Determining an individual's gender based on what mobile app they use, when they use and how long they use it for in a single day using Machine Learning Algorithms.

by
Sayanti Chatterjee
Ranit Bhowmick

Keywords

Machine Learning, Decision Tree Classifiers, Random Forest, Survey, App Usage

Abstract

A survey was conducted which asked people about their daily app usage. Other data like gender, age, professional status and field was collected. The dataset was analyzed, refined, processed and two machine learning models - Decision Tree Classifier and Random Forest were trained using the app usage data and gender identity of each specific individual. The model with the high success rate was taken into account.

Introduction

Mobile applications nowadays come packed with APIs in order to show ads to the individual using them. This allows app developers to easily monetize and adopt an ad based, free marketing model for their application.
These APIs track various personal information about the user which is then analyzed and processed in order to create shadow profiles even if they haven't created an account or opted to provide any information.
These profiles help advertisers to better target interest based ads and other promotional contents to the user.

This article focuses on those cases where information like gender, age, professional status and working field are not readily available to track or are blocked by some unknown means.
In such scenarios, the App usage data of the user in that device can be tracked, collected and analyzed in order to calculate or find out the above mentioned information using a machine learning model.

In our experiment, we try to analyze and calculate the gender identity of a user using the app usage information from their phone.

Requirements, Libraries and experimental methods

The Survey :

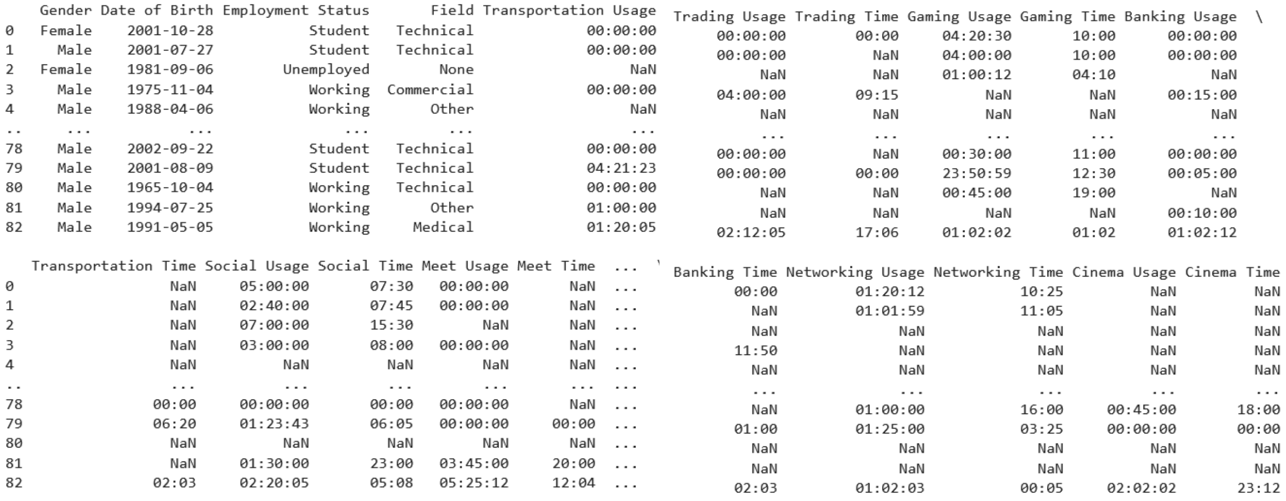
A google survey form was created (https://forms.gle/uytVGGCjGfRFqp7P9) in order to collect information about an user's age, gender, employment status, field of working and App usage data.
The survey questions on usage data is divided into 10 different sections for different type of mobile apps e.g. Transportation apps like uber, ola ; Video Gaming apps like Temple Run, Candy Crush ; Social Apps like Facebook, instagram etc. Each section asks the user for the total time duration they use the said app for in a single day and the approximate time of the day they start using the app e.g. 8:30am, 6:20pm, etc.

The survey data is then exported in a *.csv format for preprocessing and further use.

The Dataset :

The dataset contains information from 10 different categories of Android apps e.g. Transportation apps like uber, ola ; Video Gaming apps like Temple Run, Candy Crush ; Social Apps like Facebook, instagram etc. Each category has two columns underneath it for the total time duration an individual uses the said app for in a single day and the approximate time of the day they start using the app e.g. 8:30am, 6:20pm. Along with this, the dataset has the users gender, age, employment status and field of working information.

