Name: Suguru Naresh

Email address: sugurunaresh111@gmail.com

Contact number: +91 8096569188

Anydesk address: **485 090 771**

Years of Work Experience: **3 years**

Date: **12th March 2021**

Self Case Study -1: **TalkingData Mobile User Demographics**

Overview

*** Write an overview of the case study that you are working on. ***(MINIMUM 200 words)*** ***

TalkingData is China's leading third-party data intelligence solution provider which strives to empower enterprises with data-driven digital transformation.In the last seven years, TalkingData's vision of using "big data for smarter business decisions and a better world" has allowed it to gradually become China's leading data intelligence solution provider.

In July 2016 TalkData released a Kaggle Competition to predict the demographics of a Mobile User by providing Mobile Device details of 70% of daily active users in China to help its clients better understand and interact with their audiences.

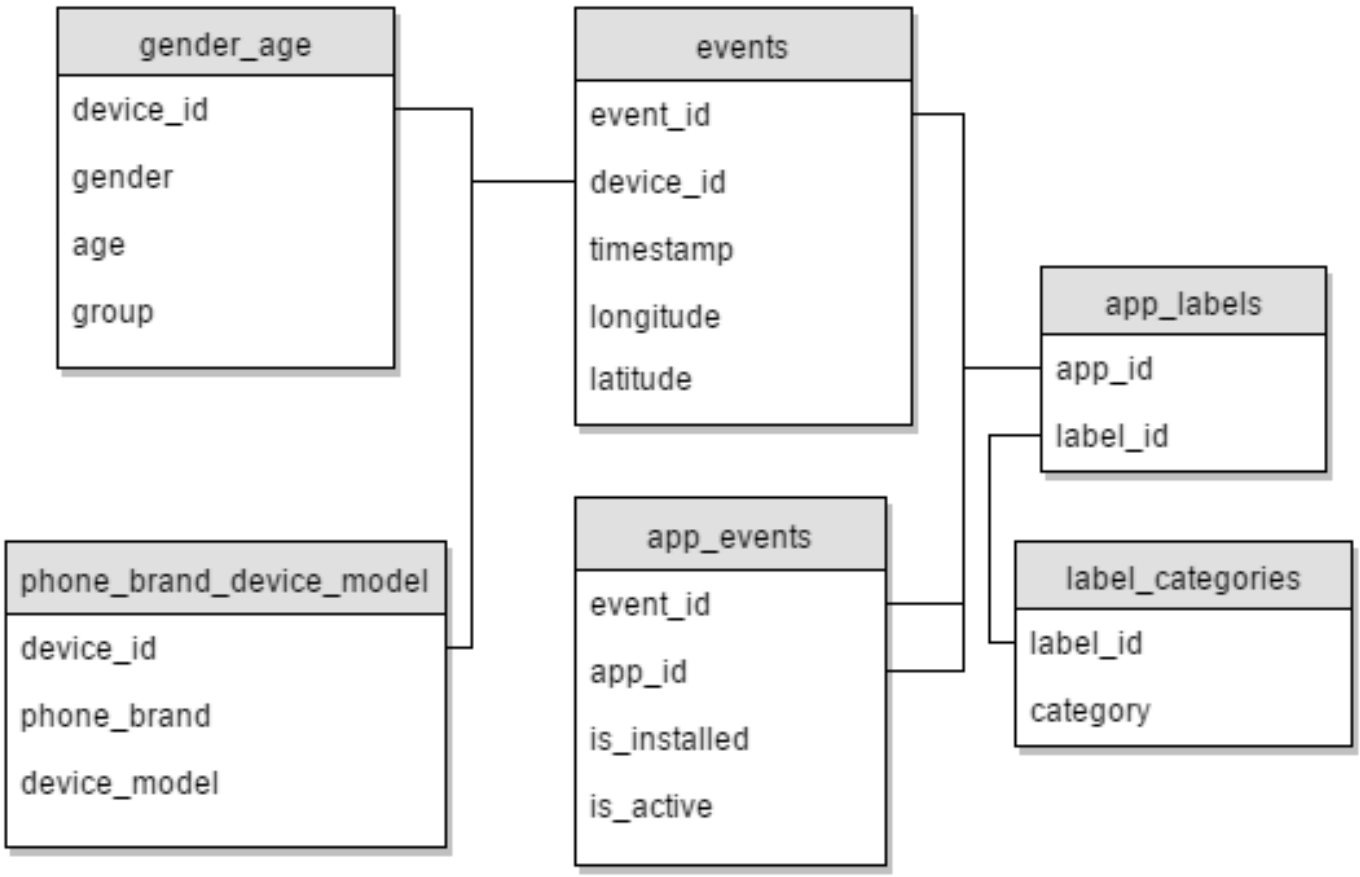
In this competition, Kagglers are challenged to build a model predicting users' demographic characteristics (Age & Gender) based on their app usage, Geo-location, and mobile device properties. Doing so will help millions of developers and brand advertisers around the world pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.

ML Formulation:

After doing EDA and Preprocessing the available training Data a Supervised machine learning model or a deep learning model can be applied on best features of the training data to predict the demographics of mobile users.

