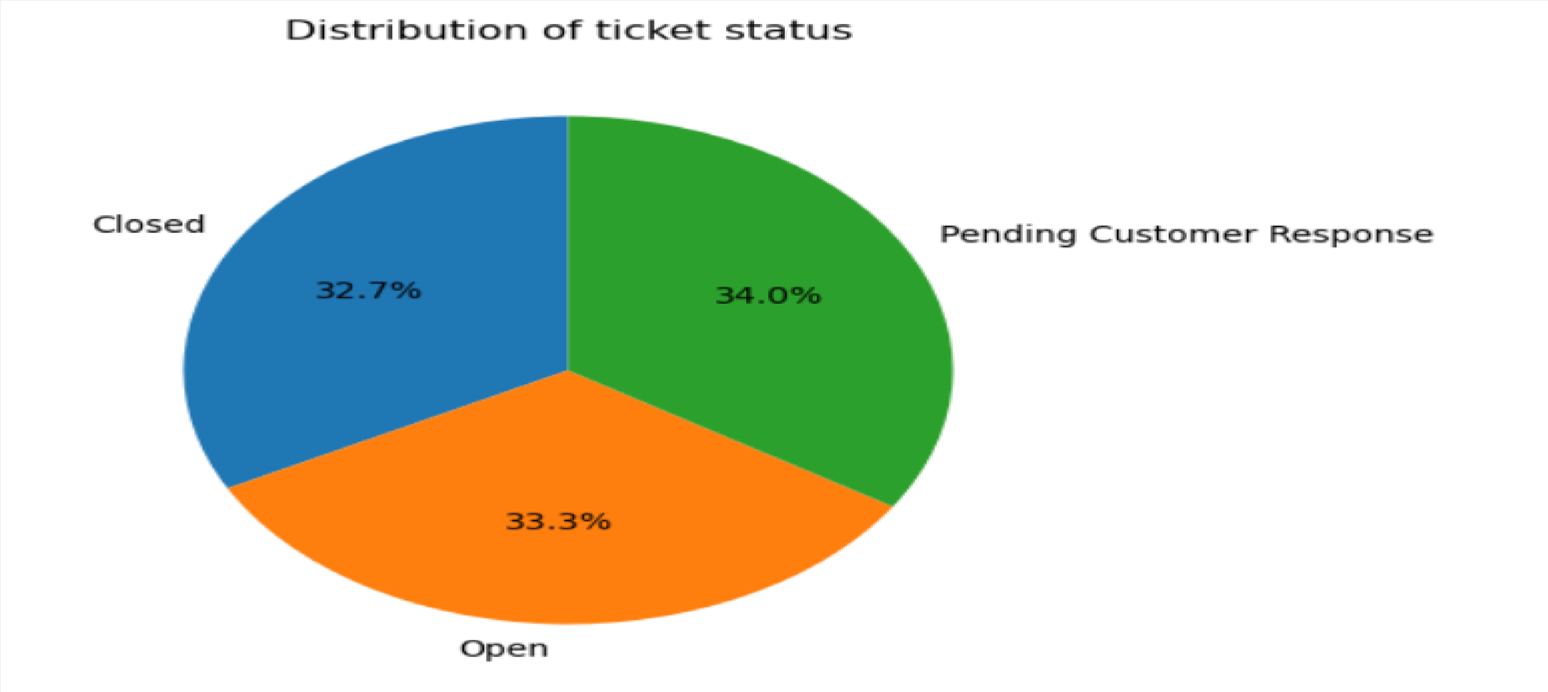
Exploratory Data Analysis(EDA) and visualisation results

A count plot is plotted to depict the channel that customers use to raise the ticket most of the time.



