Project_seconPart

April 12, 2018

0.1 1) Import Libraries

```
In [67]: import os
         import numpy as np
         import pandas as pd
         from profanity import profanity
         from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import StandardScaler
         from sklearn import model_selection
         from sklearn.metrics import classification_report
         from sklearn.metrics import accuracy_score
         from sklearn.linear_model import LogisticRegression
         from sklearn.tree import DecisionTreeClassifier
         from sklearn.neighbors import KNeighborsClassifier
         from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
         from sklearn.naive_bayes import GaussianNB
         from sklearn.neural_network import MLPClassifier
         from sklearn.svm import SVC
```

0.2 2) Import Data Set

```
In [68]: df = pd.read_csv('data_set2.csv')

In [69]: df.shape

Out[69]: (800, 20)

In [70]: df.head()

Out[70]:         Id                                                Title  \
         0   666146                                                  NaN
         1  2752683  "internal_metadta error" when using Blockchain...
         2   675850                                                  NaN
         3  2675420                                                  NaN
         4   817177                                                  NaN

                                                         Text  \
         0  How can I display an image or text whenever I ...
```

```
1  We are using Blockchain as a service on a Blue...
2  I working on a `GIS` application which uses `P...
3  I have a war file with the below structure.\r\...
4  My code has generated the search string `veri_...

                                          Comment Tags  PostTypeId  \
0                          Grammar, condensed text  NaN           1
1  I made some minor edits to grammar and spelling.  NaN           1
2                            Formatting some texts  NaN           1
3                         added missing characters  NaN           1
4              Fixed formatting and broken English  NaN           1

    LastEditDate                                              Title.1  \
0   4/8/13 13:39                            CSS tricky hover effect
1  5/23/17 10:27  "internal_metadta error" when using Blockchain...
2  4/15/13 12:05  How to set InsertCommand parameter as function...
3  8/18/16 13:06           Issue while reading a file from WAR file
4  7/22/13 20:14  How to break this while loop in apache poi get...

                                          Tags.1  Reputation   CreationDate  \
0                                <javascript><css>        8374  8/17/11 17:20
1            <ibm-cloud><blockchain><hyperledger>        1709  1/15/14 15:53
2  <c#><postgresql><ado.net><postgis><dataadapter>         748  5/16/12 10:48
3                                          <java>        1211  8/15/12 19:52
4                 <java><while-loop><apache-poi>        3292  6/18/12 17:57

   Views  UpVotes  DownVotes               WebsiteUrl  \
0   2662     2491        855  http://chrisforrence.com
1    533       52         19                      NaN
2    233      377         55                      NaN
3    110       27          4                      NaN
4   1012     5738        248                      NaN

                   Location  \
0  Atlanta, GA, United States
1                 Austin, TX
2                        NaN
3                        NaN
4                       4444

                                             AboutMe  \
0  <h2>Howdy!</h2>\n\n<p>I'm a software engineer ...
1  <p>I am a Knowledge Manager for IBM Cloud plat...
2                                                NaN
3                                                NaN
4  <p><a href="http://stackoverflow.com/users/146...

              DisplayName   ApprovalDate  RejectionDate
```

```
0            Chris Forrence   4/8/13 13:39           NaN
1  William 'Bill' Wentworth  10/3/16 21:40           NaN
2               Futuregeek   4/15/13 9:14           NaN
3               Ömer Erden  8/18/16 13:06           NaN
4                     4444  7/22/13 18:31           NaN
```

0.3 3)Data Preparation

0.3.1 - Cleaning some features:

In [71]:
```
##Comments' Features:

#There are two types of posts that can be edited
# I use 0 for Editing a question
# I use 1 for editing an answer
df.loc[df['PostTypeId'] == 1, 'PostTypeId'] = 0
df.loc[df['PostTypeId'] == 2, 'PostTypeId'] = 1

#Checks if the post was edited before
df['LastEditDate']=df['LastEditDate'].fillna(0)
df.loc[df['LastEditDate'] != 0, 'LastEditDate'] = 1

#Comments Length
df['CommentLength'] = df['Comment'].apply(len)

#Check if the title of the post was edited
df["TitleChange1"] = df['Title'].fillna('False')
df.loc[df['TitleChange1'] != 'False', 'TitleChange1'] = 'True'

df.loc[df['Title'] == df['Title.1'], "TitleChange2"] = 'True'
df.loc[df['Title'] != df['Title.1'], "TitleChange2"] = 'False'

df.loc[df['TitleChange1'] == 'False', "TitleChange1"] = 0
df.loc[df['TitleChange1'] == 'True', "TitleChange1"] = 1

df.loc[df['TitleChange2'] == 'False', "TitleChange2"] = 0
df.loc[df['TitleChange2'] == 'True', "TitleChange2"] = 1

df['TitleChange'] = df['TitleChange1']^df['TitleChange2']

# check for profanity in the comments and the editions
df['CommentProfanity'] = df['Comment'].apply(lambda x: profanity.contains_profanity(x))
df['Text']=df['Text'].fillna('0')
df['TextProfanity'] = df['Text'].apply(lambda x: profanity.contains_profanity(x))
```

