RATINGS PEDICTION

SUBMITTED BY

JUHI MISHRA

ACKNOWLEDGEMENT

To complete this project taken help from many sources. Summarised the list below for the same :
Websites from where taken help mentioned below:
1. Analytics Vidya
2. GeeksforGeeks
3. Towards data science
4. Datacamp
5. Medium.com
6. Youtube
7. Even taken help from Ms.Khosboo Garg my SME and Mr.Subham Yadav regarding all my doubts which I was not able to solve via these websites.

Above all websites used to clear my doubts regarding NLP preprocessing.

To scrap rating and review on which I was suppose to build model and predict rating accordingly taken help from :
1. Flip Robo - My SME Ms.Khosboo Garg and Mr. Subham Helped me a lot in order to understand how to scrap and exactly what data I need to scrap.
2. GeeksforGeeks and
3. Youtube

I want to thanks for all the sources from where I got help and able to finish this project.

INTRODUCTION

BUSINESS PROBLEM FRAMING :

E-commerce business emerged to be a larger scale business all over the world. Shopping become easier and convenient as anyone can shop online from anywhere. No need to preplan to go outside to shop. But as we know that customer satisfaction in any of the business is mandatory. Online shopping completely depends upon trust because what they show on site if delivered the

same customer will be 100% satisfied but if not happened so that turns to dissatisfaction and in result business will be impacted negative. Now to evaluate the same every online business do have option of customer can give their feedback in the form of review and also can rate for the same. Review and rating is one of the important aspect through which a particular site can have look on how many customers are satisfied or dissatisfied and what is the reason behind the same. This will not only helpful in improving their business but also will able to look after the loophole where they are lacking and can retain there customer accordingly. Loosing customer is a loss for any of the business so to curb the same every e-commerce business person should know the reason and try to resolve the same also should make sure that this should not be repeated later.

CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM :

This project is text based for which we need to solve using NLP and NLP is one of the NLTK tool which is used to solve basically natural language processing issues. NLP is the branch of artificial intelligence concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. Here we are suppose to predict rating based on review given by customers on any of the particular product. To understand text data we need to do the preprocessing using NLP. After preprocessing done need to use ML algorithm so that we can predict the rating on any particular review.

REVIEW OF LITERATURE :

This project is text based for which we need to solve using NLP and NLP is one of the NLTK tool which is used to solve basically natural language processing issues. NLP is the branch of artificial . Here we are working for one of our client which is Flipkart who has a website where people write different reviews for technical products. Here new feature being added where customer can not only write their review but also can add star for product how they are and how much star could be given to that particular product. Total 5 ratings are available which is 1 star, 2 stars, 3 stars, 4 stars and 5 stars. Here we have to predict rating based on reviews which was written in past and they don’t have rating available. So

here as we do have two columns which is rating and review , so this problem statement is text based where a computer have to understand the text and need to predict rating accordingly. As a data scientist we need to find the solution for the same. For ant text based problem statement NLP works best as it makes computer understand the natural language and can then accurately extract information and insights contained in the document as well as categorise and organise the documents themselves. Here using NLP will preprocess the data and will later choose ML algorithm for making prediction based on previous dataset. This is a classification based problem statement as in target we do have discrete values which are ratings. For which we need to choose classification algorithms and build model accordingly.

MOTIVATION OF THE PROBLEM UNDERTAKEN :

Any of the business house do want to have a satisfied customer which will not only help them to grow in future but also will give them an opportunity to grab more and more customer via reference. Key to get success in any of the business is having a satisfied and regular customer base. Why I started this project is because of understanding the customer reason to be dissatisfied for any of the particular product. As a data scientist my job role is to provide solution to any of the business house as per their requirement. And here my client need solution of prediction of rating based on review given by customer. Also working on different projects gives more and more experience of different solving approaches. Here I learned how to solve using NLP which was completely new topic for me. Always learning new techniques and different approaches to solve different projects makes me more confident and satisfied. As a data scientist we need to explore more and more approaches to solve any of the problem which can help our client to get solution from us.