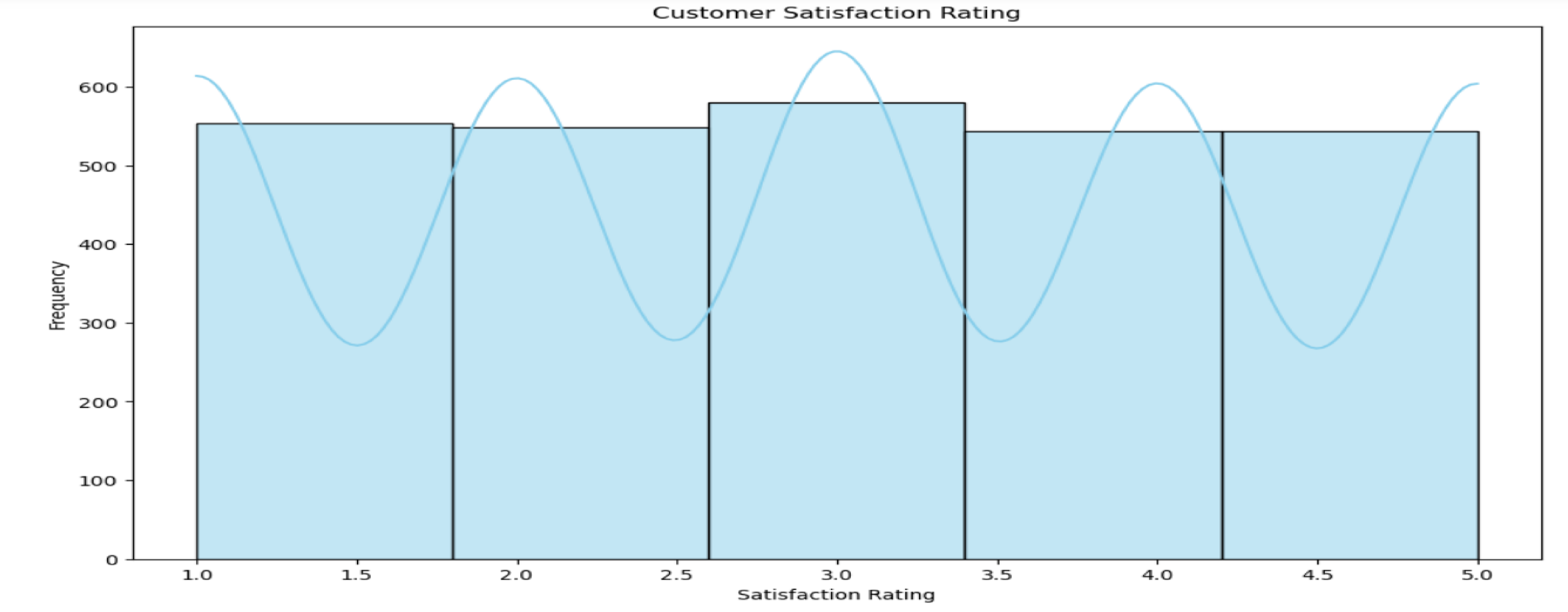
Exploratory Data Analysis(EDA) and visualisation results