Data columns/featuires Overview:

Training Data:



- Device_ID: Dataytpe=Integer.Unique number given to the User's Mobile
- Gender: Dataytpe=Categorical. Gender of the User ,M for Male & F for Female
- Age: Dataytpe=Integer.Age of the user
- Group: Dataytpe=Categorical. This is the Target class for our problem and contains the classes which we need to predict. The first letter denotes the gender of the user and it is followed by the age group to which the user belongs.
- Event_ID: Dataytpe=Integer. Unique number given to the Event
- Label_ID: Dataytpe=Integer. Unique number given to the Label
- Phone_Brand: Dataytpe=Categorical. Brand name of the Mobile
- Device_Model: Dataytpe=Categorical. Model name of the device with in the Brand
- Time_Stamp: Dataytpe=Integer. Time of the event happening
- Latitude-Longitude: Dataytpe=Integer. geo-location where event is happening
- App_ID: Dataytpe=Integer. Unique number given to the Label
- Is_Installed: Dataytpe=Categorical. Is the Application installed?
  0-not installed, 1-Installed
- Is_Active: Dataytpe=Categorical. Is the Application active

  0-Not Active, 1-Active

- Category: Dataytpe=Categorical. Category of application like finance, sports etc.

Performance matric:

Multi Class Logarithmic -Loss: Each device has been labelled with one true class. For each device, we have to predict a set of predicted probabilities (one for each class). The formula is

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of devices in the test set, M is the number of class labels, log is the natural logarithm, yij is 1 if device i belongs to class j and 0 otherwise, and pij is the predicted probability that observation ii belongs to class j.

Research-Papers/Solutions/Architectures/Kernels

*** Mention the urls of existing research-papers/solutions/kernels on your problem statement and in your own words write a detailed summary for each one of them. If needed you can include images or explain with your own diagrams. it is mandatory to write a brief description about that paper. Without understanding of the resource please don't mention it***

1.**You Are What Apps You Use: Demographic Prediction Based on User's Apps**

Eric Malmi Verto Analytics and Aalto University Espoo, Finland eric.malmi@aalto.fi
Ingmar Weber Qatar Computing Research Institute Doha, Qatar iweber@qf.org.qa

The above paper is all about understanding user demographics (gender, income etc.) based on installed apps.
The dataset is one hot encoded and logistic regression is applied on it since it is a multi-class problem.
So in similar fashion we shall onehot encode the dataset and use Logistic regression for prediction

2.A Machine Learning Approach to Demographic Prediction using Geohashes

Avipsa Roy Institute for Geoinformatics University of Muenster a roy001@uni-muenster.de
Edzer Pebesma Institute for Geoinformatics University of Muenster edzer.pebesma@uni-muenster.de

The above paper tried to predict the user demographics with geolocation data .Here the Latitude and longitude of the geo location is converted to strings with Geohashing technique.

And models are applied on it to achieve the results

Here, we can use the geohashing technique on the Geolocation features.

3. **Kaggle Solution:**

https://www.kaggle.com/c/talkingdata-mobile-user-demographics/discussion/23424:

As per the mentioned solution by Yiyun Chen Since approx 70% data doesn't contain event data ,we can have two models one for data with events and other for data without events .

1NN and 1XGB is used as Model1 for training on data without events with phone brand and phone model as features.

1 NN with 1024*128 (two hidden layers) and dropout 0.8, 0.5, activation sigmoid is used as model2 for training on data with events with below features

- one hot of app, app labels, hours, day of the week ;
- entropy of app, app label, hours, day of the week ;
- count of app, app labels ;
- average of non-zero longitude and latitude
  (if all the longitude, latitude of a device is 0,0,then mark it in a binary column called 'missing longitude latitude') ;
- phone brand and phone model;

First Cut Approach

*** Explain in steps about how you want to approach this problem and the initial experiments that you want to do. ***(MINIMUM 200 words)*** ***

*** When you are doing the basic EDA and building the First Cut Approach you should not refer any blogs or papers ***

I would like to experiment and implement the kaggle solution mentioned along with Logistic regression model approach.I shall try using tfidf features using hot encoding.We can also convert the longitude and latitude data to strings with geohashing technique and use it as a feature.

Notes when you build your final notebook:

1. You should not train any model either it can be a ML model or DL model or Countvectorizer or even simple StandardScalar
2. You should not read train data files
3. The function1 takes only one argument "X" (a single data points i.e 1*d feature) and the inside the function you will preprocess data point similar to the process you did while you featurize your train data
    a. Ex: consider you are doing taxi demand prediction case study (problem definition: given a time and location predict the number of pickups that can happen)
    b. so in your final notebook, you need to pass only those two values
    c. def final(X):
        preprocess data i.e data cleaning, filling missing values etc
        compute features based on this X

use pre trained model
return predicted outputs
final([time, location])

d. in the instructions, we have mentioned two functions one with original values and one without it
e. final([time, location]) # in this function you need to return the predictions, no need to compute the metric
f. final(set of [time, location] values, corresponding Y values) # when you pass the Y values, we can compute the error metric(Y, y_predict)

4. After you have preprocessed the data point you will featurize it, with the help of trained vectorizers or methods you have followed for your train data
5. Assume this function is like you are productionizing the best model you have built, you need to measure the time for predicting and report the time. Make sure you keep the time as low as possible
6. Check this live session: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4148/hands-on-live-session-deploy-an-ml-model-using-apis-on-aws/5/module-5-feature-engineering-productionization-and-deployment-of-ml-models