ANALYTICAL PROBLEM FRAMING

MATHEMATICAL / ANALYTICAL MODELING OF THE PROBLEM :

Rating prediction project is one of the classification problem statement where target column do have discrete values which is rating from 1 - 5 depending upon reviews given. Before solving any problem statement we need to first analyse the same and what are the approached required need to list down all. Before preprocessing checked whether there are any missing values in any of the rows or column. There are no missing values in any of the column or rows. Datatype of both the column which is review and rating is object datatype. Then whether the problem is balanced or imbalanced using value.counts because this is classification based problem statement and for classification problem statement we have to check whether the data is balanced or not. Here the data is a balanced one so no need to preprocess for the same. All the rating are equally distributed which is 3,456 each for rating 1,2,3,4,5 and 6. This is text based problem statement where I need to use NLP approach. NLP stand for natural language preprocessing which generally understand the text or speech type of data and try to solve the same. Mathematically here I used TFIDF vectoriser to convert all string form to numerical one as computer doesn't understand string datatype . TF-IDF stands for term frequency and inverse document frequency. Tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contains the word, which helps to adjust for the fact that some word appear more frequently in general. Formula for Tf-idf mentioned below :

Tf-idf

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

tf-idf(t, d) = tf(t, d) * idf(t)

Formula and mathematical approach we seen above for calculation of tf-idf where tf is used in connection with information retrieval and shows how frequently any term or word occurs in a document. While idf used to calculate the weight of rare words across all documents in the corpus. There are lot many approaches to convert string to vector which are

count vectoriser, tf-idf, word to vec(Word embedding technique). But here I used tf-idf because of more reliable result I can get using this approach here.

DATA SOURCES AND THEIR FORMATS :

As a data source I just received two documents from my internship company where mentioned all the required information like how to get the data , what are the approaches I need to follow, and what are the sources I can use to have data for further process or model building. Here I used Flipkart site to scrap rating and reviews of customers for different technical products and prepared excel for the same. I was suppose to collect 4000 rows or rating each for all the rating which were 1,2,3,4 and 5 to have balanced dataset which would help us to make better model. Total I collected 17,275 rows and for each rating 3,456 . Using pandas library imported the data for further process. Snapshot for the same mentioned below :

	Rating	Review
0	1	Useless product
1	1	Don't waste your money
2	1	Very poor
3	1	Useless product
4	1	Hated it!

DATA PREPROCESSING DONE :

To preprocess the data used NLP approach (Natural Language Preprocessing) . As this problem is text based problem statement and for text dataset NLP approaches are best . By using NLP text classifiers can automatically analyse text and then assign a set of pre - defined tags or categories based on its content. Natural language understanding helps machine read text or another input such as speech by stimulating the human ability to understand a natural language such as English, Hindi, or Japanese. Using NLTK (Natural Language Toolkit) library done preprocessing. Details of steps are mentioned below:

1. Converted string data to lower case such as ‘Like’ will be converted as ‘like’. It helps to get rid of noise.
2. Then did contraction to expansion which generally expand the word which is in contraction mode such as ‘don’t’ converted to ‘do not or does not’. Basically contraction are shorten form of words which we need to expand.
3. Remved punctuations, multiple spaces such as (“ ”) to (“ “) to avoid noise for the dataset or get rid of unhelpful parts of the data.
4. Removed stop words from the dataset as these words do not add much value to the data . These words are “ is, and, the, an” etc. By removing these words we can give more focus to important information in our dataset.
5. Tokenized sentence into words. Toeknization basically done to break sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokenization can be done into word or sentences or phrases depending upon which dataset we are working upon.
6. Then performed lemmatisation whose main job is to bring the word to base word. Like run, running, ran all will become run once lemmatisation is performed on the same. In other word we also call it as lemma which change the different form of a word into a single item called a lemma.

