XGBoost On Amazon Fine Food Reviews Analysis

Data Source: https://www.kaggle.com/snap/amazon-fine-food-reviews

EDA: https://nycdatascience.com/blog/student-works/amazon-fine-foods-visualization/

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.

Number of reviews: 568,454
Number of users: 256,059
Number of products: 74,258
Timespan: Oct 1999 - Oct 2012
Number of Attributes/Columns in data: 10

Attribute Information:

1. Id
2. ProductId - unique identifier for the product
3. UserId - unqiue identifier for the user
4. ProfileName
5. HelpfulnessNumerator - number of users who found the review helpful
6. HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

Objective:

Given a review, determine whether the review is positive (rating of 4 or 5) or negative (rating of 1 or 2).

[Q] How to determine if a review is positive or negative?

[Ans] We could use Score/Rating. A rating of 4 or 5 can be cosnidered as a positive review. A rating of 1 or 2 can be considered as negative one. A review of rating 3 is considered nuetral and such reviews are ignored from our analysis. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

[1]. Reading Data

[1.1] Loading the data

The dataset is available in two forms

1. .csv file
2. SQLite Database

Here as we only want to get the global sentiment of the recommendations (positive or negative), we will purposefully ignore all Scores equal to 3. If the score is above 3, then the recommendation wil be set to "positive". Otherwise, it will be set to "negative".

In [0]:

```
!pip install tqdm
import warnings
warnings.filterwarnings("ignore")
import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix,roc_curve,auc,roc_auc_score
from sklearn.model_selection import train_test_split,RandomizedSearchCV,TimeSeriesSplit
!pip install vaderSentiment
```

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import re
import pickle
from tqdm import tqdm
import os
from bs4 import BeautifulSoup
from scipy import sparse
import random
%env JOBLIB_TEMP_FOLDER=/tmp
!pip install xgboost
from xgboost import XGBClassifier
```

In [0]:

```
!pip install wordcloud
from wordcloud import WordCloud
```

In [0]:

```
!pip install -q kaggle
!mkdir -p ~/.kaggle
!cp kaggle.json ~/.kaggle/
!kaggle datasets download -d snap/amazon-fine-food-reviews
!unzip amazon-fine-food-reviews.zip

input_data = pd.read_csv('Reviews.csv')
```

```
Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run '
chmod 600 /content/.kaggle/kaggle.json'
Downloading amazon-fine-food-reviews.zip to /content/datalab
 99%|████████████████████████████████████████ | 248M/251M [00:01<00:00, 170MB/s]
100%|█████████████████████████████████████████| 251M/251M [00:01<00:00, 191MB/s]
Archive:  amazon-fine-food-reviews.zip
  inflating: Reviews.csv
  inflating: database.sqlite
  inflating: hashes.txt
```

[2] Exploratory Data Analysis

[2.1] Data Cleaning: Deduplication

It is observed (as shown in the table below) that the reviews data had many duplicate entries. Hence it was necessary to remove duplicates in order to get unbiased results for the analysis of the data. Following is an example:

In [0]:

```
# Removing all the neutral reviews
input_data = input_data[input_data.Score != 3]
# Sorting the data with respect to the ProductID
sorted_data=input_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicks
ort', na_position='last')
# Removing the duplicates of UserId,ProfileName,Time,Text
input_data = sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first
', inplace=False)
# Removing all the rows having the usefulness more or equal than the total number of votes
input_data = input_data[input_data.HelpfulnessNumerator<=input_data.HelpfulnessDenominator]
# Sorting the data by time in ascending order
data = input_data.iloc[input_data.Time.argsort()]
# Assigning score value 1 for reviews > 4 else 0 i.e 1 represents 'Positive' and 0 represents 'Neg
ative' Review
data.Score = data.Score.map(lambda x : 1 if (x > 3) else 0)
```

In [0]:

```
data.head(5)
```

Out[0]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Su
150523	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0	0	1	939340800	EVEI edu
150500	150501	0006641040	AJ46FKXOVC7NR	Nicholas A Mesiano	2	2	1	940809600	Th grea spe
451855	451856	B00004CXX9	AIUWLEQ1ADEG5	Elizabeth Medina	0	0	1	944092800	Ente
374358	374359	B00004CI84	A344SMIA5JECGM	Vincent P. Ross	1	2	1	944438400	A day 1
451854	451855	B00004CXX9	AJH6LUC1UT1ON	The Phantom of the Opera	0	0	1	946857600	FANT

In [0]:

```
data.Score.value_counts()
```

Out[0]:

```
1    307061
0     57110
Name: Score, dtype: int64
```

[3] Preprocessing

[3.1]. Preprocessing Review Text

Now that we have finished deduplication our data requires some preprocessing before we go on further with analysis and making the prediction model.

Hence in the Preprocessing phase we do the following in the order below:-

1. Begin by removing the html tags
2. Remove any punctuations or limited set of special characters like , or . or # etc.
3. Check if the word is made up of english letters and is not alpha-numeric
4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)
5. Convert the word to lowercase
6. Remove Stopwords

After which we collect the words used to describe positive and negative reviews

Methods For Pre-Processing

In [0]:

```
def remove_urls(sentence):
  """A method to remove URLS from the text data
     Credits to : https://stackoverflow.com/a/40823105/4084039
  """
  return re.sub(r"http\S+", "", sentence)

def remove_tags(sentence):
  """A method to remove tag elemts from the text data
     Credits to : https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an-element
  """
```

