Question1 (30 points)

1. Name 3 assumptions required for linear regression.
2. Name 3 activation functions used in neural networks.

Question2 (70 points)

Facebook Comments Volume Prediction

In this question, you will be using the dataset provided to predict the comment volume for a given facebook post. Following is an excerpt from the original paper which describes the features used to predict the comments volume for a given facebook post.

Description of the feature set:

1. Page Features: We identified 4 features of this category that includes features that define the popularity/Likes, category, checkin's and talking about of source of document. Page likes: It is a feature that defines users support for specific comments, pictures, wall posts, statuses, or pages. Page Category: This defined the category of source of document eg: Local business or place, brand or product, company or institution, artist, band, entertainment, community etc. Page Checkin's: It is an act of showing presence at particular place and under the category of place, institution pages only. Page Talking About: This is the actual count of users that were ‘ engaged’ and interacting with that Facebook Page. The users who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares by visitors to the page.
2. Essential Features: This includes the pattern of comment on the post in various time intervals w.r.t to the randomly selected base date/time demonstrated in Figure below, named as C1 to C5. C1: Total comment count before selected base date/time. C2: Comment count in last 24 hrs with respect to selected base date/time. C3: Comment count is last 48 hrs to last 24 hrs with respect to base date/time. C4: Comment count in first 24 hrs after publishing the document, but before the selected base date/time. C5: The difference between C2 and C3. Furthermore, we aggregated these features by source and developed some derived features by calculating min, max, average, median and Standard deviation of 5 above mentioned features. So, adding up the 5 essential features and 25 derived essential features, we got 30

features of this category.

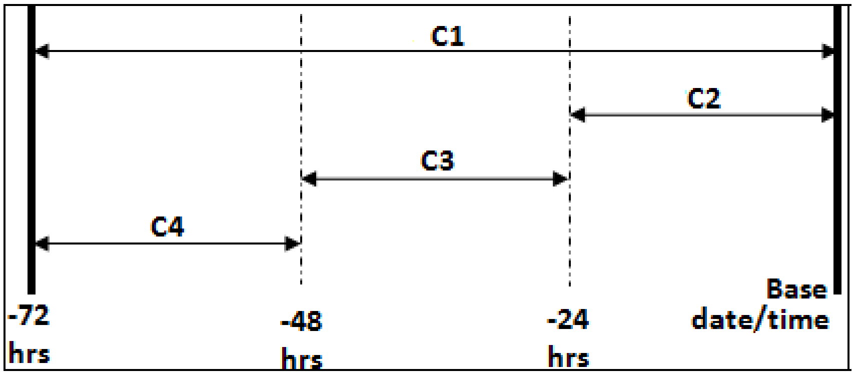

Figure 2. Demonstrating the essential feature details.

3. Weekday Features: Binary indicators (0,1) are used to represent the day on which the post was published and the day on selected base date/time. 14 features of this type are identified.
4. Other Basic Features: This include some document related features like length of document, time gap between selected base date/time and document published date/time ranges from (0,71), document promotion status values (0,1) and post share count. 5 features of this category are identified.
5. Target: Comments received for the facebook post

Answer the below questions in the order provided.

1. Load the dataset. Based on the description provided in the feature set above, create new column names for the data. Update the data to reflect the new column names. Eg: df.columns = ['likes', 'category', 'checkins', ..., 'target']
   Note: the last column is the target variable.
2. Standardize the input data.
3. Split the data into 90% training and 10% test sets.
4. Build a Linear Regressor and find the Mean Squared Error(MSE) and R2 for the test data
5. Build a Decision Tree Regressor and find the Mean Squared Error for the test data and plot the top 10 most important features.
6. Build a GBM **OR** XgBoost Regressor model and find the Mean Squared Error for the test data.
7. What model gives the best results in terms of the MSE?