Capstone Project:

Predicting mobile user demographics of TalkingData

ADARSH SRINIVAS

MENTOR: HOBSON LANE

Problem to be solved and Motivation

- Demographic information such as age, gender, and location is usually unavailable due to privacy and other reasons for mobile data providers.
- Prepaid services allow the users to be anonymous—no need to provide any user-specific information.
- 95% of mobile users in India, 80% in Latin America, 70% in China, 65% in Europe, and 33% in the United States are prepaid.
- These demographic characteristics can be widely used in marketing efforts to characterize different types of customers for mobile service providers worldwide.

Client

- ***TalkingData:*** _TalkingData is one of the largest mobile-service provider's in China. 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.
- ***CEO of LotusFlare, Mr. Sam Gadodia:*** LotusFlare is looking to solve 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 implementing certain tweaks to the algorithms, can be used by his company in their marketing efforts, personalizing experience of their users, recommendations and so on.

Feature Engineering

The features that were used in developing the models were:

- **Phone brand**
- **Device model**
- **App labels**
- **The number of apps installed by each device**

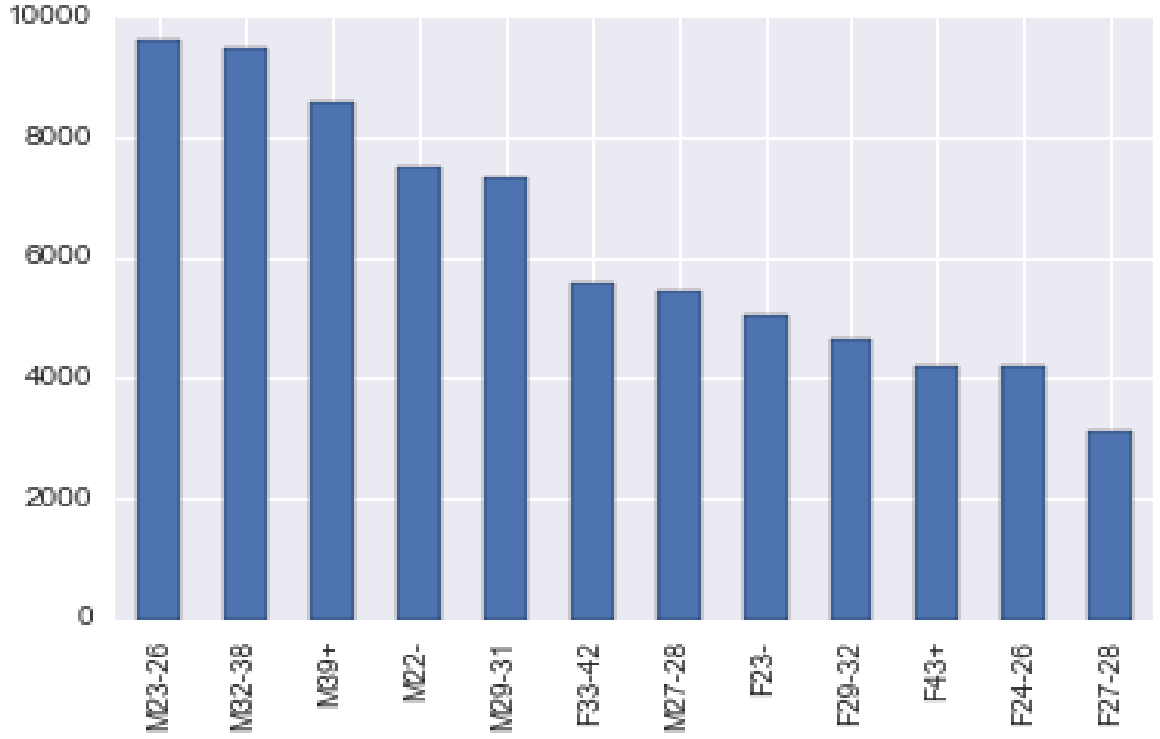
Exploratory Data Analysis

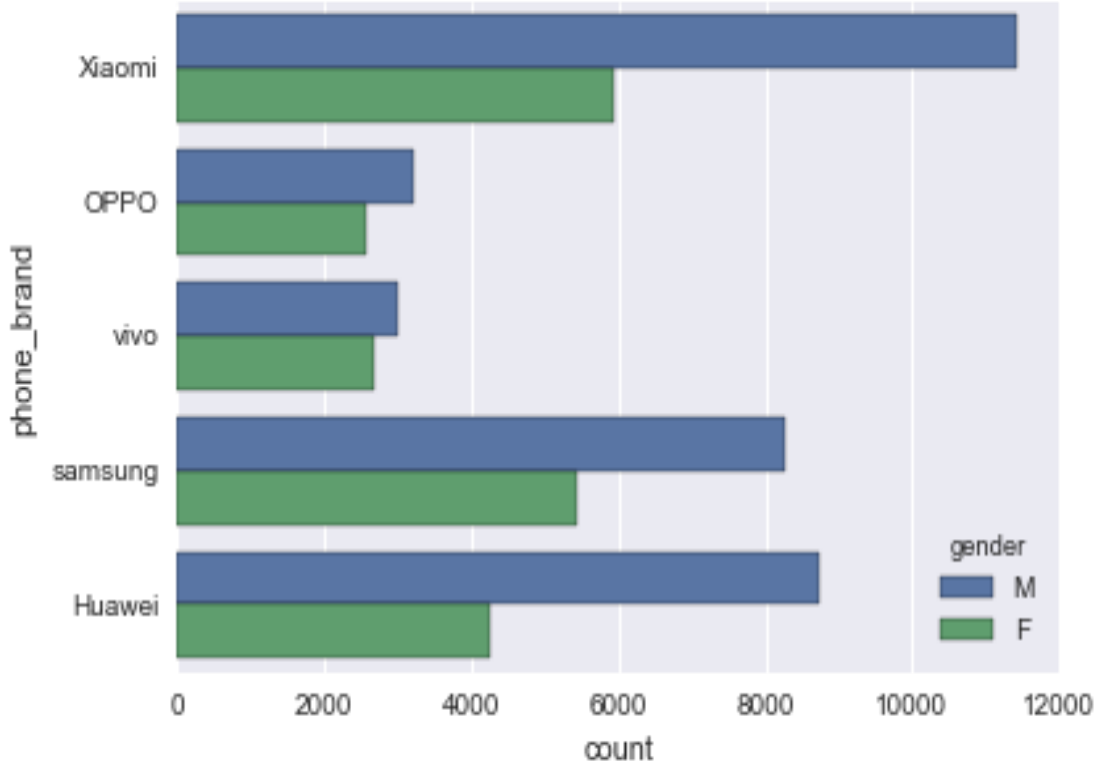
- Most of the talking data users are Males in the age group 23-26.
- Females in the age group 27-28 are the least users of TalkingData.
- Age group of 20 - 40 are the dominant age in the TalkingData dataset.
- Females at old age are more active (use more mobile devices) than males at the same age.

Exploratory Data Analysis

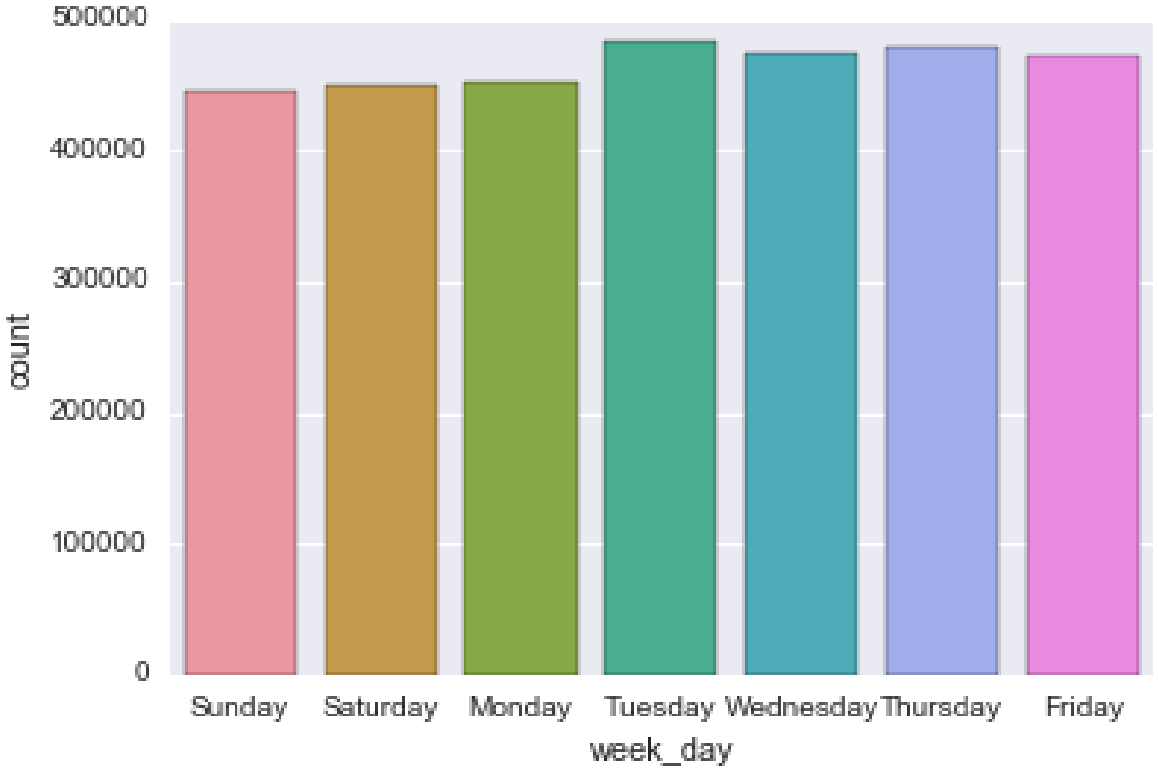
- TalkingData users have the maximum number of events at morning 10:00 AM and night 21:00 PM.
- Maximum events are recorded on a Tuesday followed by a Thursday.
- 75% of devices contain at most 21 apps.
- 75% of devices have at most 45 categories.
- The top 5 brands of TalkingData users are Xiaomi, OPPO, Vivo, Samsung and Huawei.