```
  return BeautifulSoup(sentence, 'lxml').get_text()

def deconstructing_sentence(sentence):
  """ A method to expand certain words like can't to cannot
      Credits to : https://stackoverflow.com/a/47091490/4084039
  """
  # specific
  phrase = re.sub(r"won't", "will not", sentence)
  phrase = re.sub(r"can\'t", "can not", phrase)
  # general
  phrase = re.sub(r"n\'t", " not", phrase)
  phrase = re.sub(r"\'re", " are", phrase)
  phrase = re.sub(r"\'s", " is", phrase)
  phrase = re.sub(r"\'d", " would", phrase)
  phrase = re.sub(r"\'ll", " will", phrase)
  phrase = re.sub(r"\'t", " not", phrase)
  phrase = re.sub(r"\'ve", " have", phrase)
  phrase = re.sub(r"\'m", " am", phrase)
  return phrase

def remove_words_with_numbers(sentence):
  """ A method to remove words with numbers
      Credits to : https://stackoverflow.com/a/18082370/4084039
  """
  return re.sub("\S*\d\S*", "", sentence).strip()

def remove_special_chars(sentence):
  """ A method to remove special characters in text sentence/data
      Credits to : https://stackoverflow.com/a/18082370/4084039
  """
  return re.sub('[^A-Za-z0-9]+', ' ', sentence)

# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
# <br /><br /> ==> after the above steps, we are getting "br br"
# we are including them into stop words list
# instead of <br /> if we have <br/> these tags would have revmoved in the 1st step

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "y
ou're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', '
while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under'
, 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll'
, 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "do
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
"mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])


def clean_sentence(list_of_sent):
  preprocessed_text = []
  for i in tqdm(list_of_sent):
    sent = remove_urls(i)
    sent = remove_tags(sent)
    sent = deconstructing_sentence(sent)
    sent = remove_words_with_numbers(sent)
    sent = remove_special_chars(sent)
    sentence = ' '.join(e.lower() for e in sent.split() if e.lower() not in stopwords)
```

```
    preprocessed_text.append(sentence.strip())
  return preprocessed_text

# Using Vader Pre Trained Corpus Of Strings
analyser = SentimentIntensityAnalyzer()
def vaderAnalysis(sentence):
  """Method to return the positive score of the summary review"""
  return (analyser.polarity_scores(sentence)['pos'])
```

In [0]:

```
data['Cleaned'] = clean_sentence(data.Text.values)
```

```
100%|██████████| 364171/364171 [02:47<00:00, 2180.58it/s]
```

[3.2] Preprocessing Review Summary

In [0]:

```
data['Cleaned_Summary'] = clean_sentence(data.Summary.astype(str).values)
```

```
100%|██████████| 364171/364171 [01:51<00:00, 3252.49it/s]
```

[4] Featurization

In [0]:

```
# Omitting null values rows in the speicified columns
data = data[data[['Cleaned','Cleaned_Summary','Score']].notnull()]
# Counting the number of words in the reviews
data["rev_len"] = data['Cleaned'].str.split().str.len()


data['sum_pos_score'] = [vaderAnalysis(i) for i in data.Cleaned_Summary.values]
```

In [0]:

```
x_train,x_test,y_train,y_test = train_test_split(data[['Cleaned','sum_pos_score','rev_len']],data.
Score.values,shuffle=False,test_size=0.3)
```

[4.1] BAG OF WORDS

In [0]:

```
from sklearn.feature_extraction.text import CountVectorizer
BoW_dict = CountVectorizer(min_df=8,max_features = 5000,ngram_range = (1,2)).fit(x_train.Cleaned)
BoW_train = BoW_dict.transform(x_train.Cleaned)
BoW_test = BoW_dict.transform(x_test.Cleaned)

rev_lens_train = x_train.rev_len.values.reshape(-1,1)
sum_pos_score_train = x_train.sum_pos_score.values.reshape(-1,1)
training_data = sparse.hstack((BoW_train, rev_lens_train,sum_pos_score_train))
feat_names = []
feat_names = BoW_dict.get_feature_names()
feat_names.append('review_length')
feat_names.append('pos_score')

rev_lens_test = x_test.rev_len.values.reshape(-1,1)
sum_pos_score_test = x_test.sum_pos_score.values.reshape(-1,1)
test_data = sparse.hstack((BoW_test, rev_lens_test,sum_pos_score_test))
```

[4.3] TF-IDF

In [0]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
TFIDF_dict = TfidfVectorizer(min_df=10,max_features = 5000,ngram_range = (1,2)).fit(x_train.Cleaned
)
TFIDF_train = TFIDF_dict.transform(x_train.Cleaned)
TFIDF_test = TFIDF_dict.transform(x_test.Cleaned)

rev_lens_train = x_train.rev_len.values.reshape(-1,1)
sum_pos_score_train = x_train.sum_pos_score.values.reshape(-1,1)
training_data = sparse.hstack((TFIDF_train, rev_lens_train,sum_pos_score_train))
feat_names = []
feat_names = TFIDF_dict.get_feature_names()
feat_names.append('review_length')
feat_names.append('pos_score')

rev_lens_test = x_test.rev_len.values.reshape(-1,1)
sum_pos_score_test = x_test.sum_pos_score.values.reshape(-1,1)
test_data = sparse.hstack((TFIDF_test, rev_lens_test,sum_pos_score_test))
```