After preprocessing done scaling using standard scaler on all the feature dataset which is review here.

DATA INPUTS - LOGIC - OUTPUT RELATIONSHIPS :

Data input here is Review where all the customer gives their valuable feedback based on their experience. All the reviews scrapped from Flipkart for different technical products in an excel sheet. Review basically given by customer to show how much they are satisfied or dissatisfied in a descriptive form . In order to have more clarity company added ratings which a customer can use to show how much rating they can give to a particular product. Here total 5 ratings being given as an option to customer which is 1,2,3,4 and 5. Now along with review they can provide rating too which will be very helpful in filtering those product which is highly recommended and also for those which is not at all recommended. While review will help in understanding the reason behind dissatisfaction or not recommending the product later on. This will help a company to understand the reason and to improve the same so to gain more profit and make revenue.

HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED :

To build a good model we should have a cleaned dataset and if not then need to work on same so that will get best possible outcome. Data Processing and data cleaning are one of the vital job role of data scientist. In this problem .
Hardware tools which being used while coding
were :

1. CPU - This is one of the important part while performing machine learning task because most of the computation will be done in learning

environment which is most likely done on CPU.

2. Power Supply - This is also one of the important hardware while performing any machine learning task as if power supply will not be

there , will not able to perform the task.

3. RAM - While performing coding on machine learning minimum 8 GB

ram is required to perform the task.

Coming to software part the important libraries and packages required

was for machine learning algorithm :

1. Pandas - One of the important library which helps in importing the

data to perform data manipulation and analysis.

2. Numpy - This library is used to perform all the mathematical

operations required during coding

3. Seaborn and Matplotlib - Used for visualisation analysis. As we know that to understand any data in a more proper way and for better presentation, visualisation is one of the important technique used.

4. NLTK - This library is used to perform all NLP based problems. It helps the

computer to analyse, preprocess and understand the writer text.

MODEL/S DEVELOPMENT AND EVALUATION

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS) :

Data input here is Review where all the customer gives their valuable feedback based on their experience. All the reviews scrapped from Flipkart for different technical products in an excel sheet.. Here I need to predict rating based on reviews provided former. Which means using NLTK library I have to follow the NLP approach to solve this problem. Preprocessing done based on NLP approach here. Which means brought all the uppercase words to lower case to avoid noise. Then converted the words from contraction to expansion means shorten words have been expanded. Then removed multiple spaces and removed stop words. Stopwords are here not required because they do not add much value to our dataset. And by removing these we can have more important words much visible. Then lemmatised alll the words which means converting the words to their base form. Then converted all the strings to numerical form using tfidf approach as our machine algorithm doesn't understand the string datatype . After preprocessing done used classification based algorithms to build model. This problem statement is classification based because target column were having discrete value. Before preprocessing checked whether my data is imbalanced or not as for any classification problem statement we need to check this . If the data is imbalanced one then our model performance will be hampered. So to avoid such we need to treat them using sampling method means either we need to oversample the data or undersample. But here in this dataset no such treatment was required as every rating were equal in number.

Also there were no null values wither in any of the row or column.

TESTING OF IDENTIFIED APPROACHES (ALGORITHM) :

Data input here is Review where all the customer gives their valuable feedback based on their experience. All the reviews scrapped from Flipkart for different technical products in an excel sheet. Review . Based on problem statement selected 3 classified algorithms listed below :
1. Random Forest Classifier
2. SGD Classifier
3. LinearSVC

For evaluation selected metrics are :
1. Classification report
2. Accuracy Score

To train and test , hyper parameter tuning and for validation score import few of the algorithm from sklearn model selection . These are:
1. train_test_split
2. GridsearchCV
3. Cross Validation Score.

To scale the features used standard scaler to bring all the features to same unit with the help of which model will have better performance.

