Classified

Machine Learning in Practice, 2017

Team members:
Jordi Beernink
Gerdriaan Mulder
Thijs Werrij
Jeffrey Luppes
Roel Bouman

Outline

1. Introduction
2. Approach
3. Results
4. Other Ideas
5. Future
6. Conclusion

Introduction: Team members and roles

- Jordi Beernink
  - Coordination, pre-processing, XGBoost implementation
- Gerdriaan Mulder
  - Pre-processing, repository manager
- Thijs Werrij
  - Deep learning
- Jeffrey Luppes
  - Classification
- Roel Bouman
  - Pipeline and classification

Introduction: Problem description

- *TalkingData* is the largest third-party mobile data platform
- Seeking to leverage behavioral data for more than 70% of the 500 million mobile devices active daily in China to help its clients better understand

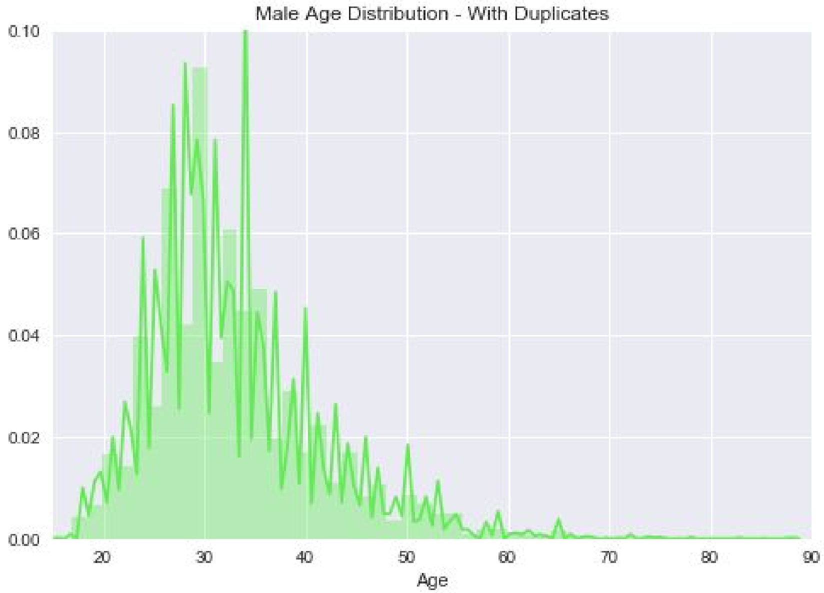
  "*Nothing is more comforting than being greeted by your favorite drink just as you walk through the door of the corner café.*" – TalkingData competition page

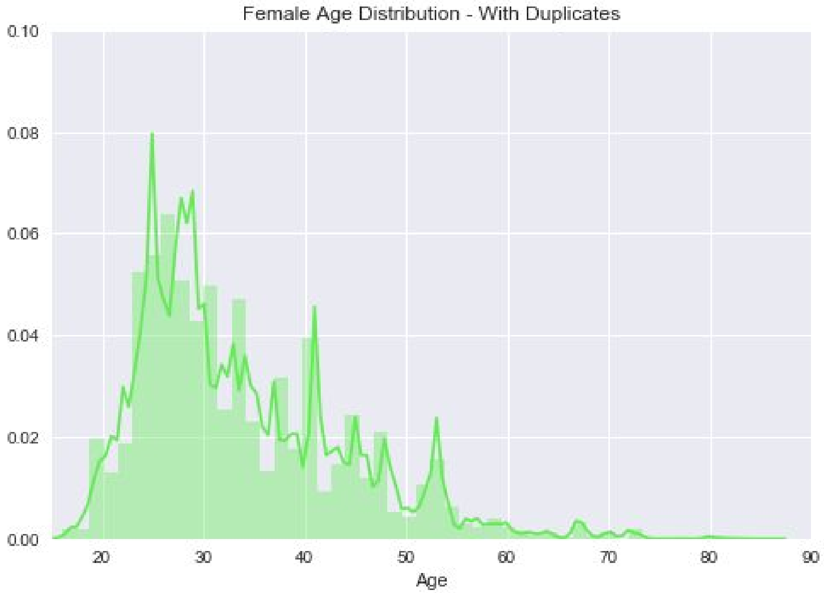
- Predict user demographic based on:
  - application usage, phone brand and location
- Demographic to predict given a device:
  - Gender (male, female)
  - Age class (6 categories per gender, e.g.: 23-, 29-33, 43+ etc.)