[4.4] Word2Vec

In [0]:

```
# Train your own Word2Vec model using your own text corpus
list_of_sentence=[]
for sent in tqdm(x_train.Cleaned):
    list_of_sentence.append(sent.split())
```

```
100%|██████████| 254919/254919 [00:04<00:00, 57028.49it/s]
```

In [0]:

```
!pip install gensim
import gensim
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

w2v_model=gensim.models.Word2Vec(list_of_sentence,min_count=5,size=50, workers=8)
w2v_words = list(w2v_model.wv.vocab)
```

[4.4.1] Converting text into vectors using Avg W2V, TFIDF-W2V

[4.4.1.1] Avg W2v

In [0]:

```
sent_vectors_train = []
for sent in x_train.Cleaned.values:
    sent_vec = np.zeros(50)
    cnt_words =0
    for word in sent.split():
        try:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
        except:
            pass
    sent_vec /= cnt_words
    sent_vectors_train.append(sent_vec)

sent_vectors_train = np.nan_to_num(sent_vectors_train)

rev_lens_train = x_train.rev_len.values.reshape(-1,1)
sum_pos_score_train = x_train.sum_pos_score.values.reshape(-1,1)
training_data = np.hstack((sent_vectors_train, rev_lens_train,sum_pos_score_train))
```

In [0]:

```
sent_vectors_test = []
for sent in x_test.Cleaned.values:
    sent_vec = np.zeros(50)
    cnt_words =0
    for word in sent.split():
        try:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
        except:
            pass
    sent_vec /= cnt_words
    sent_vectors_test.append(sent_vec)

sent_vectors_test = np.nan_to_num(sent_vectors_test)

rev_lens_test = x_test.rev_len.values.reshape(-1,1)
sum_pos_score_test = x_test.sum_pos_score.values.reshape(-1,1)
test_data = np.hstack((sent_vectors_test, rev_lens_test,sum_pos_score_test))
```

[4.4.1.2] TFIDF weighted W2v

In [0]:

```
!kaggle datasets download -d sanjeev5/tfidfw2vtrain
!unzip tfidfw2vtrain.zip
tfidf_w2v_train = pickle.load( open( "Tfidf_W2V_Train.txt", "rb" ) )
!kaggle datasets download -d sanjeev5/tfidfw2vtest
!unzip tfidfw2vtest.zip
tfidf_w2v_test = pickle.load( open( "Tfidf_W2V_Test.txt", "rb" ) )
```

In [0]:

```
rev_lens_train = x_train.rev_len.values.reshape(-1,1)
sum_pos_score_train = x_train.sum_pos_score.values.reshape(-1,1)
training_data = np.hstack((tfidf_w2v_train, rev_lens_train,sum_pos_score_train))

rev_lens_test = x_test.rev_len.values.reshape(-1,1)
sum_pos_score_test = x_test.sum_pos_score.values.reshape(-1,1)
test_data = np.hstack((tfidf_w2v_test, rev_lens_test,sum_pos_score_test))
```

[5] Assignment 9: Random Forests

1. **Apply Random Forests & GBDT on these feature sets**
    - SET 1:Review text, preprocessed one converted into vectors using (BOW)
    - SET 2:Review text, preprocessed one converted into vectors using (TFIDF)
    - SET 3:Review text, preprocessed one converted into vectors using (AVG W2v)
    - SET 4:Review text, preprocessed one converted into vectors using (TFIDF W2v)
2. **The hyper paramter tuning (Consider any two hyper parameters)**
    - Find the best hyper parameter which will give the maximum AUC value
    - Find the best hyper paramter using k-fold cross validation or simple cross validation data
    - Use gridsearch cv or randomsearch cv or you can also write your own for loops to do this task of hyperparameter tuning
3. **Feature importance**
    - Get top 20 important features and represent them in a word cloud. Do this for BOW & TFIDF.
4. **Feature engineering**
    - To increase the performance of your model, you can also experiment with with feature engineering like :
        - Taking length of reviews as another feature.
        - Considering some features from review summary as well.

5. **Representation of results**

   - You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure
     with X-axis as **n_estimators**, Y-axis as **max_depth**, and Z-axis as **AUC Score** , we have given the notebook which explains how to plot this 3d plot, you can find it in the same drive *3d_scatter_plot.ipynb*

(or)

   - You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure
     seaborn heat maps with rows as **n_estimators**, columns as **max_depth**, and values inside the cell representing **AUC Score**
   - You choose either of the plotting techniques out of 3d plot or heat map
   - Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
   - Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps.

6. **Conclusion**

   - You need to summarize the results at the end of the notebook, summarize it in the table format. To print out a table please refer to this prettytable library link

Note: Data Leakage

1. There will be an issue of data-leakage if you vectorize the entire data and then split it into train/cv/test.
2. To avoid the issue of data-leakag, make sure to split your data first and then vectorize it.
3. While vectorizing your data, apply the method fit_transform() on you train data, and apply the method transform() on cv/test data.
4. For more details please go through this link.

