CS498 AML,AMO HW7

Shrashti Singhal, Ankush Singhal

TOTAL POINTS

100 / 100

QUESTION 1

1 Show the distribution graph of words counts vs word rank. **5 / 5**

QUESTION 2

2 List the stop words you choose as well as the frequency threshold. **5 / 5**

QUESTION 3

3 After chosing the stop words, show the distribution graph of words counts vs word rank. **5 / 5**

QUESTION 4

4 Show the snippet of your code that you convert all the reviews into bag-of-words formulation using your chosen stop words and your code for nearest-neighbours with cos-distance. **15 / 15**

QUESTION 5

5 Show the original reviews with the distance scores **10 / 10**

QUESTION 6

6 Show your query results and explain the reasons that you choose them. **10 / 10**

QUESTION 7

7 Show your code for creating classifier. Report the accuracy on train and test dataset with threshold 0.5. **10 / 10**

QUESTION 8

8 Show your code for plotting predicted scores and show the figure. **10 / 10**

QUESTION 9

9 Report the accuracy on train and test dataset with a different threshold. Explain why you choose that threshold. Plot the ROC curve. **20 / 20**

QUESTION 10

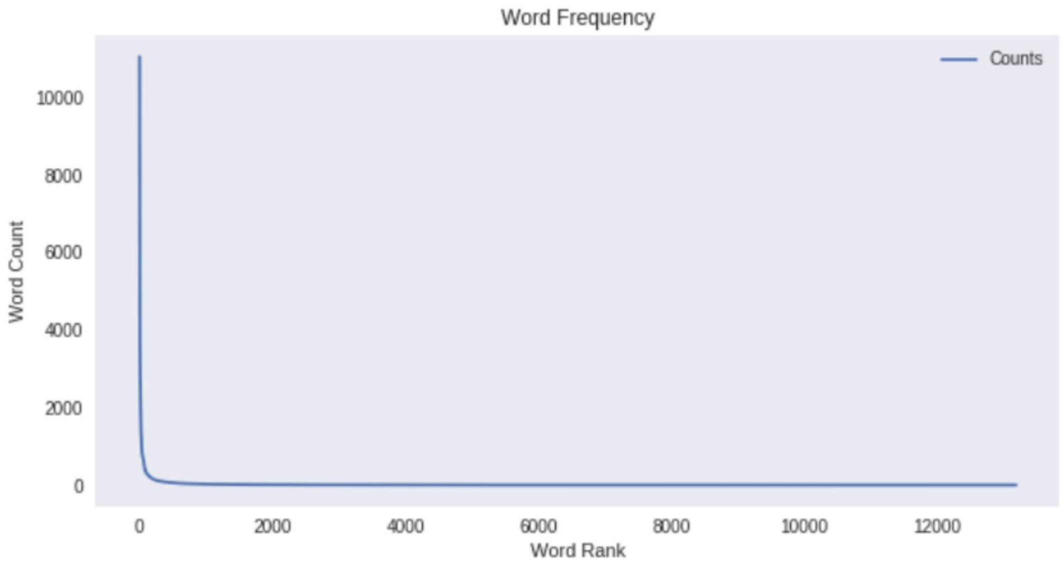
10 Choose the threshold that minimizes false positives while maximizing true positives. Explain your reason. **10 / 10**

QUESTION 11

11 Late Penalty **0 / 0**

Distribution graph

- Extracted text and stars columns as X (data) and y (label), from the dataset.
- Converted the text into lower case then into bag-of-words representation.
- Plotted the Graph of distribution of words counts vs word rank.