RUN AND EVALUATE SELECTED MODELS :

First classifier algorithm I used is RandomForestClassifier is an ensemble technique which consist of many decision trees and combines many classifiers to provide solutions to complex problems.This algorithm generally gives output based on the predictions of the many decision trees. Increasing the number of trees increases the precision of the outcome. It reduces the overfitting of datasets and increases precision. Below provided the snapshot of the

RandomForest Classifier algorithm performance :

```
rfc = RandomForestClassifier()
rfc.fit(x_train,y_train)
pred =rfc.predict(x_test)
acc = classification_report(y_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           1       0.99      0.99      0.99       874
           2       0.78      0.98      0.87       845
           3       1.00      0.82      0.90       873
           4       0.92      0.87      0.89       874
           5       0.95      0.93      0.94       853

    accuracy                           0.92      4319
   macro avg       0.93      0.92      0.92      4319
weighted avg       0.93      0.92      0.92      4319
```

```
print(cross_val_score(rfc,x,y,cv=5).mean())
```

```
0.9165267727930535
```

For rating 1 all the precision , recall and f1-score is 99% which means score for rating 1 is really good. While for Rating 2 precision score is too low which is 78% compare to recall which is 98% because of which f1 score shows 87%, as f1 score is the combination of both false positive and false negative where precision is having false positive and recall is having false negative. Formula of f1 score is 2*(precision * Recall)/Precison + Recall. Similarly for rating 3 we can see that false positive which is precision is having 100% while recall which basically have focus on false negative is 82% and the avg of both which is f1 score is 90%. In this problem statement recall score is much helpful compare to precision because here we need rating prediction and if false positive will be higher will not able to identify genuine rating of customers how many are satisfied and how many are dissatisfied. For rating 4 precision score is 92% while recall is 87% and avg of both is 89%. While for rating 5 there are not much difference between scores of precision recall and f1 score . As we can see that false positive(precision) score is 95% while false negative score (recall) is 93% . F1 score for the same is 94% which is multiplication of 2(p*r)/(p+r). Overall accuracy score we got as 92% . While cross validation score on total 5 foldings is 91.65% which is less than accuracy score .

Second used SGD Classifier algorithm.SGD classifier is a Linear classifier optimized by SGD. These are two different concepts. SGD is a optimisation method while Logistic regression or Linear support vector machine is a machine learning algorithm / model. Because of few downsides of Gradient Descent algorithm we use SGD. SGD randomly picks one data point from the whole data set at each iteration

to reduce the computation enormously. Below is the snapshot of classification report of SGD classifier :

```
sgd = SGDClassifier()
sgd.fit(x_train,y_train)
pred =sgd.predict(x_test)
acc = classification_report(y_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           1       0.99      0.99      0.99       874
           2       0.99      0.82      0.90       845
           3       0.79      0.98      0.87       873
           4       0.91      0.88      0.89       874
           5       0.97      0.93      0.95       853

    accuracy                           0.92      4319
   macro avg       0.93      0.92      0.92      4319
weighted avg       0.93      0.92      0.92      4319
```

From above classification report we came across each rating precision, recall, f1 and accuracy score. Here rating 1 is having all 99% means false positive, false negative and average of all which is f1 score are equal. Having false positive report high for this problem statement will not give us accurate report . Because if will get high false positive then most of the rating will be shown as positive result which will not be helpful in identifying the dissatisfied customer. Rating 2 precision score is 99% (false positive) and recall is 82% (false negative) . Overall combination of both which is f1 score is 90%. For rating 3 precision score is 79% while recall is pretty high which 98% means here the possibility to have false negative is high. And so the f1 score affected and showing including both false positive and negative report is 87%. For rating 4 precision score is 91% while recall score is 88% while f1 score is 89%. For rating 5 precision score is 97% while recall is 93% and f1 score is 95%. Over all we can say that rating 3 is having huge difference between precision and recall in which recall score is high which means possibility of getting false negative options will be high. Overall Rating 5 doesn't have much fluctuation in score. While overall accuracy score is 92% which is almost equal to random forest classifier. While CV score for this algorithm is 91.97% . Not much difference between accuracy score which is our classification metrics and CV score. Only 0.25% difference we observe.

```
print(cross_val_score(sgd,x,y,cv=5).mean())
```

```
0.9197684515195368
```

