A Report on User Adoption Prediction

By

Name: Vigneshwaran M

Email: vicky12799@gmail.com

Ph.No.:9171989693

Predict future user adoption

- Imported the necessary packages
- The takehome_users.csv dataset is not encoded in utf8 so used latin1 encoding to read the dataset
  users = pd.read_csv("C:/Users/user/Downloads/takehome_users.csv",encoding='latin1')
  df = pd.read_csv("C:/Users/user/Downloads/takehome_user_engagement.csv")
- To remove not adopted users initially removed users which exists less than 3 times in the dataset and stored that in user_id_counts variable.
- And stored those user id in a list (potential_adopted_users)
- Then from potential-adopted_users if the user id login in 3 separate days in seven day period that user id is appended to adopted_users
- By checking length of adopted_users I got to know there are 1656 adopted_users in the dataset
- Created an extra column in users dataframe named adopted. If the user is adopted it will be true else false.
- Then added a new column to specify the email domain and calculated the value_count(). If the value count is more than 5 then that email domain is stored in common_emails and rest all stored as others.
- Checking the value counts of each column to eliminate null and duplicate values.
- Created another column to find whether the user is invited or not invited.
- If the user has last login time, then the user is marked as loggedin else didn't_login
- Using LabelEncoder converted the categorical variables to int numbers
- Tried to figure out the correlation "users.corr()" and plotted that.
- On checking the correlation graph plotted few graphs to compare the adopted and not adopted users where 0 refers to not adopted and 1 refers to adopted
- Created a graph for adopted column and opted to mailing list, enabled for marketing drip, creation source, adjusted email and invited or not invited columns
- For training used sklearn train_test_split model, where
  x = users.drop(['object_id','creation_time','email','name','email_domain','adopted'],axis = 1)
  y = users['adopted']
- Checked whether all column data types are int
- Using LogisticRegression Model tried to predict the values where I got 89.96% accuracy
- Later using confusion matrix plotted the actual and predicted counts of adopted and not adopted users.