[5.2] Applying GBDT using XGBOOST

In [0]:

```
len(y_train[y_train==0])/len(y_train[y_train==1])
```

Out[0]:

```
0.17533772880261883
```

In [0]:

```
param_grid = {'max_depth': list(range(2,7,1)) ,'n_estimators': list(range(10,210,10))}
tss = TimeSeriesSplit(n_splits=3)
def best_param_search(train_data,train_label,params,tss):
  """ To choose the best hyperparamters for the XGBoost"""
  xgbc = XGBClassifier(subsample=0.5, colsample_bytree=0.5, seed=1,scale_pos_weight = 1/17)
  rscv = RandomizedSearchCV(xgbc,params, scoring = 'roc_auc', cv=tss, n_jobs = -1,verbose = 1,n_ite
r = 20)
  rscv.fit(train_data, train_label)
  params = rscv.cv_results_['params']
  train_scores = rscv.cv_results_['mean_train_score']
  cv_scores = rscv.cv_results_['mean_test_score']
  pr = []
  for i in params:
    depth_val = str(i['max_depth'])
    sample_split_val = str(i['n_estimators'])
    pr.append(depth_val+','+sample_split_val)
  df_cm = pd.DataFrame(data = [train_scores,cv_scores],index=['Train','CV'], columns=pr)
  plt.figure(figsize=(25, 7))
  heatmap = sns.heatmap(df_cm, annot=True, fmt="f")
  heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right')
  heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=45, ha='right')
  plt.ylabel('Type')
```

```
    plt.xlabel('(max_depth,n_estimators) Params')
    plt.title('Results Heat Map')
    plt.show()

  # Confusion Matrix
def confusion_matrix_display(conf_mtrx,tst_labels,Title):
    """Printing the confusion matrix
       Reused from previous assignments
    """
    class_names = [0,1]
    df_cm = pd.DataFrame(conf_mtrx, index=class_names, columns=class_names)
    TN, FP, FN, TP = conf_mtrx.ravel()
    heatmap = sns.heatmap(df_cm, annot=True, fmt="d")
    heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right')
    heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=45, ha='right')
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
    plt.title(Title + ' Confusion Matrix')
    plt.show()
    print('\nThe TPR is : ',TP/(TP+FN))
    print('The TNR is : ',TN/(TN+FP))
    print('The FPR is : ',FP/(FP+TN))
    print('The FNR is : ',FN/(TP+FN),'\n')

def on_test(train_data,train_label,test_data,test_label,md,n_estimators):
    """XGB on test data"""
    xgb = XGBClassifier(subsample=0.5, colsample_bytree=0.5, seed=1,max_depth=md,n_estimators = n_est
imators,scale_pos_weight = 1/17)
    xgb.fit(train_data,train_label)
    print('The ROC AUC score for the params max_depth' ,md, ' and n_estimators ',n_estimators, 'is '
, roc_auc_score(test_label,xgb.predict_proba(test_data)[:,1]))
    confusion_matrix_display(confusion_matrix(train_label,xgb.predict(train_data)),train_label,'Train
Data Confusion Matrix')
    confusion_matrix_display(confusion_matrix(test_label,xgb.predict(test_data)),test_label,'Test
Data Confusion Matrix')
    roc_curve_draw(train_label,test_label,xgb.predict_proba(train_data)[:,1],xgb.predict_proba(test_d
ata)[:,1])
    return xgb

def Wordcl(title,val):
    """Print the wordcloud for top important features"""
    wordcloud = WordCloud(
                            background_color='white',
                            max_words=200,
                            max_font_size=40,
                            random_state=42
                           ).generate(str(val))
    fig = plt.figure(1)
    plt.imshow(wordcloud)
    plt.axis('off')
    plt.title(title)
    plt.show()

def roc_curve_draw(y_train,y_test,train_predict_proba,test_predict_proba):
      """Method to draw the roc curve for the train and the test date
      This code was reused from KNN assignment
      """
      fpr_train, tpr_train, thresholds_train = roc_curve(y_train,train_predict_proba)
      auc_train =  auc(fpr_train, tpr_train)
      print('\n The auc of train is : ',auc_train)
      fpr_test, tpr_test, thresholds_test = roc_curve(y_test,test_predict_proba)
      auc_test =  auc(fpr_test, tpr_test)
      print('\n The auc of train is : ',auc_test)
      plt.title('Receiver Operating Characteristic')
      plt.plot(fpr_train, tpr_train, 'b', label = 'Train AUC = %0.2f' % auc_train)
      plt.plot(fpr_test, tpr_test, 'g', label = 'Test AUC = %0.2f' % auc_test)
      plt.legend(loc = 'lower right')
      plt.plot([0, 1], [0, 1],'r--')
      plt.xlim([0, 1])
      plt.ylim([0, 1])
      plt.ylabel('True Positive Rate')
      plt.xlabel('False Positive Rate')
      plt.show()
```