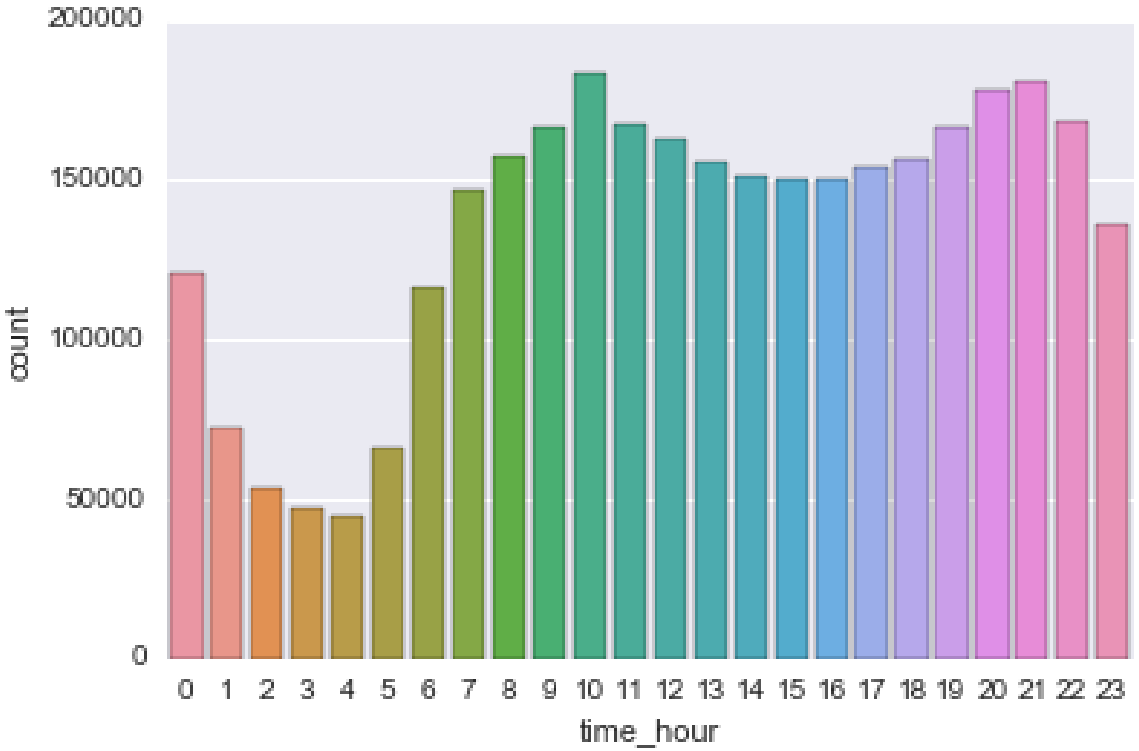
Visualizing TalkingData





Visualizing TalkingData





Algorithms and Results

Models (Algorithms)	Log loss score on Kaggle
Stochastic Gradient Descent (SGD) Classifier	2.40891
Logistic Regression	2.27331
Multinomial Logistic Regression	**2.26713**

Recommendations for the client

- Use this critical demographic information of users to design better marketing strategies (identify potential customers and prevent customer churns)
- Supply users with more personalized services and focus on enhancing the communication experience.
- Built a recommendation engine, implement behavioral targeting based on specific age groups and gender.
- These demographic prediction results 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.
- Produce personalized app selections, mobile data plans more likely to appeal to specific age group users using these machine learning techniques to analyze individual consumption patterns.

Future Research

- **Try non-linear models:** The models that were used in here were all linear models. Non-linear models could be implemented to see if better results can be achieved.

- **New features:** New features could be created to help us generalize better on the test dataset thereby achieving better results.


DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
TINGKI
BİYAN
SHUKRIA
GRACIAS
THANK
SUKSAMA
EXHMET
ARIGATO
YOU
SHUKURIA
GRAZIE
MEHRBANI
PALDIES
GOZAIMASHITA
EFCHARISTO
KOMAPSUMNIDA
JUSPAXAR
BOLZÏN
MERCI