Third algorithm used is Linear SVC. Objective of this algorithm is to fit to the data that we provide, returning a best fit line which is hyperplane that decides or categorise our data. Below is the snapshot of classification report for Linear SVC performance on this particular dataset:

```
svc = LinearSVC()
svc.fit(x_train,y_train)
pred =svc.predict(x_test)
acc = classification_report(y_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           1       1.00      0.99      0.99       874
           2       0.78      0.98      0.87       845
           3       0.99      0.82      0.90       873
           4       0.92      0.87      0.89       874
           5       0.96      0.94      0.95       853

    accuracy                           0.92      4319
   macro avg       0.93      0.92      0.92      4319
weighted avg       0.93      0.92      0.92      4319
```

Here precision score is 100% while recall and f1 score is 99%. Which means both the false positive and false negative reports are almost equal which in result f1 score too are equal to recall which is false negative. For rating 2 precision score is 78% which is comparatively too low compare to recall which is 98% means here false negative is high. So f1 score which mostly affected by recall score is 87% . If recall score is high then f1 score will also shows high score because of giving much importance to recall and of bias nature. Rating 3 precious score is 99% while recall is 82% , f1 score is 90% which is average of precision and recall means including both false positive and false negative score is been given. But as in rating 2 we seen that due to recall high score f1 score increased but in rating 3 precision score is high but that not affected much the f1 score because of f1 score mostly rely on recall score. Rating 4 precision score is 92% and recall is 87% and overall f1 score is 89% which is near to recall score. While Rating 5 precision is 96%, recall is 94% and f1 score is 95%. Overall the score of Rating 5 is balanced compare to all other ratings. Overall accuracy score for all the three algorithm are same which is 92%. While cv score which being trained on 5 foldings are 91.74%.

```
print(cross_val_score(svc,x,y,cv=5).mean())
```

```
0.917452966714906
```

Over all the difference between accuracy score and cv score is less for SGD so did hyperparameter tuning on the same using GridSearchCV. Usually we do

Hyperparameter tuning to improve the score if possible.

```
param = {'penalty' :['l2', 'l1', 'elasticnet'],
        'alpha':[0.001,0.002,0.004],
        'l1_ratio':[0.15,0.20,0.30],
        'max_iter':[100,200,500]}
```

```
GC = GridSearchCV(sgd,param,cv=5)
```

```
GC.fit(x_train,y_train)
```

```
GridSearchCV(cv=5, estimator=SGDClassifier(),
             param_grid={'alpha': [0.001, 0.002, 0.004],
                         'l1_ratio': [0.15, 0.2, 0.3],
                         'max_iter': [100, 200, 500],
                         'penalty': ['l2', 'l1', 'elasticnet']})
```

nd output; double click to hide output

```
GC.best_params_
```

```
{'alpha': 0.002, 'l1_ratio': 0.2, 'max_iter': 200, 'penalty': 'l1'}
```

```
final_sgd = SGDClassifier(alpha=0.02,l1_ratio=0.2,max_iter =200, penalty='l1' )
final_sgd.fit(x_train,y_train)
pred = final_sgd.predict(x_test)
acc = accuracy_score(pred,y_test)
print(acc*100)
```

```
91.75735123871267
```

Did hyperparameter using 4 parameters which is penalty, alpha, l1 ratio and max iteration. Provided to each parameter 3 values on which I was suppose to train the model. One by one using each value the model get trained and will provide accuracy score. After training got best parameters and used the same on my best performing algorithm which is SGD and predicted the score . As a result got 91.75% . Means score has not improved . Even 0.25% score been reduced. Here might be chance that if we increase the parameter and values inside that we can have much better score.

KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION :

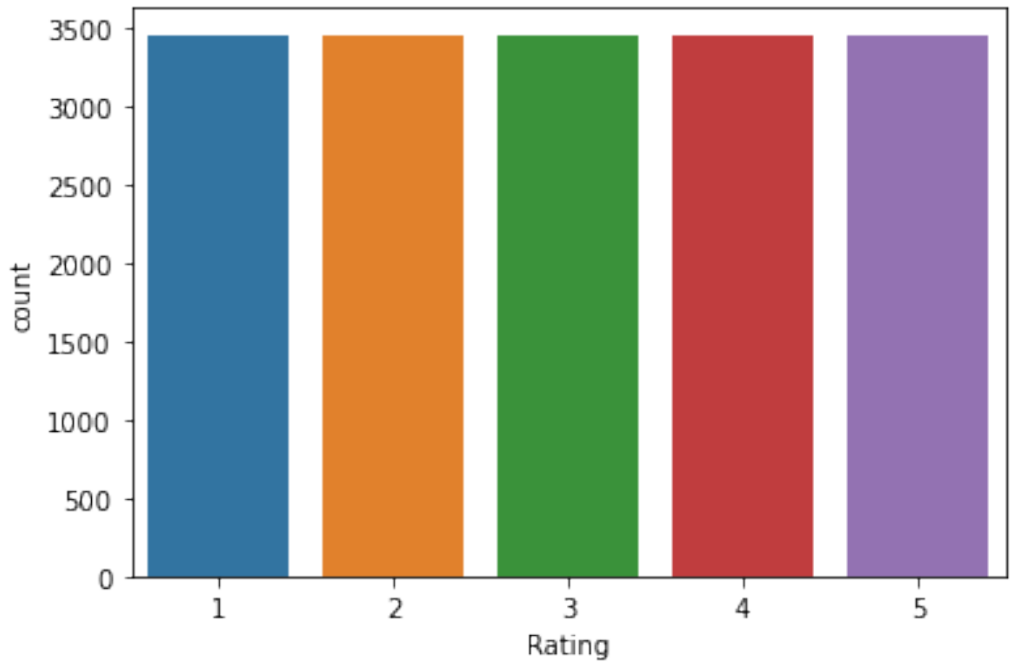
.Key metrics used here are classification report and accuracy score. Classification report usually we use to show scores of all the other metrics such as precision, Recall, F1 score and accuracy score. Means summarisation of all the score. Precision generally gives us false positive . Formula for the same is TP/TP+FP. While recall highly associated to false negative and its formula is

TP/TP+FN. F1 score is a function of Precision and Recall. This one we use generally to balance between precision and recall . Formula for the same is 2*(Precision * Recall / Precision + Recall). Next coming to accuracy score which gives us the result after prediction. And this score is the overall score calculated based on training dataset. While F1 score is considered to be better than Accuracy score as it is a combination of both false positive and false negative

VISUALISATIONS :

Visualisation gives us more clear idea of what the information we require by it's visual context through graph or map. It is easier to identify trends, patterns and outliers within large dataset. To have visualisation here I used matplotlib, seaborn and wordcloud as this is text based problem statement.

Below is the countplot of all the rating given based on reviews by customer :



As all the ratings do have equal number of data so we are getting all the bar plot on same scale which 3456.

Using word cloud we can have visualisation of which words are more important . As using this visualisation technique words which are having high value will be more visible and which are having low value will have less visibility. Below is the snapshot for the same :

Using word cloud we can have visualisation of which words are more im. Here we can see that bad quality , useless product are larger in size which means these words are having high value while money poor, product hated useless all these words are too small which means they do less value in this dataset.