Identify the stop words

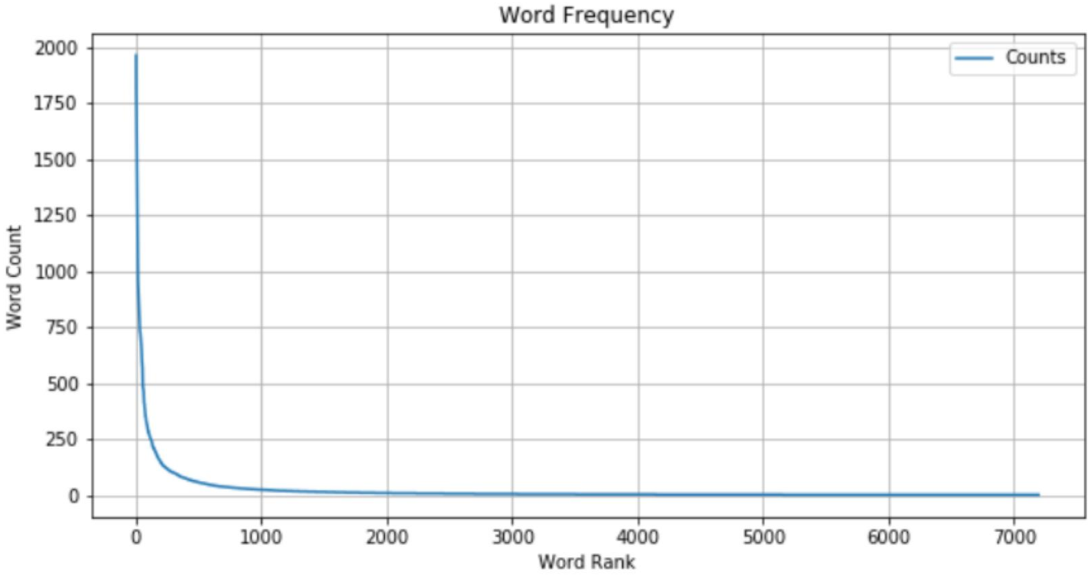
- Stop words selected are the most frequently used words, so that we can remove common English language words and retain the useful words.
- Identified set of common stop words by plotting a graph of 50 most frequently used words.
- From the graph, we identified that, top 16 frequently used words have a frequency of 2000 and above.
- **Therefore Frequency Threshold selected is 2000 to retrieve top 16 stop words**

List of Stop Words

1. The
2. And
3. I
4. To
5. A
6. Was
7. It
8. Of
9. For
10. In
11. My
12. Is
13. That
14. They
15. this
16. we

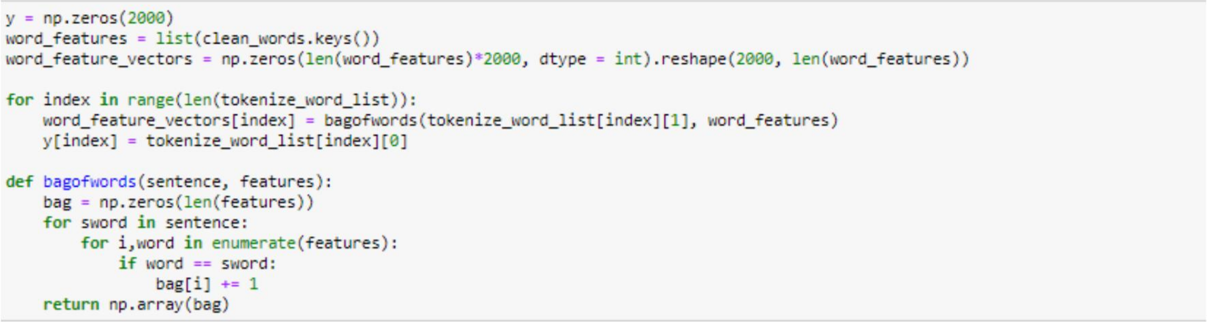
Distribution graph again

- Removed all the stop words(Words appeared more than 2000 times)
- Kept the words appeared exactly 1 time.
- Reprocessed your data using the stop-words list, the max document frequency and the minimum word occurrence.
- Graphed the updated words counts vs word rank.