Introduction: Dataset

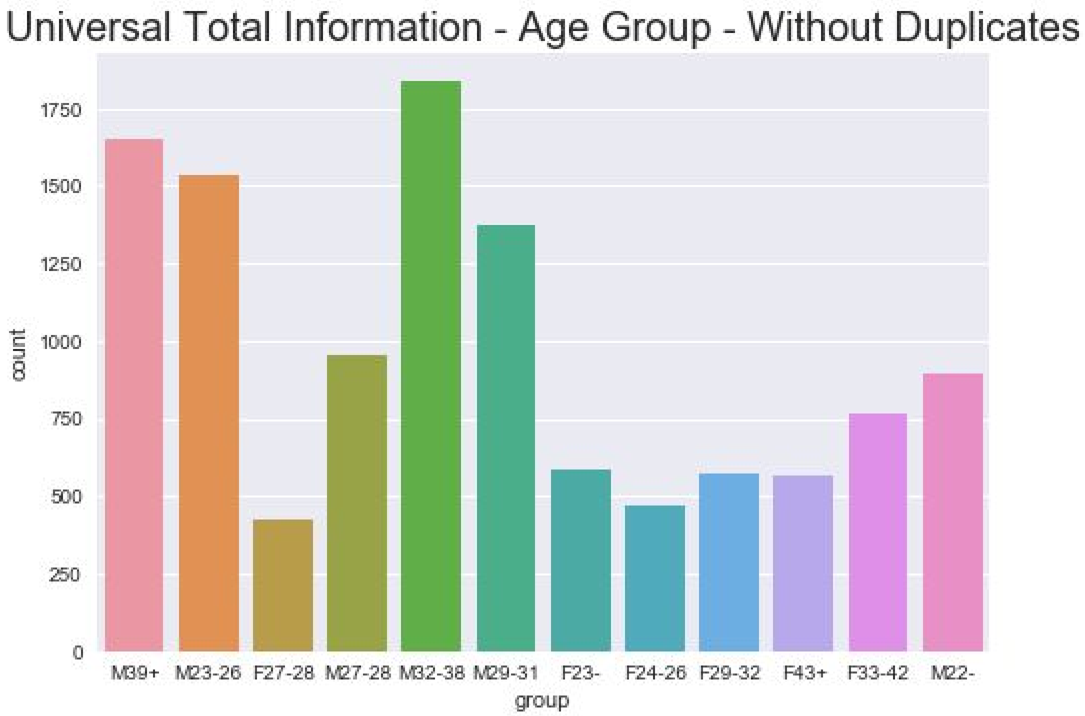
- About 1,2GB of CSV files
- Structure per CSV file:
  - App_events (event_id/app_id/is_installed/is_active)
  - App_labels (app_id/label_id)
  - Events (event_id/device_id/timestamp/coordinates)
  - Gender_age_train (device_id/gender/age/group)
  - Gender_age_test (device_id)
  - Phone_brand_device_model (device_id/phone_brand/device_model)

Introduction: Dataset (continued)