Distribution of ticket status



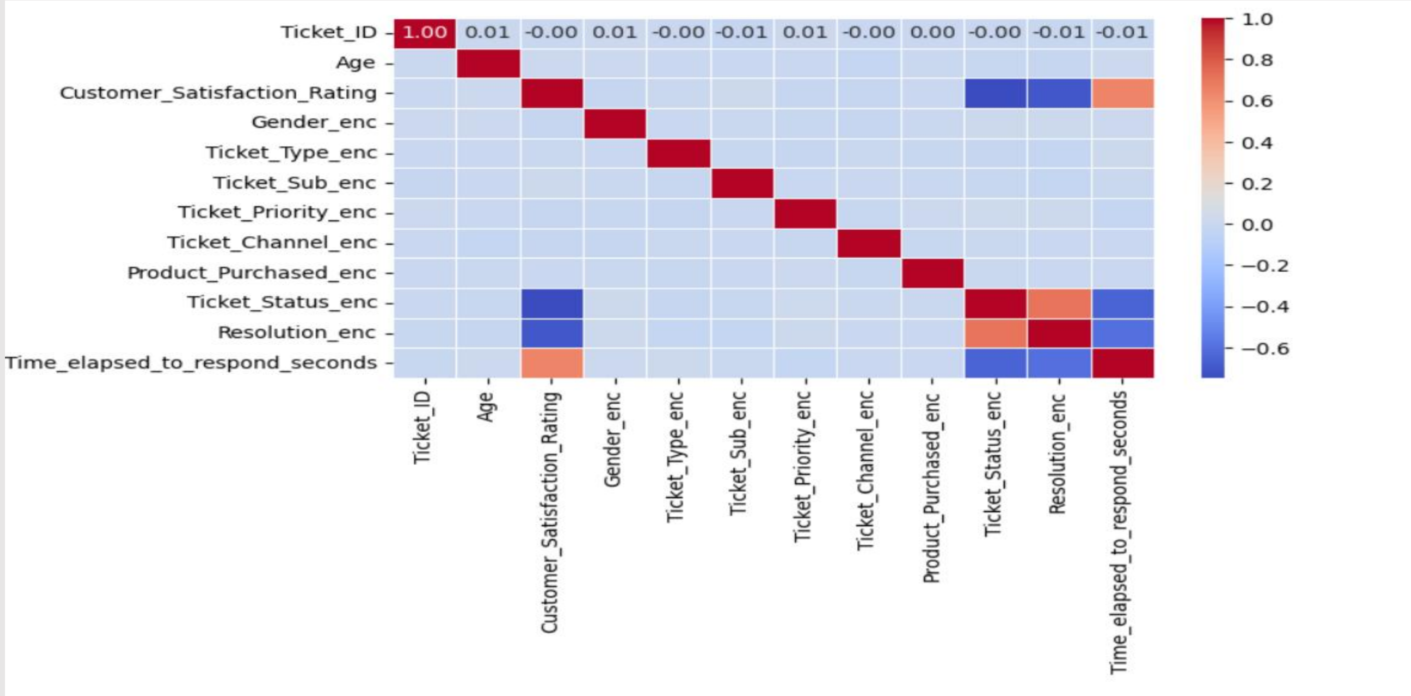
Exploratory Data Analysis(EDA) and visualisation results