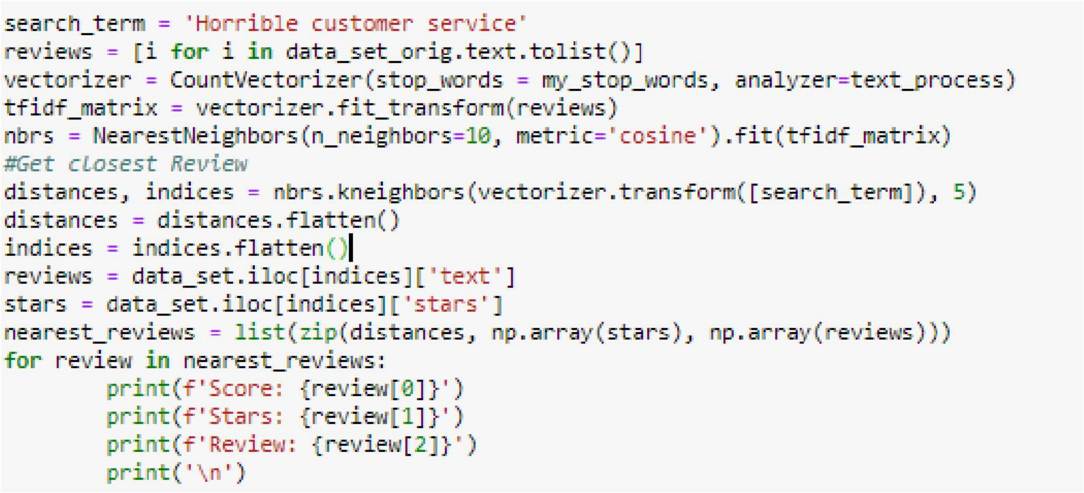
Code snippets

Convert into Bag of Words



```
y = np.zeros(2000)
word_features = list(clean_words.keys())
word_feature_vectors = np.zeros(len(word_features)*2000, dtype = int).reshape(2000, len(word_features))

for index in range(len(tokenize_word_list)):
    word_feature_vectors[index] = bagofwords(tokenize_word_list[index][1], word_features)
    y[index] = tokenize_word_list[index][0]

def bagofwords(sentence, features):
    bag = np.zeros(len(features))
    for sword in sentence:
        for i,word in enumerate(features):
            if word == sword:
                bag[i] += 1
    return np.array(bag)
```