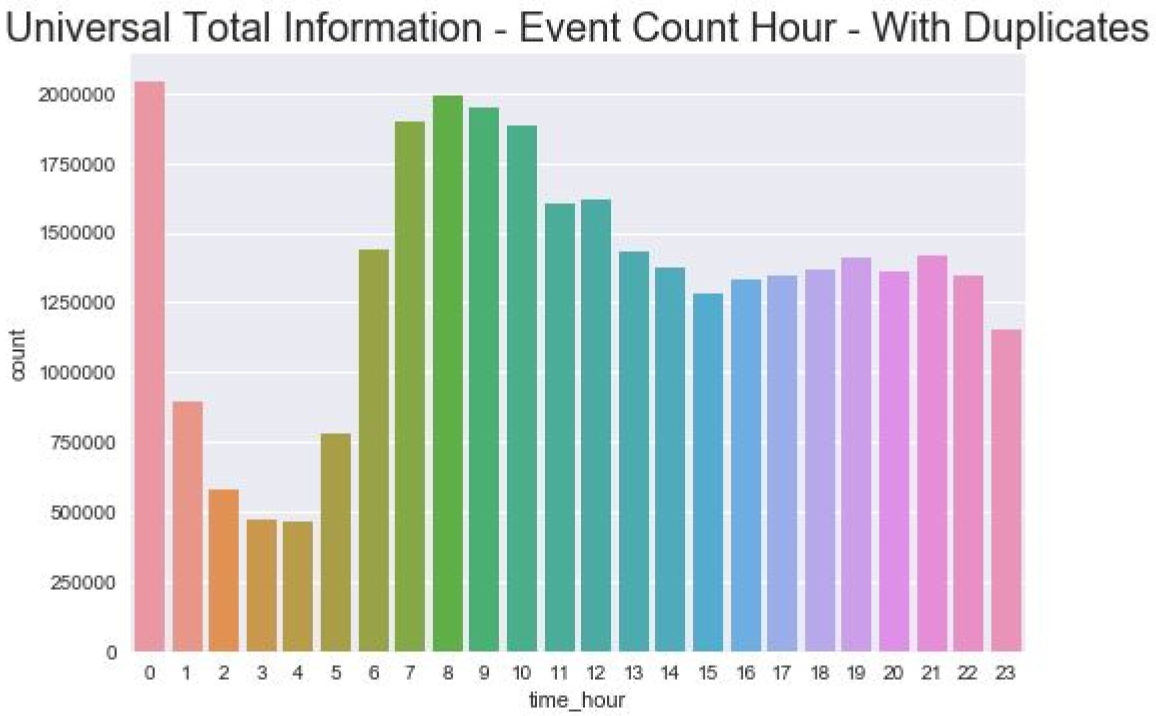




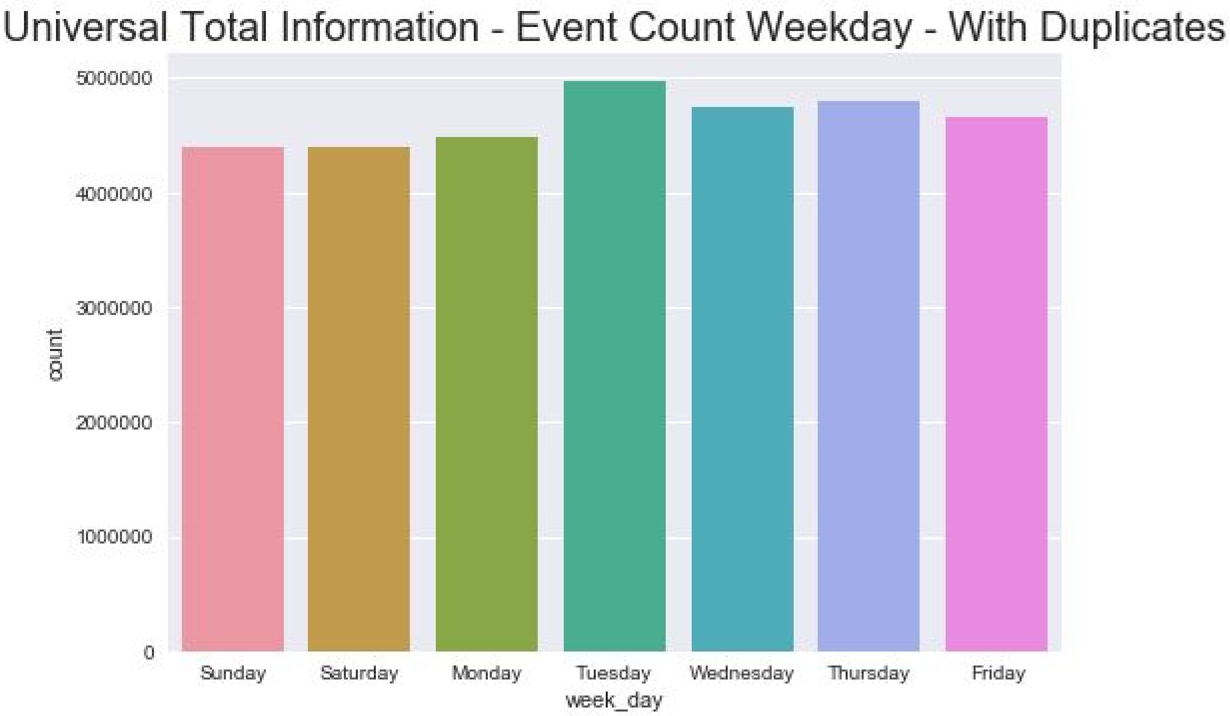
Introduction: Dataset (continued)



Introduction: Dataset (continued)