In [72]:
```
##User's Features

#The user has a WebstieURL
df['WebsiteUrl']=df['WebsiteUrl'].fillna(0)
```

```
df.loc[df['WebsiteUrl'] != 0, 'WebsiteUrl'] = 1

#The user stated a Location
df['WebsiteUrl']=df['WebsiteUrl'].fillna(0)
df.loc[df['Location'] != 0, 'Location'] = 1

#the user wrote an AboutMe
df['AboutMe']=df['AboutMe'].fillna(0)
df.loc[df['AboutMe'] != 0, 'AboutMe'] = 1
```

In [73]:
```
## Output
#output 0 notApprove, 1 approve
df['Y'] = df['ApprovalDate'].fillna(0)
df.loc[df['Y'] != 0, 'Y'] = 1
```

0.3.2 - Organizing the data in a new dataframe

In [74]:
```
data = pd.DataFrame()
```

In [75]:
```
##Comments' Features:
#Qestion 0, answer 1
data['PostType']= df['PostTypeId']
# Not Edited before 0, Edited before 1
data['Edited']= df['LastEditDate']
#length of comment
data['LenComment']= df['CommentLength']
#Title Change
data['TitleChange'] = 'Nan'
data.loc[df['TitleChange'] == True, 'TitleChange'] = 1
data.loc[df['TitleChange'] == False, 'TitleChange'] = 0
#CommentProfanity
data['ComProf'] = 'Nan'
data.loc[df['CommentProfanity'] == True, 'ComProf'] = 1
data.loc[df['CommentProfanity'] == False, 'ComProf'] = 0
#TextProfanity
data['TxtProf'] = 'Nan'
data.loc[df['TextProfanity'] == True, 'TxtProf'] = 1
data.loc[df['TextProfanity'] == False, 'TxtProf'] = 0
```

In [76]:
```
##User's Features

#Total Reputation
data['Reputation']= df['Reputation']
#totalUpvotes
data['UpVotes']= df['UpVotes']
#totalDownVotes
data['DownVotes']= df['DownVotes']
#Completion of profile 0 nothing 3 all complete
data['ProfileCompletion'] = df['Location'] + df['AboutMe'] + df['WebsiteUrl']
```

```
In [77]: #output
         data['Output'] = df['Y']

In [78]: data.shape

Out[78]: (800, 11)

In [79]: data.head()
```

Out[79]:

	PostType	Edited	LenComment	TitleChange	ComProf	TxtProf	Reputation	\
0	0	1	23	0	0	0	8374	
1	0	1	48	0	0	0	1709	
2	0	1	21	0	0	0	748	
3	0	1	24	0	0	0	1211	
4	0	1	35	0	0	0	3292	

	UpVotes	DownVotes	ProfileCompletion	Output
0	2491	855	3	1
1	52	19	2	1
2	377	55	1	1
3	27	4	1	1
4	5738	248	2	1

0.4 4) Make a test/train split of the data

```
In [80]: X = data.drop('Output',axis=1)
         Y = data['Output']

In [81]: seed = 7
         X_train, X_test, Y_train, Y_test = train_test_split(X, Y,test_size=0.30, random_state=s
         Y_train=Y_train.astype('int')
         Y_test=Y_test.astype('int')
```

0.5 5) Normalise data

```
In [82]: scaler = StandardScaler()
         scaler.fit(X_train)
         X_train = scaler.transform(X_train)
         X_test = scaler.transform(X_test)
```

0.6 6) Testing many ML algorithms

```
In [83]: # Spot Check Algorithms
         scoring = 'accuracy'
         models = []
         models.append(('LR', LogisticRegression()))
         models.append(('LDA', LinearDiscriminantAnalysis()))
         models.append(('KNN', KNeighborsClassifier()))
         models.append(('CART', DecisionTreeClassifier()))
```

```
        models.append(('NB', GaussianNB()))
        models.append(('SVM', SVC()))
        models.append(('MLP',MLPClassifier(alpha=10,hidden_layer_sizes=(40,40,40),max_iter=1000
        # evaluate each model in turn
        results = []
        names = []
        print('accuracy score')
        for name, model in models:
            kfold = model_selection.KFold(n_splits=100, random_state=seed)
            cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, sco
            results.append(cv_results)
            names.append(name)
            msg = "%s: %f " % (name, cv_results.mean())
            print(msg)
```

```
accuracy score
LR: 0.705000
LDA: 0.694333
KNN: 0.645667
CART: 0.587333
NB: 0.646667
SVM: 0.693667
MLP: 0.694333
```