[5.2.1] Applying XGBOOST on BOW, SET 1

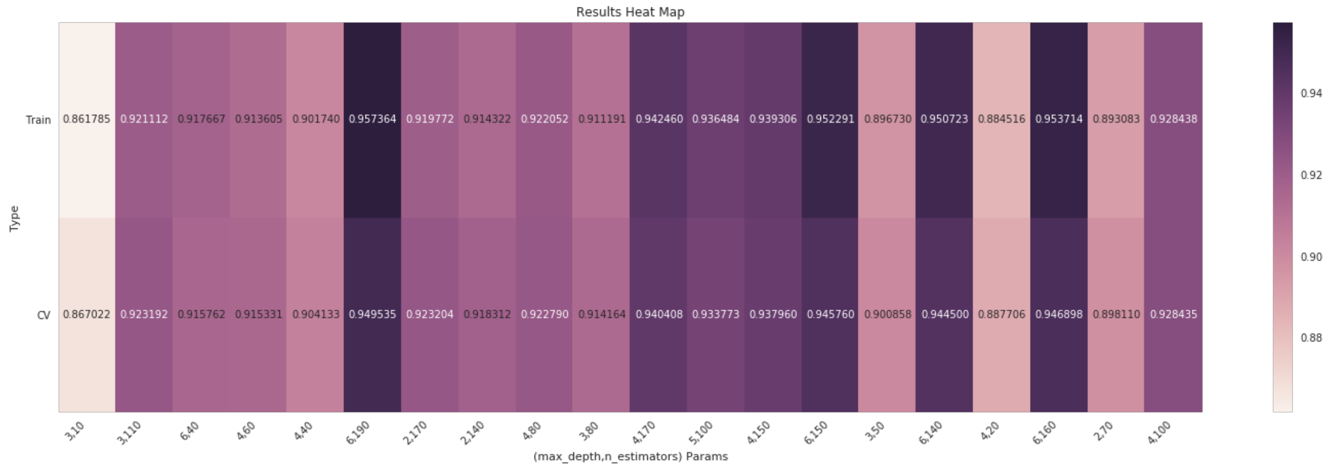
In [0]:

```
best_param_search(training_data,y_train,param_grid,tss)
```

```
Fitting 3 folds for each of 20 candidates, totalling 60 fits
```

```
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed: 17.3min
[Parallel(n_jobs=-1)]: Done  60 out of  60 | elapsed: 30.2min finished
```

Out[0]:



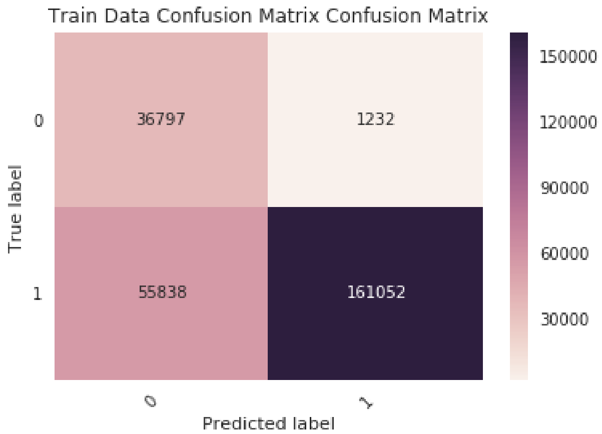
On Test Data

In [0]:

```
xgb = on_test(training_data,y_train,test_data,y_test,6,190)
```

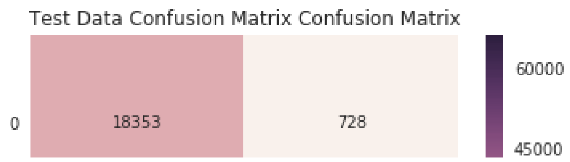
```
The ROC AUC score for the params max_depth 6  and n_estimators  190 is  0.9542202322037244
```

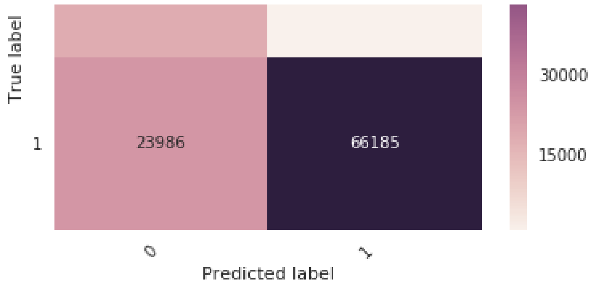
Out[0]:




```
The TPR is :  0.7425515238139149
The TNR is :  0.9676036708827473
The FPR is :  0.032396329117252626
The FNR is :  0.25744847618608513
```

Out[0]:



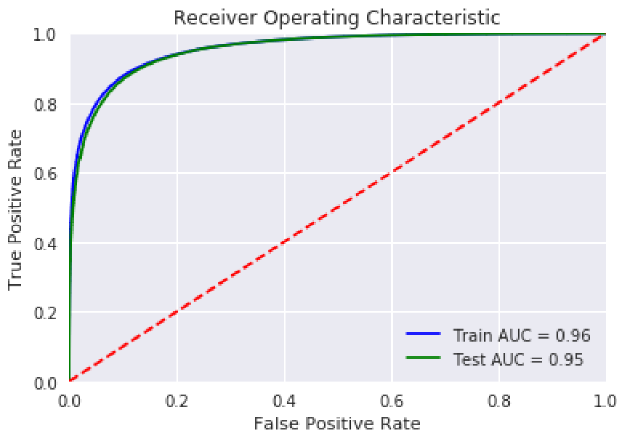



```
The TPR is :  0.733994299719422
The TNR is :  0.9618468633719407
The FPR is :  0.038153136628059324
The FNR is :  0.266005700280578


 The auc of train is :  0.9575131628248775

 The auc of train is :  0.9542202322037244
```

Out[0]:



Wordcloud

In [0]:

```
top_20 = np.argsort(xgb.feature_importances_)[-20:].tolist()
vals  = [feat_names[i] for i in top_20]
Wordcl('BoW Important Features',vals)
```

Out[0]:

