Capstone Project Milestone Report

**Predicting Mobile User Demographics of TalkingData**

Adarsh Srinivas

Mentor: Hobson Lane

**Problem to be solved and Motivation:**

Demographics are widely used in marketing to characterize different types of customers. However, in practice, demographic information such as age, gender, and location is usually unavailable due to privacy and other reasons. The task of this challenge is to predict users’ demographics (age and gender) based on their app download, usage, geolocation, and mobile device properties. A postpaid mobile user is required to create an account by providing detailed demographic information (e.g., name, age, gender, etc.). However, a recent report indicates that there is still a large portion of prepaid users (also commonly referred to as pay-as-you-go) who are required to purchase credit in advance of service use. Statistics show that 95% of mobile users in India are prepaid, 80% in Latin America, 70% in China, 65% in Europe, and 33% in the United States. Even in the U.S., the switch to prepaid plans is accelerating during the economic recession from 2008. Prepaid services allow the users to be anonymous—no need to provide any user-specific information. However, building demographic profiles for all customers is critical to mobile service providers. This can help them make better marketing strategies (e.g., identify potential customers and prevent customer churns). Moreover, by using demographic information, service providers can supply users with more personalized services and focus on enhancing the communication experience. Demographic prediction is important for many applications, such as recommendation, personalization and behavioral targeting. 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.

**Client:**

**TalkingData** is seeking to leverage behavioral data from more than 70% of the 500 million mobile devices active daily in China to help its clients better understand and interact with their audiences.

Also, one of my friend’s relative is a founder of a similar mobile data service providing company in the Silicon Valley. They are looking to solve some similar problems and using my algorithms developed for TalkingData, I hope to present to him my findings and see if they can be applied to solve some of their business problems. The same algorithms or after certain tweaks to the algorithms, the findings can be used by his company in their marketing efforts, personalizing experience of their users, recommendations and so on.

**Data:**

The data will be obtained from the ongoing Kaggle Competition.

<https://www.kaggle.com/c/talkingdata-mobile-user-demographics/data>

**Important field and Information in the Dataset:**

The dataset provided by kaggle consists of seven different csv files. These are events.csv, app\_events.csv, app\_labels.csv, label\_categories.csv, phone\_brand\_device\_model.csv, gender\_age\_train.csv and gender\_age\_test.csv. A detailed description of the various data files is given below:

* **gender\_age\_train.csv, gender\_age\_test.csv** - The training and test set
  + group:This is the target variable that I am going to predict
* **events.csv, app\_events.csv** - When a user uses TalkingData SDK, the event gets logged in this data. Each event has an event id, location (lat/long), and the event corresponds to a list of apps in app\_events.
  + timestamp: When the user is using an app with TalkingData SDK
* **app\_labels.csv** - Apps and their labels, the label\_id's can be used to join with label\_categories
* **label\_categories.csv** - Apps' labels and their categories in text
* **phone\_brand\_device\_model.csv** - Consists of device ids, brand, and models

The objective of this project is to predict the demographics i.e. the age group and gender for each device\_id. Some of the important features given in the dataset that can be used for this prediction are location (longitude and latitude) and timestamp in the events.csv. The category that a particular app belongs to from label\_categories can also be used. Other variables like phone\_brand and device\_model from phone\_brand\_device\_model.csv will also be helpful. As we see, there are already a lot of meaningful features given in the dataset that may be used to achieve our goal.

**Data Wrangling and Cleaning:**

The main challenge is that the data given by Kaggle come in the form of different csv files. In order to use it to build models, we need to perform merge operations as well as other manipulations. I have performed a number of merge operations to get the dataset in a format that can be used for predictions. This involves combining the app\_labels and the label\_categories files, merging app\_events and events files to get a dataset with only active users, joining the active users file with the label\_categories to get the device\_id along with the different useful features that can be used for building the models. Secondly, the phone\_brand names in the phone\_brand\_device\_model dataset were in chinese. I have created a pandas dictionary to map and translate these chinese brand names to english for sake of easy visualizations and data exploration. Some other important features were created, such as, extracting date, month, year, hour and minute from the timestamp as those could be significant features in predicting the gender and age of users. I have also performed group by operations and manipulations to find the maximum count of apps and categories for each device. These additional features would be used in training our model to obtain accurate predictions.

No additional sources or datasets are used to build models and do predictions. This is because this competition is hosted by a company TalkingData and they want the participants to use their company provided data to obtain accurate results.

**Preliminary Findings and Observations:**

There are 74,645 users in the training data set of Talking Data. Among these, 47,904 are males (around 64%) and 26,741 are females (around 36%).

The mean, median and mode for the variable Age are not equal as well as since the skewness and kurtosis are well over our accepted range of -1 to +1, we can say that the variable age is not normally distributed. Age group of 20 - 40 are the dominant age as we see a high peak for that age category. This means most of (maximum) Talkingdata's users fall in this age category. We also see that, females at old age are more active (use more mobile devices) than males at the same age.

Most of the talking data users are Males in the age group 23-26. The next age group is Males 32-38 with the difference between the two groups not significant. Females in the age group 27-28 are the least users of Talkingdata.

Creating a new feature, ‘hour’ from the timestamp variable in the initial dataset, we notice that hours 10 and 21 have the maximum number of events recorded in the talking data dataset. Hours 10 and 21 correspond to 10:00 AM and 9:00 PM. Such a visualisation makes sense as users are more likely to be active at 10:00 AM when they travel for work or start their day. Also, 9:00 PM represents a time after dinner and users are more likely to use their mobile and browse apps as they unwind and relax just before sleep.

Maximum events are recorded on a tuesday followed by a thursday. This is an interesting insight as you would assume maximum events to be occurring on weekends. This is further surprising as we see that the weekends (Saturday and Sunday) have the least number of events.

Filtering out the devices that are active, we find that there are 60669 active apps on devices, we notice that the maximum number of app on a single device is 1342. Finally, we observe that 75% of devices contain at most 21 apps.

Performing some data manipulation to get the number of categories for each device, we can see that 75% of devices have at most 45 categories. Maximum categories per device which are basically the outliers has 332 categories. Each user has an average of 31 categories on their device.

Performing a merge operation to see the popular phone brands of all the users, we observe that the top 5 brands of the Talkingdata users are Xiaomi, OPPO, Vivo, Samsung and Huawei. We can further see that the number of males in the top three phone brands i.e. Xiaomi, Samsung and Huawei are twice as much as the number of females for these brands.

**Further Approach:**

Our initial exploration and findings is concurrent with our claims that the features given in the dataset as well as the new features obtained through feature engineering would help us achieving good accurate results in predicting the demographics of the TalkingData users. The next step is to build machine learning models and obtain predictions on our test set. The first machine learning algorithm I intend to apply is the Stochastic Gradient Descent (SGD) classifier. Following the scikit learn algorithm cheat sheet provided by my mentor, Mr. Hobson, this is the classifier that the sheet tells us me to start with according to the dataset provided by Kaggle. Once the model is built using this algorithm, we can try to improve further trying out some other algorithms to see which model gives us the best results. Finally, we can upload the sample submission file based on our best model to Kaggle as part of the competition.