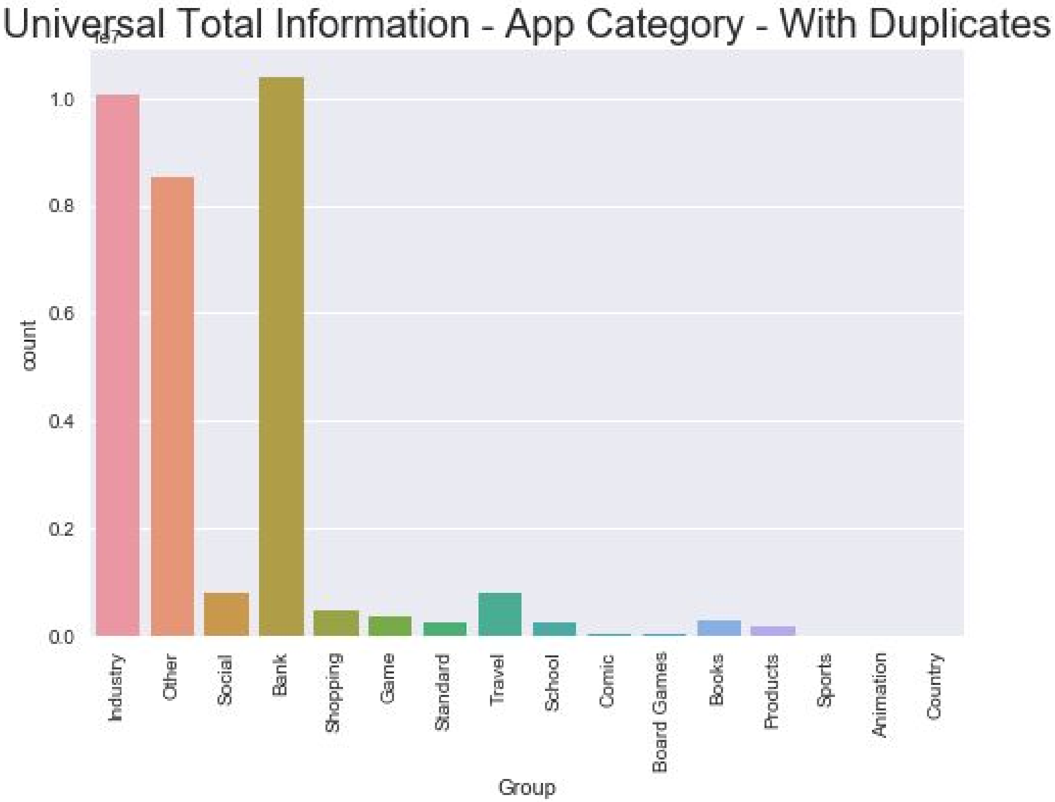


Introduction: Dataset (continued)



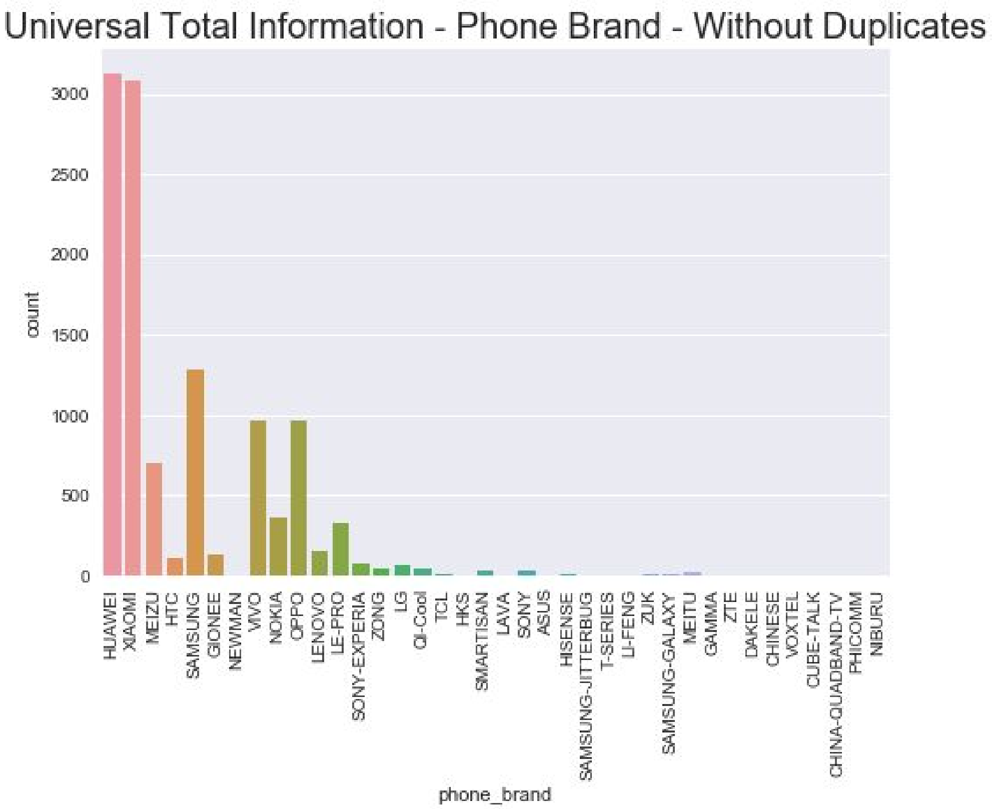
Introduction: Dataset (continued)