INTERPRETATION OF THE RESULTS :

From visualisation came across that almost all the ratings are having equal number of data means no imbalance problem statement which we seen above in our contplot. Next using World cloud came across which word do have high value and which is having low value. Highly important words are those which are large in size while low valued words are those which are smaller in size. Larger words also we can say as a reader that more focused words are much more visible in larger and bold format. Word cloud make sure that we are focusing on the right words from our problem statement.
Coming to preprocessing here I used NLP technique to preprocess the data as my dataset was text based and text dataset NLP which is a subfield of artificial intelligence and computer science concerned with the interactions between computers and human language. Mainly focus on programming the computers

to process and analyse large amounts of natural language data. As a preprocessing followed the below steps one by one :
1. Applied lowercase to convert all the capital letter words to small letter.
2. Used contraction to expansion where if any word used as short form will be expanded. Like don't will be converted as do not
3. Removed punctuation and special character
4. Removed multiple spaces
5. Removed stop words . Those words which are frequently been used in data but have less value. Such as(is, and, the, are).
6. Then Tokenized the sentence into word.
6. Using lemmatisation converted words to base word . Such as run, ran, running all will have base root "Run".
7. Finally using tfidf vectorisation vectorised the words which means converting string datatype to numerical one which our computer can understand.

Once Preprocessing done .Split feature and label into x and y respectively. Then used standard scaler on all the feature to bring them to same unit. Later selected 3 algorithms to build the model. As this is classification based problem statement so used classification algorithms. These are :
1. Random Forest Classifier
2. LinearSVC
3. SGDClassifier
And for Hyperparameter tuning used Grid Search CV

Metrics used to check the score of all the models are :
1. Precision which shows false positive report
2. Recall which shows false negative report
3. F1 score which is the combination of both false positive and false negative score. Used to make balance between precision and recall.
4. Accuracy score is one metric which shows the over performance of model on a particular dataset . Means how much a particular trained dataset been understood by model and as a result shown using test dataset .

All the models accuracy score were same which is 92% only the difference was in CV score based on which decided the best performing model. Means difference between accuracy score and CV score for which model was less selected that as best performing model. SGD classifier here is the best performing model among all. Did hyperparameter tuning on the same using GridSearchCV to improve the score more if possible. But there were no

improvement in score. May be using more parameters or increasing value within each parameter would have improved the score.

CONCLUSION

KEY FINDINGS AND CONCLUSIONS OF THE STUDY :

Working for a client who has a website where people write different reviews for technical products. Client now added a new feature to the website where reviewer not only can write their reviews but also can add stars for a particular product. Rating which they are suppose to give is out of 5 and there are 5 options available which is 1 star, 2 star, 3 star, 4 star and 5 star. Here what our client requirement is to predict ratings for those reviews written in past and they don't have a rating. So here as data scientist we need build an application which can predict rating by seeing the review. For this first we needed the data which I scrapped from Flipkart site for different technical products. And tried to have equal number of data for each review to balance the dataset. If the dataset is not balanced model performance will be affected and the score will not be accurate as expected. Once scrapped all the required data saved the same to excel format. Now to build the model import dataset using pandas library into Jupiter notebook. Total number of rows available here is 17,275 and column is 2 . Using info method checked datatype for both the column. Rating datatype here is integer while review is having object form means string datatype. There are no missing values either in any of the column or rows. This is classification based problem statement as in target column there are discrete values means which are fixed one. For preprocessing used NLTK (Natural Language Tool Kit) library from where import NLP (Natural Language processing) Which makes computer understand the text data in natural way . Interpretation of text or speech language become easier for computer by using natural language processing as it tries to understand like how we human think. Coming to preprocessing converted all the uppercase word to lowercase , using contractions to expansion expanded the short written words such as "don't" to "do not" , then removed all the punctuations and special character which were

present in sentence, removed multiple spaces, then removed stop words . These words are frequently occurring words but do have less contribution. So to avoid noise removing such words. Then tokenised all the words means breaking sentence into words which is known as tokens. After tokenisation lemmatisation done on dataset . Lemmatization also known as lemma convert all the words to it's base form means "Run, Ran, Running all will be converted to base word which is Run". After lemmatisation using Tfidf (Term frequency - inverse document frequency) which is one of the statistical measure that evaluated how relevant a word is to document in a collection of documents. Term frequency calculated of a word in a document while inverse frequency calculates of the word across a set of documents. Which means how common or rare a word is in the entire document set. Tf-Idf formula mentioned below :