Here is a condensed version of the dataset :



```
    Gender Date of Birth Employment Status       Field Transportation Usage  Trading Usage Trading Time Gaming Usage Gaming Time Banking Usage  \
0   Female    2001-10-28           Student   Technical             00:00:00       00:00:00        00:00     04:20:30       10:00      00:00:00
1     Male    2001-07-27           Student   Technical             00:00:00       00:00:00          NaN     04:00:00       10:00      00:00:00
2   Female    1981-09-06        Unemployed        None                  NaN            NaN          NaN     01:00:12       04:10           NaN
3     Male    1975-11-04           Working  Commercial             00:00:00       04:00:00        09:15          NaN         NaN      00:15:00
4     Male    1988-04-06           Working       Other                  NaN            NaN          NaN          NaN         NaN           NaN
..     ...           ...               ...         ...                  ...            ...          ...          ...         ...           ...
78    Male    2002-09-22           Student   Technical             00:00:00       00:00:00          NaN     00:30:00       11:00      00:00:00
79    Male    2001-08-09           Student   Technical             04:21:23       00:00:00        00:00     23:50:59       12:30      00:05:00
80    Male    1965-10-04           Working   Technical             00:00:00            NaN          NaN     00:45:00       19:00           NaN
81    Male    1994-07-25           Working       Other             01:00:00            NaN          NaN          NaN         NaN      00:10:00
82    Male    1991-05-05           Working     Medical             01:20:05       02:12:05        17:06     01:02:02       01:02      01:02:12

   Transportation Time Social Usage Social Time Meet Usage Meet Time  ...  \ Banking Time Networking Usage Networking Time Cinema Usage Cinema Time
0                  NaN     05:00:00       07:30   00:00:00       NaN  ...         00:00         01:20:12           10:25          NaN         NaN
1                  NaN     02:40:00       07:45   00:00:00       NaN  ...           NaN         01:01:59           11:05          NaN         NaN
2                  NaN     07:00:00       15:30        NaN       NaN  ...           NaN              NaN             NaN          NaN         NaN
3                  NaN     03:00:00       08:00   00:00:00       NaN  ...         11:50              NaN             NaN          NaN         NaN
4                  NaN          NaN         NaN        NaN       NaN  ...           NaN              NaN             NaN          NaN         NaN
..                 ...          ...         ...        ...       ...  ...           ...              ...             ...          ...         ...
78               00:00     00:00:00       00:00   00:00:00       NaN  ...           NaN         01:00:00           16:00     00:45:00       18:00
79               06:20     01:23:43       06:05   00:00:00     00:00  ...         01:00         01:25:00           03:25     00:00:00       00:00
80                 NaN          NaN         NaN        NaN       NaN  ...           NaN              NaN             NaN          NaN         NaN
81                 NaN     01:30:00       23:00   03:45:00     20:00  ...           NaN              NaN             NaN          NaN         NaN
82               02:03     02:20:05       05:08   05:25:12     12:04  ...         02:03         01:02:03           00:05     02:02:02       23:12
```

Libraries :

- *Numpy* : Numpy is used to easily create and handle arrays inside of python and as a dependency of Pandas.
- *Pandas* : Pandas is a data analysis and manipulation tool which is used to read the csv files and manipulate, process and present the data to the machine learning algorithm.
- *Sklearn* : Sklearn offers efficient tools for predictive data analysis. It is used in this project to train, test and score the machine learning model.
- *datetime :* Datetime library is used to easily manipulate and access the date and time data in our dataset.

Data preprocessing :

All the time duration data in the dataset are scanned for any possible outliers and are eliminated. Since the duration a person uses an app cannot be greater than 24 hours in a day, it is recognised as an human typing error and replaced by 0 instead.
As the next step, we fill all the NaN values in all the time duration columns by 0, since we had instructed the survey takers to leave inputs blank if their answer was 0.

In the Pre Processing stage, we convert all the time duration values from type string to panda datetime which is then converted to total seconds in the next step. This value is in integer format and can easily be used to train the model in sklearn.

All of these steps are repeated for the App usage 'time' columns in the dataset as well as the Date of birth column which is converted to float type : total number of days by subtracting it from the present date in order to find out the age of the Individual.

As the last step, the gender, employment status and field of the working column is processed and are 'one hot encoded' into their different categorical columns since they are all nominal variables.

Train test split :

The dataset is split in two in the ratio 0.2 where we use 80% of the original Dataset for training and 20% for testing our model.

Training the Model :

- *Decision Tree Classifier :*

We use the training dataset to train a Decision Tree Classifier Model. Followed by the use of a testing set to check and score the trained model.
The highest accuracy we reach using the Decision Tree Classifier is approximately 71%.

- *Random Forest :*

We use the training dataset to train a Random Forest Model with the number of n_estimators being 15. Followed by a testing set to check and score the trained model.
The highest accuracy we reach using Random Forest is approximately 82%

Conclusion

Through this experiment it is evident that it is possible to find patterns in one’s App usage activities in order to determine their gender, employment status and field of working.
We can conclude that a person's app usage and preferences are directly or indirectly affected by their gender, professional background and age.

Links

Google form link to this Survey : https://forms.gle/uytVGGCjGfRFqp7P9
Github : https://github.com/Kawai-Senpai/Info-Through-App-Usage