[5.2.2] Applying XGBOOST on TFIDF, SET 2

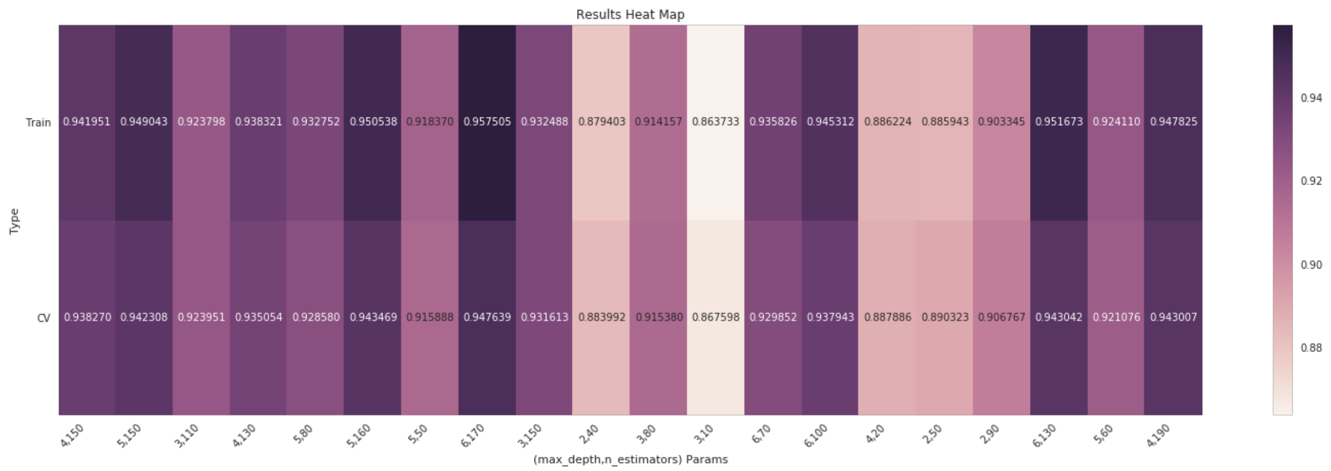
In [0]:

```
best_param_search(training_data,y_train,param_grid,tss)
```

```
Fitting 3 folds for each of 20 candidates, totalling 60 fits
```

```
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed: 20.2min
[Parallel(n_jobs=-1)]: Done  60 out of  60 | elapsed: 32.9min finished
```

Out[0]:



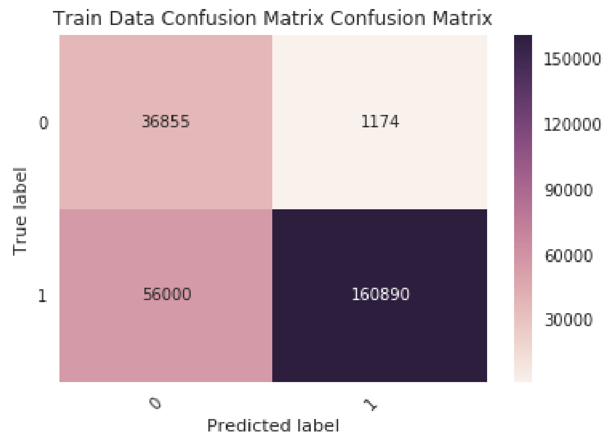
On Test Data

In [0]:

```
xgb = on_test(training_data,y_train,test_data,y_test,6,170)
```

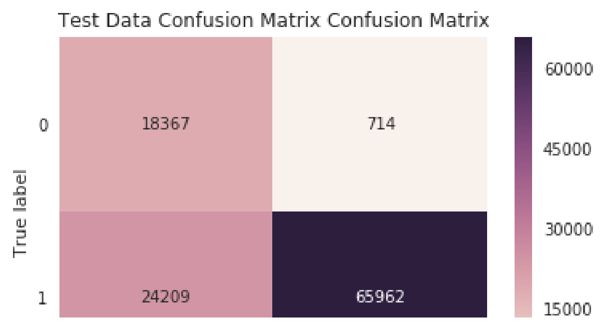
```
The ROC AUC score for the params max_depth 6  and n_estimators  170 is  0.9533927888623748
```

Out[0]:




```
The TPR is :  0.7418046014108535
The TNR is :  0.9691288227405401
The FPR is :  0.030871177259459887
The FNR is :  0.2581953985891466
```

Out[0]:




```
The TPR is :  0.7315212207916071
The TNR is :  0.9625805775378649
The FPR is :  0.03741942246213511
The FNR is :  0.2684787792083929


 The auc of train is :  0.957676035838325

 The auc of train is :  0.9533927888623748
```

Out[0]:

Wordcloud

In [0]:

```
top_20 = np.argsort(xgb.feature_importances_)[-20:].tolist()
vals  = [feat_names[i] for i in top_20]
Wordcl('TFIDF Important Features',vals)
```

Out[0]:

