Converted.in

USA Mobile App

Reported by/ Data scientist: Ahmed Hisham Elsayed

If you want to dig in much deeper in details please find here:

1) **The Code of The analysis and modeling phase (in extension html..etc)**
2) **The code of Getting the Correlations between categorical variables (in extension html..etc)**
3) **The presentation phase**

In this GitHub Repository

Note: it is better to download the files in the repository to be viewed correctly

0) Introduction:

'xxxx' is information technology company built few Apps, has Data related to These Apps they built with existing users Info.

the management and Marketing teams want to get all possible benefits from these data like..

- what is the most OS they need to be focus most?
- how to increase the profit by targeting users?
- what are the relations between variables or how the affect each other?
- how to depend on these data in hiring strategy too?

1) Data:

- Data called: 'convertedin USA Mobile App Data Small Sample. xlsx'
- Contains: Info of

1) 'MAID': ID of the Record/ Row
2) 'AppBundle': The bundle of the APP that the Cx/User subscribed
3) 'AppName': the APP name itself
4) 'Timestamp': date of Transaction
5) 'UserAgent': user agent code/ID that acts as a client/user
6) 'Model': Model of the Phone used
7) 'Operating System': operation System used
8) 'Device's Brand': the handset Brand
9) 'Language': Language Setting like..EN, ES
10) 'Lat': Latitude
11) 'Lon': longtude
12) 'Zip Code'
13) 'City'
14) 'Country'
15) 'IP'
16) 'Carrier'

2) Methods:

I assessed the data using Python packages like pandas/numpy then, I cleaned the data by Python too, then I made an interactive Dashboards/Story in Tableau and get a help using Python language too, , after that I built a Mode using Python packages like Tanserflow.. let's get details

3) Analysis&Results:

- Assessing Data:
  - Data Consist of
  - 30000 rows, 16 columns
  - 12 Categorical column, 2 integer and 2 float
    and there are missing values in 'AppName', 'AppBundle', 'Language', 'IP' and 'Carrier'

- Cleaning Data and Results:
  - No duplicates in this Data

- we shouldn't drop all rows that contain missing data now, because column like language has around 30% missing data, and column like Carrier has around 60% percent missing , so if we made that action we gonna lose more than 50% of our data and this is not acceptable
- so, I followed an approach, like.. fill missing values in 'language' with the common value in that column which is 'en'
- Carrier column has around 60% missing data, it is better to drop it
- so now we can drop any row contains missing data
- now we have 29056 records an 15 columns and no missing data
- 'Timestamp' column here is integer and ofcourse not readable, so it is better to convert it to be datetype
- Now data is cleaned
- I saved this cleaned data frame in an Excel file 'cleaneddf.xlsx'

4) Exploring Data

- I made few graphs and plots to get meaningful Info
  And The where is the most users came from or the most OS used and so on and I found that
  the most users are with:
  - OS: Android&IOS Rocu
  - languages: EN and ES
  - Device Brand:apple, samsung, lg and motorola
  - APP name: TextNow: Free Texting & Calling App
  - Wordscapes
  - iFunny :)
  - Musi - Simple Music Streaming
- I wanted to check the correlation between variables, but most of our data are categorical, so I made a Pearson/Spearman correlation matrix, and found that:

  1) AppBundle has a strong positive relation with.. Operation System, Devices Brand

2) Operation System has strong positive relation with AppBundle, Device Brand & strong negative relation with Model
3) Devices Brand has strong positive relation with .. AppBundle, Operation System
4) Otherwise above... there is no Correlation

5) Modeling:

- we can say that as software company, we can take advantage of having this Data, to predict which OS needs a focus and to be invested most without having any info related to OS nor Model at any future Data.

- apostrophe in column(Device's brand) gonna make issues with me, so i'm gonna change it to be like.. 'Devices Brand' by an approach Find & Replace

- create new columns and eliminate others:
    - create new Features/columns:
      'AppB-N' which is a combination of ['AppBundle'] + ['AppName']
      'DeviceM-B' which is a combination of ['Model'] + ['Devices Brand']
      'Location' which is a combination of ['Country'] + ['City']

    - Elimination:
      Eliminate column ['AppBundle', 'AppName', 'Model', 'Devices Brand', 'Country', 'City'] and 'NormalDate' too because the difference between date in all records is in the 'second' digit, so this will not be important in my model.

- taking a Random Sample:
  I will take a sample with size n = 1000 to build the model easily on my machine.

- Selecting Target:
  i want to set the target of my model to be the operation system, without info about DeviceM-B(device model, device bundle), but this column is categorical, so i'll proceed with Approach "Find and Replace" which mean i'll replace each categorical variable with a number like this {"Apple iOS": 1, "Android": 2, 'Roku OS':3, 'Tizen':4, 'Unknown':0, 'Chrome OS':6, 'Linux': 6}

- selecting "Features" of the model
  data_features = ['Zip Code','IP', 'AppB-N', 'Location', 'Lat', 'Lon']

- Choosing the model && Fix categorical variables:
  - so now, we should start building the model, but, models hate categorical variables, so we have 3 approaches 1) drop categorical columns, but this will decrease of course the accuracy of the model 2) "Label Encoding" for ordinal variable, but our categorical Variables here are nominal ! 3)"One-Hot Encoding" and i will go proceed with this approach

  - mmmm, lt's make a last assess..
    - our data here is:
      1) structured
      2) there is no lack of data or issues like that
      3) supervised

    - it came in mind 'descision tree' but Decision trees leave you with a difficult decision. A deep tree with lots of leaves will

overfit, But a shallow tree with few leaves will perform poorly because it fails to capture as many distinctions in the raw data.

- mmmmm, now it came in mind "Random forest", The random forest uses many trees, and it makes a prediction by averaging the predictions of each component tree. It generally has much better predictive accuracy than a single decision tree and it works well with default parameters

- so I got: MAE from Approach 3 (One-Hot Encoding):

  0.06228

- Conclusion

  from now to the future we can know the OS of the user without any info about mobile model/brand

6) Recommendations:

so,

now i can say that, we need to focus on where our profit come from and from what and where we can get more and more income to our company and check hiring needs as well:

and the most users are with:
OS: Android&IOS Rocu
languages: EN and ES
Device Brand:apple, samsung, lg and motorola
APP name: TextNow: Free Texting & Calling App
Wordscapes
iFunny :)
Musi - Simple Music Streaming

Recommendations:

1) we can increase Advertisment on these APPs that has the alot of users
2) we can compare between Good cities and other Bad cities that didn't reach the KPIs(which is each city should have at least Average count of users) to investigate deeper in Cx behavior befor we invest in Bad cities
3) we now recognized the our most Cx are EN/ES speakers
4) we now know that we need to wide our android and IOS teams more than other teams to cover all updates and issues with most of our Users

to help:

1) **AppBundle** has a strong positive relation with.. Operation System, Devices Brand
2)**Operation System** has strong positive relation with AppBundle, Device Brand & strong negative relation with Model
3)**Devices Brand** has strong positive relation with .. AppBundle, Operation System
4)Otherwise above... there is no Correlation