HW3: BINARY CLASSIFICATION　　SUBMITTED BY: MAYANKA JHA

Problem 1

Downloaded the dataset and loaded it in Jupyter Notebook. I will be using Python to build my classifier models.

Problem 2

a) Build the best classifier you can with the given data, documenting the choices that you make.

I experimented with the given features set and build an initial Logistic Regression classifier using the features, ‘page’ and ‘answer’. The model gave me a 76% accuracy on splitting the given training dataset into train and test for validation purposes. However, the actual test dataset does not contain the ‘answer’ column and also the ‘page’ column which is a categorical variable has different values in both training and test dataset.

Hence, I decided to build my classifier models using ‘body score’, ‘tournaments’, ‘answer_type’ and ‘inlinks’. The following table displays the Accuracy Score and Area Under Curve (auc) obtained from the ROC curve.

Model Name	Accuracy Score	Area Under Curve (auc)
Logistic Regression	66.7	0.69
Adaptive Boosting Classifier	78.0	0.83
Decision Tree	71.7	0.71
Random Forest	77.7	0.83
SVM	78.0	0.82

b) Look at where you’re making mistakes. Can you see any patterns?

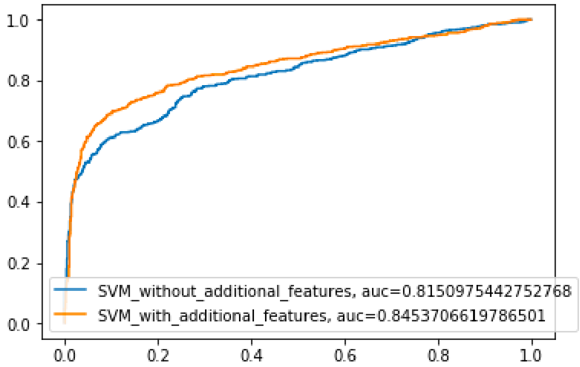
I could have improved the accuracy by computing a new body score using Wikipedia API but due to the limited scope of the assignment I am unable to go into details. Also, since answer column is not present in the test dataset it is not useful for prediction and page values are different in test and training dataset.

Problem 3

I have extracted two additional features from the given dataset:

1. length of the text: The text column represents what part of answer is revealed. Hence, the longer the length of text for each row might indicates a better guess.
2. Tournament year: I’m taking the tournament year to see if there is any relation of the dependent variable to the year the tournament was held on, i.e. does the tournament become more difficult with each year.

a) The following plot displays the area under curve before and after adding two additional features in my model.



b) The additional features have improved the accuracy by 3%. The new accuracy score of my final model is 81% whereas the previous accuracy score was 78%. Please note that I have used SVM to build this model as from the accuracy score table we can observe that SVM and Adaptive Boosting classifiers yields in the best result for our problem statement.

Problem 4

a) Kaggle Score= 0.0 and Kaggle Username: Mayanka Jha
b) Submitted a separate error analysis file as instructed on Piazza.