Customer Satisfaction Rating



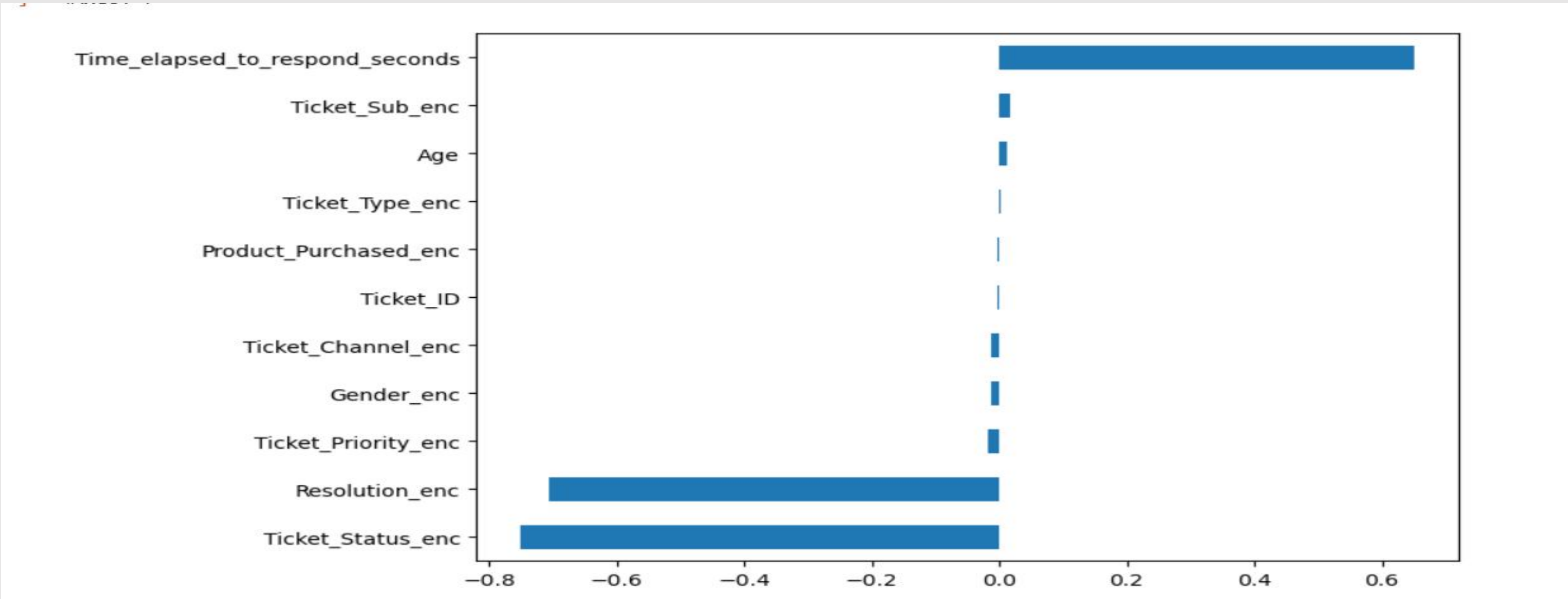
Exploratory Data Analysis(EDA) and visualisation results