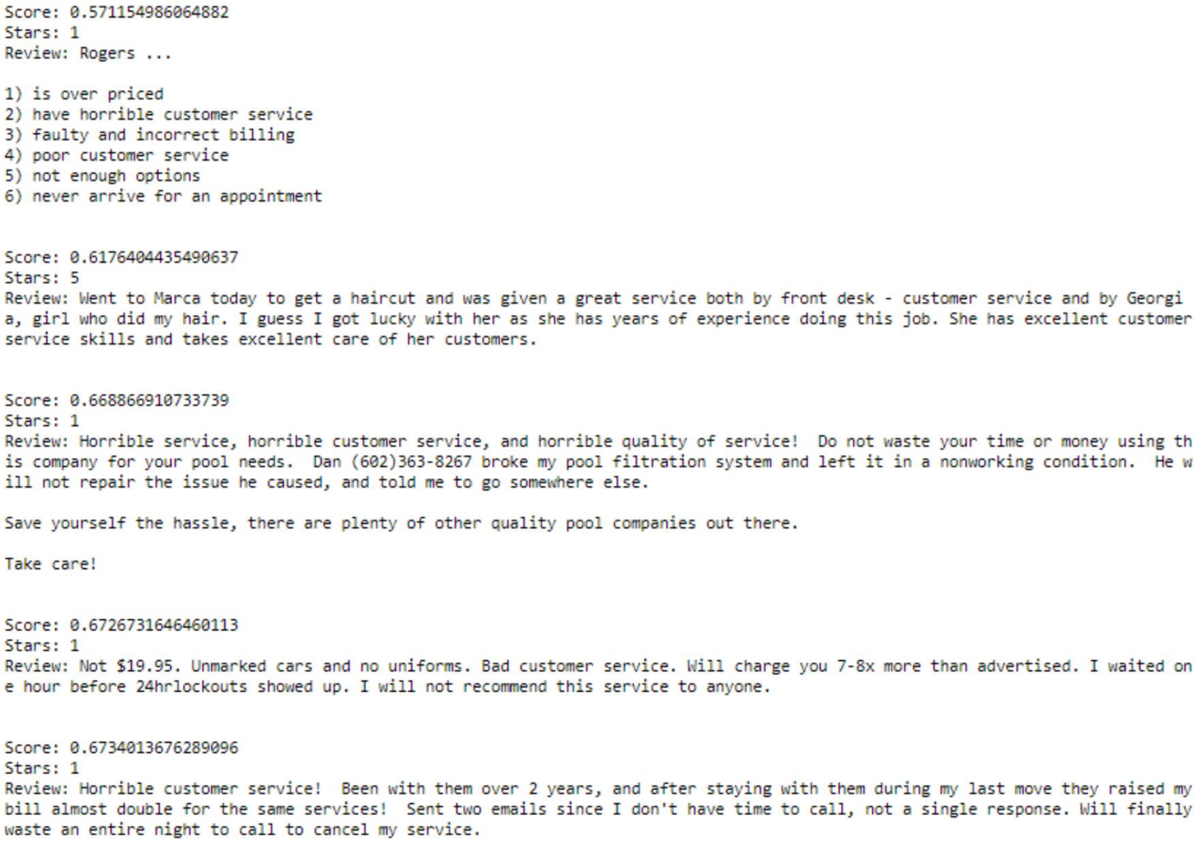
Cosine of Nearest Neighbour



```
search_term = 'Horrible customer service'
reviews = [i for i in data_set_orig.text.tolist()]
vectorizer = CountVectorizer(stop_words = my_stop_words, analyzer=text_process)
tfidf_matrix = vectorizer.fit_transform(reviews)
nbrs = NearestNeighbors(n_neighbors=10, metric='cosine').fit(tfidf_matrix)
#Get closest Review
distances, indices = nbrs.kneighbors(vectorizer.transform([search_term]), 5)
distances = distances.flatten()
indices = indices.flatten()
reviews = data_set.iloc[indices]['text']
stars = data_set.iloc[indices]['stars']
nearest_reviews = list(zip(distances, np.array(stars), np.array(reviews)))
for review in nearest_reviews:
        print(f'Score: {review[0]}')
        print(f'Stars: {review[1]}')
        print(f'Review: {review[2]}')
        print('\n')
```

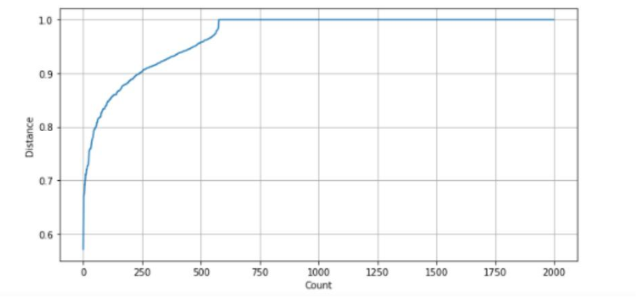
Reviews with score

- Used nearest neighbor with a cos-distance metric, to find documents matching the query phase "*Horrible customer service*".
- We found 5 reviews matching to the query phrase.
- Below are the original reviews with scores

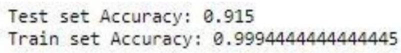


```
Score: 0.571154986064882
Stars: 1
Review: Rogers ...

1) is over priced
2) have horrible customer service
3) faulty and incorrect billing
4) poor customer service
5) not enough options
6) never arrive for an appointment


Score: 0.6176404435490637
Stars: 5
Review: Went to Marca today to get a haircut and was given a great service both by front desk - customer service and by Georgi
a, girl who did my hair. I guess I got lucky with her as she has years of experience doing this job. She has excellent customer
service skills and takes excellent care of her customers.


Score: 0.668866910733739
Stars: 1
Review: Horrible service, horrible customer service, and horrible quality of service!  Do not waste your time or money using th
is company for your pool needs.  Dan (602)363-8267 broke my pool filtration system and left it in a nonworking condition.  He w
ill not repair the issue he caused, and told me to go somewhere else.

Save yourself the hassle, there are plenty of other quality pool companies out there.

Take care!


Score: 0.6726731646460113
Stars: 1
Review: Not $19.95. Unmarked cars and no uniforms. Bad customer service. Will charge you 7-8x more than advertised. I waited on
e hour before 24hrlockouts showed up. I will not recommend this service to anyone.


Score: 0.6734013676289096
Stars: 1
Review: Horrible customer service!  Been with them over 2 years, and after staying with them during my last move they raised my
bill almost double for the same services!  Sent two emails since I don't have time to call, not a single response. Will finally
waste an entire night to call to cancel my service.
```