[5.2.3] Applying XGBOOST on AVG W2V, SET 3

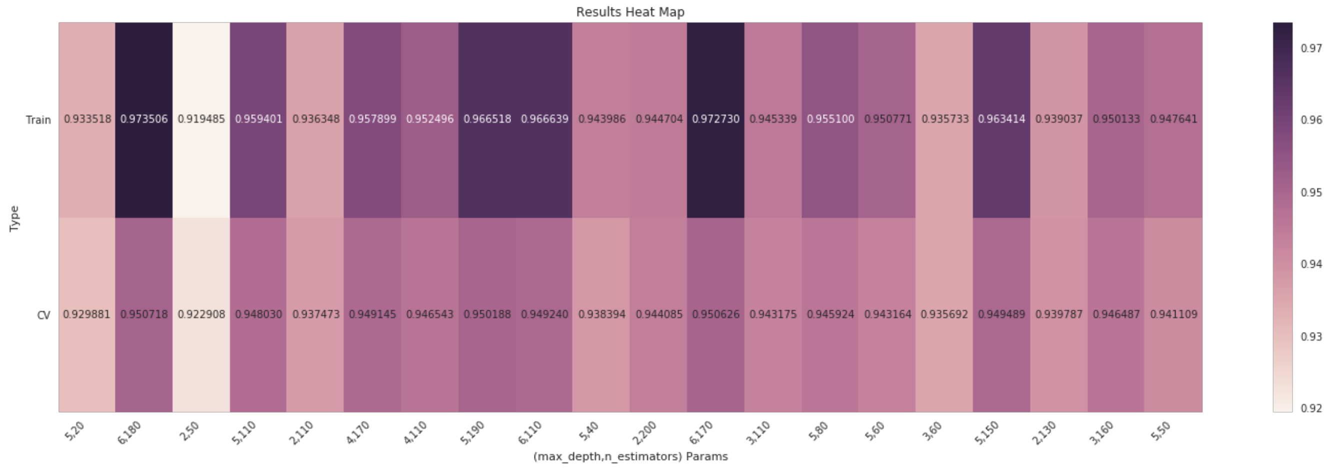
In [0]:

```
best_param_search(training_data,y_train,param_grid,tss)
```

```
Fitting 3 folds for each of 20 candidates, totalling 60 fits
```

```
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed: 23.0min
[Parallel(n_jobs=-1)]: Done  60 out of  60 | elapsed: 35.2min finished
```

Out[0]:



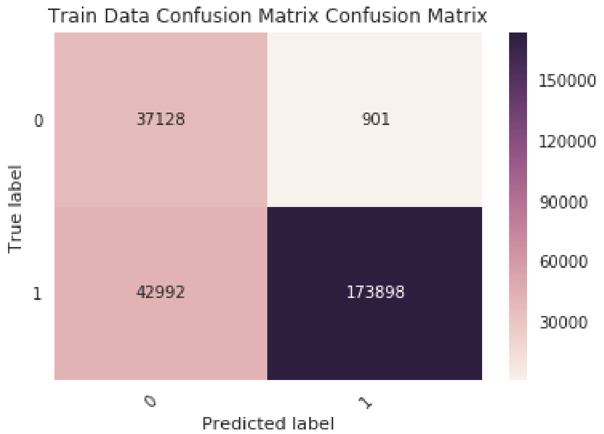
On Test Data

In [0]:

```
xgb = on_test(training_data,y_train,test_data,y_test,6,180)
```

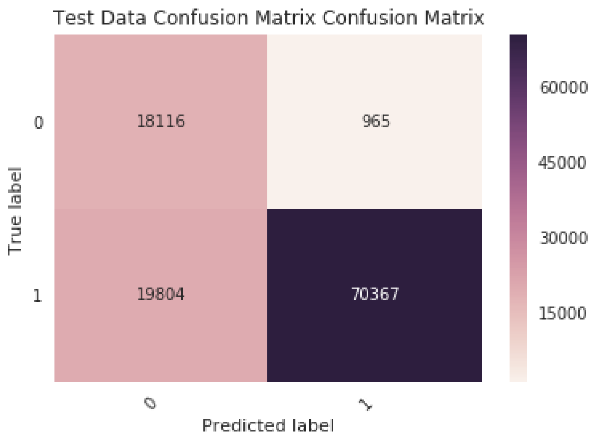
```
The ROC AUC score for the params max_depth 6  and n_estimators  180 is  0.9517887253787127
```

Out[0]:




```
The TPR is :  0.801779703997418
The TNR is :  0.9763075547608404
The FPR is :  0.023692445239159587
The FNR is :  0.19822029600258195
```

Out[0]:



```
The TPR is :  0.7803728471459782
The TNR is :  0.9494261307059378
The FPR is :  0.05057386929406216
The FNR is :  0.2196271528540218


 The auc of train is :  0.9663886730552633

 The auc of train is :  0.9517887253787127
```

Out[0]:

[5.2.4] Applying XGBOOST on TFIDF W2V, SET 4

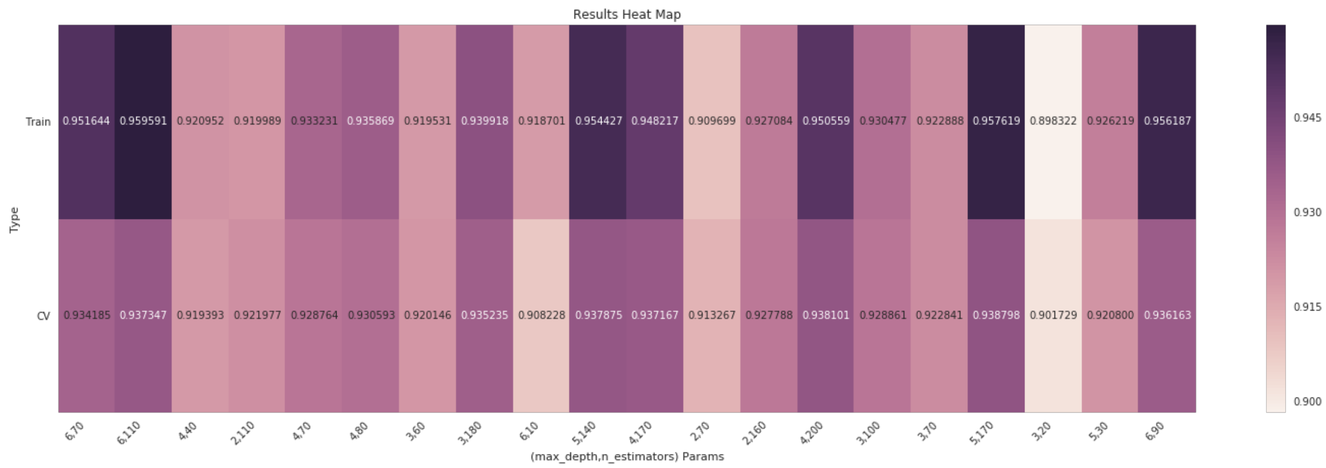
In [0]:

```
best_param_search(training_data,y_train,param_grid,tss)
```

```
Fitting 3 folds for each of 20 candidates, totalling 60 fits
```

```
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed: 16.7min
[Parallel(n_jobs=-1)]: Done  60 out of  60 | elapsed: 28.3min finished
```