1. Bank
2. Industry
3. (Other)
4. Travel
5. Social



Introduction: Dataset (continued)

1. Huawei
2. Xiaomi
3. Samsung
4. Vivo / Oppo
5. Meizu



Introduction: Dataset (continued)

- Predictors
  - Apps
  - App Category
  - Brands
  - Models

- Problems
  - Missing data
  - Geographic coordinates did not make sense
  - Class imbalance
    - 40% of device users were male older than 32
    - 15% of device users were female

Approach: Pre-processing

- First situation
  - All the CSV files were combined (resulting in 4,2GB dataset)
  - Sparse data
    - Missing properties of the mobile devices (owner, brand…)
    - About 66% of the devices did not participate in any event

- Second situation
  - Combing .csv files based on device properties and installed apps
  - Creating sparse csr matrix for the features

Approach: Feature Extraction

- First situation
  - Creating normalized co-occurrence matrices of possible combinations
    - Timestamp
    - Brands
    - Device
    - Location
    - Apps
- Second situation
  - One Hot Encoding the different device properties
    - Apps installed
    - Phone brand
    - Phone device model

Approach: Classifiers

- Random Forests/Logistic Regression
- XGBoost
- Deep learning
  - Keras (library for Theano and Tensorflow)
  - Three layers (Dense, Dropout and PReLU)
    - Dense: standard densely-connected neural network
    - Dropout: compensation for overfitting
    - PReLU: Parametric Rectified Linear Unit; adaptively learns the parameters of the rectifiers and improves accuracy

Results: Kaggle Rankings

Random Forest

Score: 3.2 Rank: 1667/1689

LogisticRegression

Score: 2.265 Rank: 666/1689

XGBoost

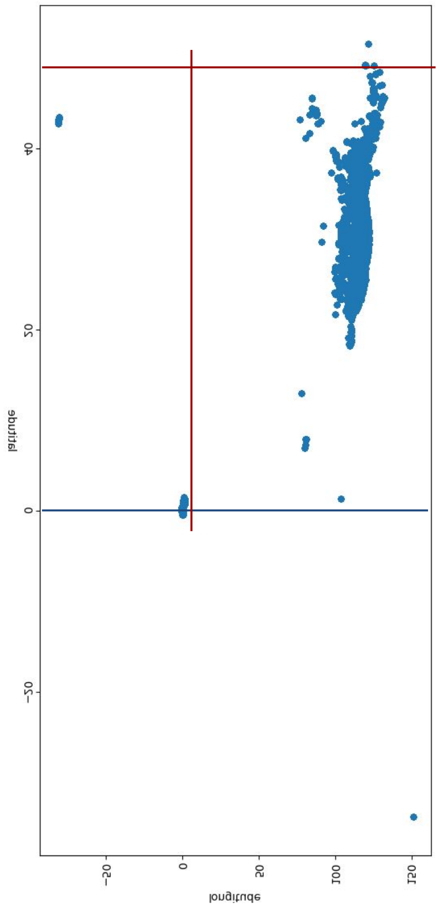
Score: 2.26853 Rank: 777/1689

Deep learning

Score: 2.25992 Rank: 645/1689

Other Ideas

- Geolocation (distance travelled, travel patterns)
  - No guarantees on accuracy of lat/long data
  - Odd points (e.g. 0,00N 0,00E)
  - Distance travelled varied from 0 to 160km / day
  - Reverse geocoding (i.e. extract addresses)
- Weighted categories



For the next competition

- Focus on pre-processing
- Optimalization of parameters
- Exploring the possibilities of XGBoost and DeepLearning

Conclusion

Thanks for listening!