Query results



Using Threshold 0.7 from above plot we received following results:



```
Test set Accuracy: 0.915
Train set Accuracy: 0.9994444444444445
```

Score: 0.571154986064882

Stars: 1
Review: Rogers ...

1) is over priced
2) have horrible customer service
3) faulty and incorrect billing
4) poor customer service
5) not enough options
6) never arrive for an appointment

Score: 0.6176404435490637
Stars: 5
Review: Went to Marca today to get a haircut and was given a great service both by front desk - customer service and by Georgia, girl who did my hair. I guess I got lucky with her as she has years of experience doing this job. She has excellent customer service skills and takes excellent care of her customers.

Score: 0.668866910733739
Stars: 1
Review: Horrible service, horrible customer service, and horrible quality of service! Do not waste your time or money using this company for your pool needs. Dan (602)363-8267 broke my pool filtration system and left it in a nonworking condition. He will not repair the issue he caused, and told me to go somewhere else.

Save yourself the hassle, there are plenty of other quality pool companies out there.

Take care!

Score: 0.6726731646460113
Stars: 1
Review: Not $19.95. Unmarked cars and no uniforms. Bad customer service. Will charge you 7-8x more than advertised. I waited one hour before 24hrlockouts showed up. I will not recommend this service to anyone.

Score: 0.6734013676289096
Stars: 1
Review: Horrible customer service! Been with them over 2 years, and after staying with them during my last move they raised my bill almost double for the same services! Sent two emails since I don't have time to call, not a single response. Will finally waste an entire night to call to cancel my service.

Score: 0.6797436923898257
Stars: 1
Review: Worse customer service . trying to ask for help .... No one volunteer to help. This is ridiculous

Score: 0.6913933000758161
Stars: 1
Review: They shut down. Makes sense, they had terrible service and subpar food..should have listened to your customer base.

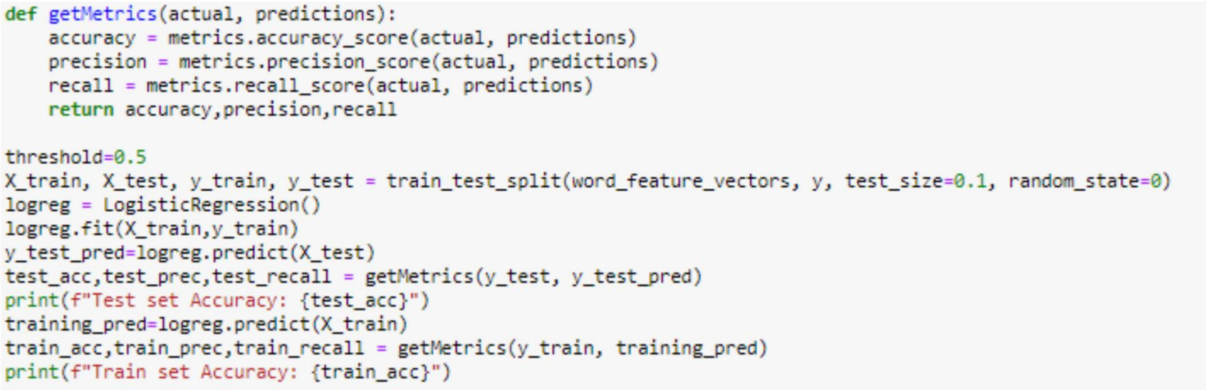
Score: 0.6913933000758161
Stars: 5
Review: Bravo! Wonderful world-class customer service and great products. I went their last week and happily blew my diet.