Correlation analysis is done and a heat map plotted to identify correlation between variables.



Exploratory Data Analysis(EDA) and visualisation results

A horizontal bar graph is plotted to visualise positive and negative correlation. Time elapsed to respond, ticket subject, age are positively correlated to customer satisfaction.



Modelling and predictive analysis

The data is subjected to machine learning algorithms such as RandomForestClassifier and Logistic Regression for multi class prediction. These models are fitted on training data and results are predicted.

RandomForestClassifier:- RandomForestClassifier object is created, hyperparameters are tuned with GridSearchCV and stratifiedKFold cross validation is done. The model thus is fitted with training data and after that the results are predicted.

As a part of evaluation a classification report is generated. A confusion matrix display is visualised.

Logistic Regression(One Vs All) for multi class classification- Logistic Regression one vs all model is created and fitted with training data and the results are predicted on unseen data.

As a part of evaluation a classification report is generated. A confusion matrix display is visualised.

Modelling classification models-Results

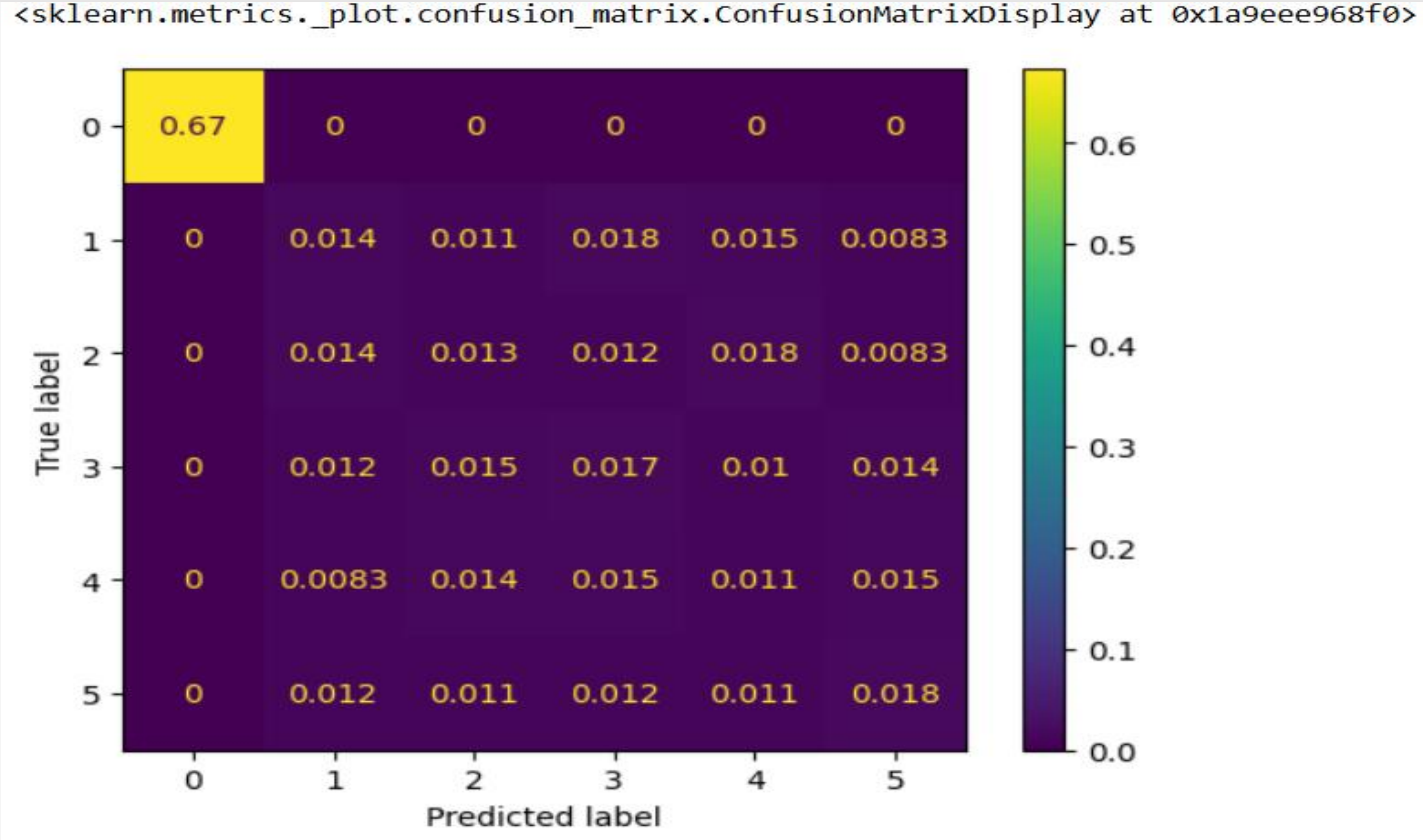
Random Forest Classifier- Accuracy score is 74.55%

Classification Report

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1140
1.0	0.23	0.21	0.22	111
2.0	0.20	0.20	0.20	110
3.0	0.23	0.25	0.24	116
4.0	0.17	0.18	0.17	108
5.0	0.28	0.28	0.28	109
accuracy			0.75	1694
macro avg	0.35	0.35	0.35	1694
weighted avg	0.75	0.75	0.75	1694

