ML MAJOR PROJECT

1. DESCRIPTION OF THE DATASET

The dataset used(information.csv) was provided in the mail already and all the analysis have been done on the same.

The dataset contains the following fields:

- **unitid**: a unique id for user
- **_golden**: whether the user was included in the gold standard for the model; TRUE or FALSE
- ***unit*state**: state of the observation; one of *finalized* (for contributor-judged) or *golden* (for gold standard observations)
- ***trusted*judgments**: number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations
- ***last*judgment_at**: date and time of last contributor judgment; blank for gold standard observations
- **gender**: one of *male*, *female*, or *brand* (for non-human profiles)
- **gender**:**confidence**: a float representing confidence in the provided gender
- **profile_yn**: "no" here seems to mean that the profile was meant to be part of the dataset but was not available when contributors went to judge it
- **profile_yn**:**confidence**: confidence in the existence/non-existence of the profile
- **created**: date and time when the profile was created
- **description**: the user's profile description
- **fav_number**: number of tweets the user has favorited
- **gender_gold**: if the profile is golden, what is the gender?
- **link_color**: the link color on the profile, as a hex value

- **name**: the user's name
- **profile*yn*gold**: whether the profile y/n value is golden
- **profileimage**: a link to the profile image
- **retweet_count**: number of times the user has retweeted (or possibly, been retweeted)
- **sidebar_color**: color of the profile sidebar, as a hex value
- **text**: text of a random one of the user's tweets
- **tweet_coord**: if the user has location turned on, the coordinates as a string with the format "[*latitude*, *longitude*]"
- **tweet_count**: number of tweets that the user has posted
- **tweet_created**: when the random tweet (in the text column) was created
- **tweet_id**: the tweet id of the random tweet
- **tweet_location**: location of the tweet; seems to not be particularly normalized
- **user_timezone**: the timezone of the user

Here, gender is taken as the dependent variable.

This dataset contains 20,000 rows and 26 columns. Each row containing user name, random tweet, account profile image, location along with link and sidebar color.

2. DATA PROCESSING AND VISUALIZATIONS

'latin-1' encoding is used because in the text and description columns contain special characters like (é,Û,Ï, etc) which do not fit in default UTF-8.

As most of the columns were not relevant for predicting the gender of the user, we will focus on the following features only:

- link_color
- sidebar_color
- text
- description
- gender
- gender:confidence

- About 13926 gender categories which were classified out of 20,050 were with 100% confidence. Hence, in order to predict the gender of the users I would be using only those genders who have 100% confidence level.

- There are 4 categories which exists in the gender feature. We are focusing this project in classifying whether the user is female, male or brand. We are not considering unknown category of gender because it is lesser in count and it doesn't provide much information about the user.

Data exploration has been done on the following features and graphs have been plotted as well:

1. Link_Color and Sidebar_Color Features Exploration

   On Twitter, user can personalize their account by changing the colors of their links or their sidebars.

   Assumption: There is a pattern in which people from different genders personalize their page differently.

   Eg. Female use pink color on the links or the sidebar. And Male use blue color on the links or the sidebar.

   We would be exploring on two features link_color and sidebar_color for three categories of the gender (Male, Female and Brand).

2. Text and Description Features Exploration:

Exploring which words are mostly used by each gender category (Male, Female and Brand). I would be exploring two attributes of the dataset - text from user description and text from tweets.

Description field contains description of the user's profile. This field contains 3744 values as NaN (ie. Not a Number). It would cause problem while trying to clean 'description' field. Therefore changing NaN to blank (ie. '')

3. Data Cleaning

In order to explore the texts, it is necessary to normalize the text. We would follow below steps for cleaning the texts.

1. Normalize the text into lowercase
2. Remove punctuations
3. Remove numbers
4. Remove double spaces
5. Remove hyperlinks

After this we were able to answer few questions on our dataset. For instances:

1. What are the most used words by each gender category on the tweets posted?

Males:

1. just
2. like
3. don't

Females:

1. just
2. like
3. love

Brands:

1. weather
2. channel
3. updates

2. Which color is most used for links and sidebars by each gender category?

Links:

1. Male - Olive
2. Female - Amethyst
3. Brand - Curoius Blue

Sidebars:

1. Male - Woodsmoke
2. Female - Tradewind
3. Brand - Algae Green

4. Data Modeling

We will use machine learning algorithms to predict the user's gender based on color features and text features. For the data modelling phase of the project, we will perform below steps

1. Creating dummy variables for gender attribute in the dataset.
2. Splitting data into training and test set by using 70:30
3. I will run following models on the training data and predicting on the test data

- SVC()
- KNeighborsClassifier()
- LogisticRegression()

4. Evaluate model based on Accuracy.

Evaluating the relationships between link color and user gender, sidebar color and user gender along with graphical representation we found out that all the models yield approximately the same accuracy around 30-42% for both the link_color and sidebar_color

features. 41% is not significant improvement in prediction of user's gender since baseline model is predicting with 38% accuracy.

Hence, we would explore texts of the user to see if the texts are providing better results

Now considering only the description text found collected from the user's profile the accuracy was of the baseline model was found to be 38.88% accurate and the results for description feature used for predicting user gender was that logistic Regression was performing best to predict the gender. Similarly, the Tweets text posted by the user and the combined results were analyzed and the respective prediction accuracies were found.

We councluded that:

1. Logistic Regression is performing best in predicting gender of the users among the three models.
2. Combining texts from tweet and profile description yeilds better accuracy when compared to considering text from each separately.
3. Text feature is better in predicting users' gender when compared to color features