Score: 0.6993415887988683
Stars: 5
Review: This was the highlight of our trip to the Outlet Mall.

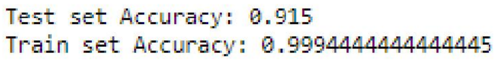
We had great customer service, and that goes a long ways. We were not in the market to buy, but we were in the market for information, and they delivered.

If we lived in the area, I'd be back just for the customer service. Bose makes a great product, and here the product was enhanced as the information about the products was delivered with a smile and answers.

Accuracy with threshold 0.5

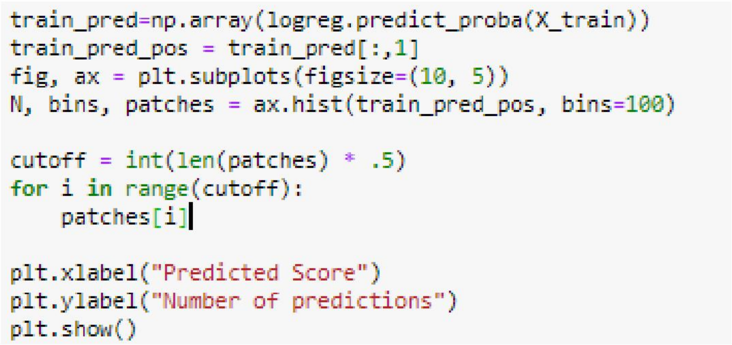


```
def getMetrics(actual, predictions):
    accuracy = metrics.accuracy_score(actual, predictions)
    precision = metrics.precision_score(actual, predictions)
    recall = metrics.recall_score(actual, predictions)
    return accuracy,precision,recall

threshold=0.5
X_train, X_test, y_train, y_test = train_test_split(word_feature_vectors, y, test_size=0.1, random_state=0)
logreg = LogisticRegression()
logreg.fit(X_train,y_train)
y_test_pred=logreg.predict(X_test)
test_acc,test_prec,test_recall = getMetrics(y_test, y_test_pred)
print(f"Test set Accuracy: {test_acc}")
training_pred=logreg.predict(X_train)
train_acc,train_prec,train_recall = getMetrics(y_train, training_pred)
print(f"Train set Accuracy: {train_acc}")
```

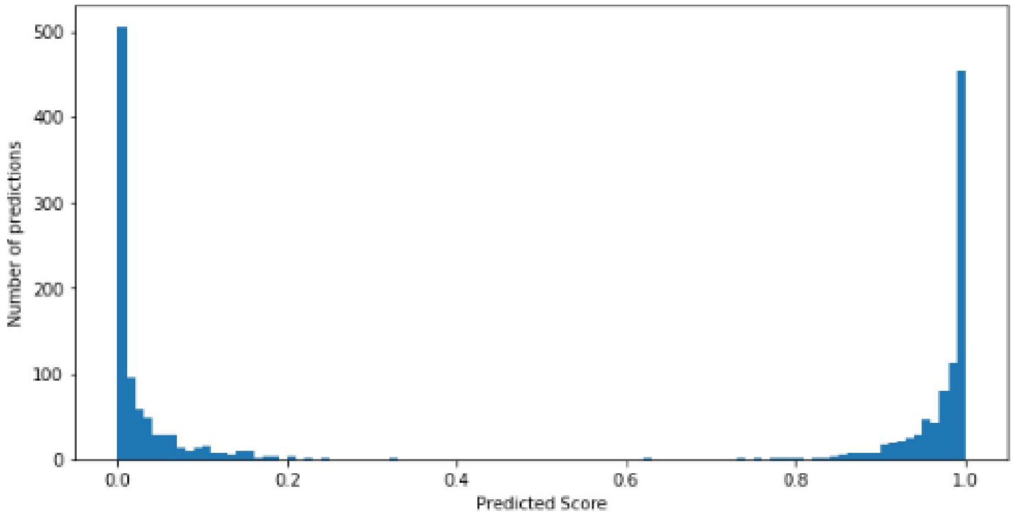


```
Test set Accuracy: 0.915
Train set Accuracy: 0.9994444444444445
```