Modelling classification models-Results

Random Forest Classifier- Normalised Confusion Matrix Display



Modelling classification models-Results

Logistic Regression – Accuracy score of 74%

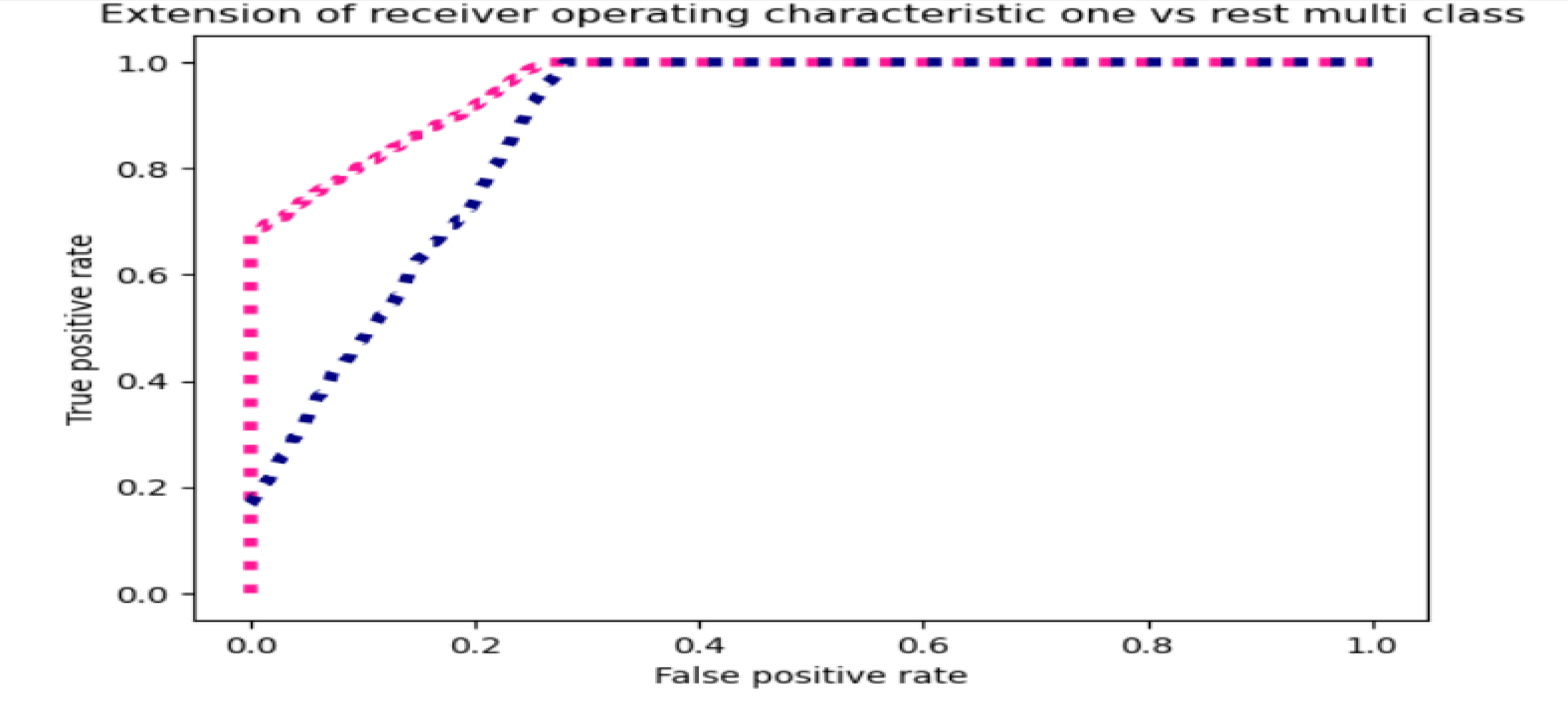
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1140
1.0	0.21	0.20	0.21	111
2.0	0.20	0.13	0.16	110
3.0	0.23	0.33	0.27	116
4.0	0.19	0.22	0.21	108
5.0	0.23	0.19	0.21	109
accuracy			0.74	1694
macro avg	0.34	0.34	0.34	1694
weighted avg	0.74	0.74	0.74	1694