Tf-idf

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

tf-idf(t, d) = tf(t, d) * idf(t)

will be converted to base . Using this transformed all the strings to numbers which our machine algorithm can understand. For visualisation used word cloud to know which is more importance and this we can conclude just by looking after the image. Words which are larger and bold are caring high weightage and which are small in size will have less weightage. This gives us more clarity that on which word we need to focus more.
Once preprocessing and visualisation done separated feature and label into x and y respectively for model building. Using standard scaler scaled all the feature to bring them to same unit. After this selected three algorithm for model building and prediction. These are :
1. RandomForest Classifier
2. Linear SVC
3. SGDClassifier

And for hyper parameter tuning used GridSearchCV which helps in improving score .

Metrics used to check the score of all the models are :
1. Precision which shows false positive report

2. Recall which shows false negative report
3. F1 score which is the combination of both false positive and false negative score. Used to make balance between precision and recall.
4. Accuracy score is one metric which shows the over performance of model on a particular dataset . Means how much a particular trained dataset been understood by model and as a result shown using test dataset .

After evaluating all the models came across that accuracy score for all the models are same which is 92% there are difference in score we can see in precision, recall and f1. Mostly f1 score considered to be best in case of text dataset. But there are not much difference even in f1 score too which is the combination of both precision and recall. Then calculated CV score on 5 folding for each model to conclude which model is the best performing model. Even CV score was almost near to each other but with minor difference considered SGD classifier as the best performing model. Because the difference between CV and accuracy score was less for this model. Did hyperparametr tuning on SGD Classifier but score not improved. This may be because of I used here only 4 parameters and in parameter there passed 3 value each. As how many parameters we select and the value inside we give training time increases. So to avoid such selected less parameters and values. May be possibility if would have increased the parameter or values inside that score would have increased . Or if used RandomizedSearchCV or any other hyperparameter to tune the model score would have increased. Atlast saved the model and using test dataset which I passed 25% and 75% for training dataset predicted the model.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE :

As this is text based problem statement so to know which word to give more importance and which to less, used WordCloud with the help of which get visualisation , and could interpret from there that the words which is larger in size and are in bold style need to emphasise more while those which are less visible or smaller in size need to focus less. Using NLP cleaned the dataset. For text dataset generally we use natural language processing as it understand human language and try to interpret speech and text dataset more accurately. Using NLTK library import NLP for preprocessing. Then used 3 algorithms to

build the model in which SGDclassifier is the best performing model as the difference between CV and accuracy score for this model is less. Also the F1 score comparatively to all other model is good for all the 5 ratings. Mainly the issue while doing this project I faced was how to clean the data as NLP was completely new project for me. By taking the help of my SME Ms.Khosboo Garg and using various websites and YouTube videos came across how to solve this issue. NLP preprocessing do have various steps when and where to use what in chronological manner was difficult. Even understanding which algorithm will work best for this text dataset was little tough. But with help of all the videos and websites solved these issues . Also thanks to my SME who supported a lot during this project work.

LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK :

As such no limitations found to solve this problem statement. While doing hyperparameter tuning score not imporved . There felt if would have added more parameters or values inside that because value to each parameter passed 3 and total 4 parameters used for SGD Algorithm where each value been trained one by one and pass score accordingly. Score was not at all improved , instead of improving by few points it diminished only. So either I would have used some other hyperparameter tuning like randomised search cv to check if score would have improved.