Predicted scores:



```
train_pred=np.array(logreg.predict_proba(X_train))
train_pred_pos = train_pred[:,1]
fig, ax = plt.subplots(figsize=(10, 5))
N, bins, patches = ax.hist(train_pred_pos, bins=100)

cutoff = int(len(patches) * .5)
for i in range(cutoff):
    patches[i]

plt.xlabel("Predicted Score")
plt.ylabel("Number of predictions")
plt.show()
```



Accuracy again and curve

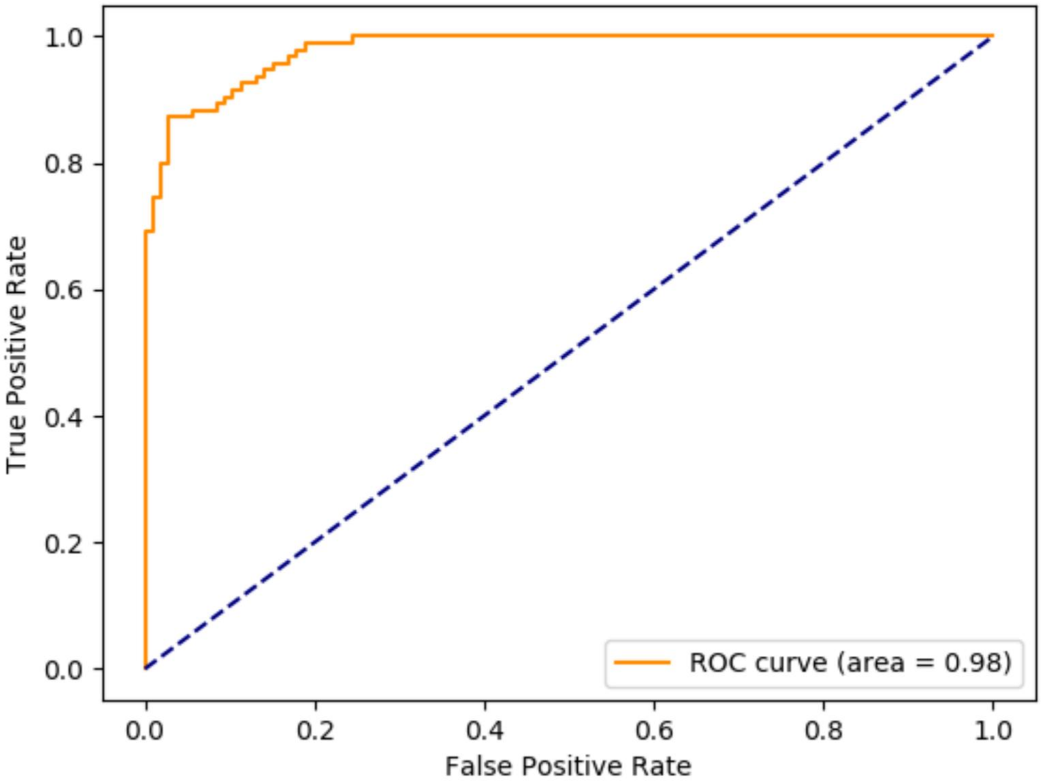
On Threshold of 0.7

Test set Accuracy: 1.0

Train set Accuracy: 0.905

We selected 0.7 as the threshold because histogram gave better results with threshold 0.7 than 0.5

Yes, the threshold is improved.



Best threshold

Maximum True positive rate when False negative rate is minimum

False Negative Rate= 0.2

Maximum True positive = 0.98

Threshold = 0.98/0.2 = 0.49

Threshold is considered good when we get the best accuracy. It is also considered good when we have minimum false negatives.

If only best accuracy would be considered 0.7 was better than 0.49. But If we consider both the parameters to minimize false negatives and maximize the true positives 0.49 is best.