Modelling classification models-Results

Logistic Regression –

Micro averaged One-vs-Rest ROC AUC score : 0.958

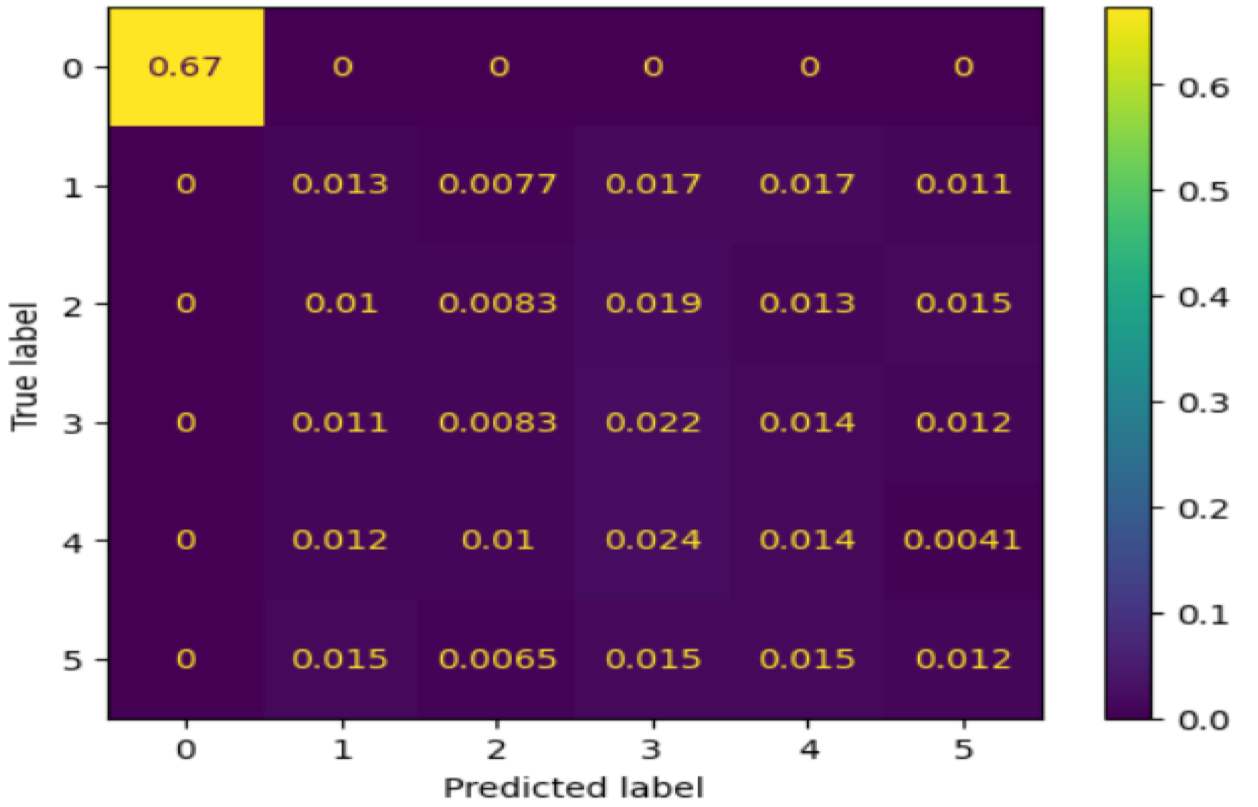
Macro averaged One-vs-Rest ROC AUC score : 0.885



Modelling classification models-Results

Logistic Regression- Normalised Confusion Matrix Display

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1a9f1c66020>



Findings and Implications

- The people of the ages 44,34,48,52 raised tickets most.
- The customer satisfaction rating of 3.0 was given by most of the customers about 580 of them.
- The customers who use Email as the channel for raising tickets was high about 2143 customers use Email as channel to raise the tickets.
- The product Canon EOS raised highest number of issues of about 240.
- Number of cancellation requests with medium priority is greatest with 460 of them. Number of billing inquiry with high priority is lowest with 382 of them.
- Technical issue raised was 20.6%, Refund request raised was 20.7%, and cancellation request raised was 20%.
- The issues such as battery life, display issue, peripheral compatibility, product set up, and refund request were of critical priority.
- The number of critical priority issues was maximum with 2129.
- The number of customers who gave a customer satisfaction rating of 5.0 was 544.

Conclusion

The different customer satisfaction ratings given by customers were almost equally distributed. The customer satisfaction rating was average for maximum number of issues. The issues had an increasing trend over the period. The customer satisfaction was high for issues that were resolved in least time. It also depends on ticket subject of critical priority. If ticket subjects of critical priority is resolved in least time then customer satisfaction will be high. The age of the customer also influences customer satisfaction since most of the issues were raised by older age groups.

Future implications and suggestions to improve customer satisfaction

- To improve customer satisfaction the issues need to be resolved on priority wise and in least time.
- The timeline of first response and resolution of different priority issues has to be fixed within a range.