Out[0]:



On Test Data

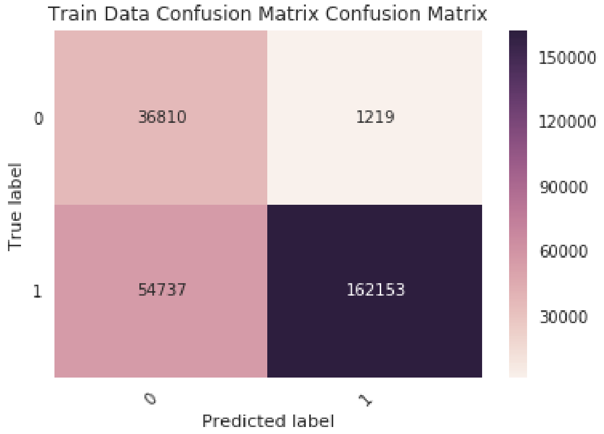
In [0]:

```
xgb = on_test(training_data,y_train,test_data,y_test,5,170)
```

```
The ROC AUC score for the params max_depth 5  and n_estimators  170 is  0.9378780405159435
```

```
/usr/local/envs/py3env/lib/python3.5/site-packages/sklearn/preprocessing/label.py:151:
DeprecationWarning: The truth value of an empty array is ambiguous. Returning False, but in future
this will result in an error. Use `array.size > 0` to check that an array is not empty.
  if diff:
```

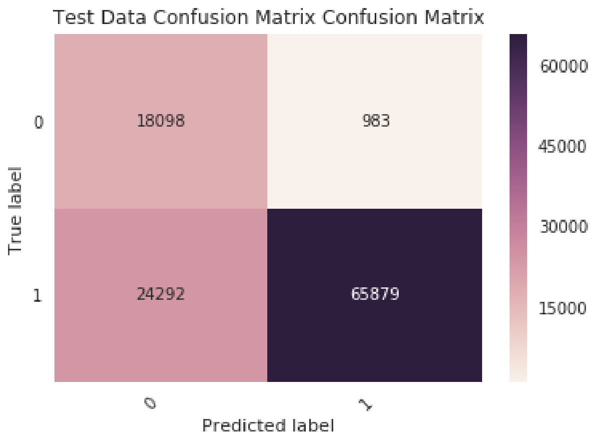
Out[0]:




```
The TPR is :  0.7476278297754622
The TNR is :  0.9679455152646664
The FPR is :  0.03205448473533356
The FNR is :  0.2523721702245378
```

```
/usr/local/envs/py3env/lib/python3.5/site-packages/sklearn/preprocessing/label.py:151:
DeprecationWarning: The truth value of an empty array is ambiguous. Returning False, but in future
this will result in an error. Use `array.size > 0` to check that an array is not empty.
  if diff:
```

Out[0]:

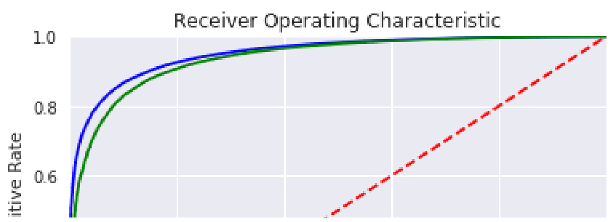


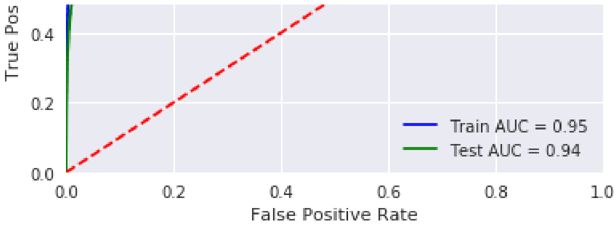

```
The TPR is :  0.7306007474686984
The TNR is :  0.9484827839211781
The FPR is :  0.05151721607882186
The FNR is :  0.26939925253130165
```

```
 The auc of train is :  0.9511554843133205

 The auc of train is :  0.9378780405159435
```

Out[0]:





[6] Conclusions

Model	Max Depth - n_estimators	Train FPR	Train FNR	Test FPR	Test FNR
Bag Of Words	6 - 190	0.0324	0.2574	0.0381	0.2660
TFIDF	6 -170	0.0309	0.2582	0.0374	0.2685
Avg W2V	6 - 180	0.0237	0.1982	0.0506	0.2196
TF-IDF W2V	5 - 170	0.0320	0.0515	0.252	0.269

The best is : Avg W2V with 6 depth and 180 estimators