JUNE 27, 2018

AGE PREDICTION BASED ON USER'S MOBILE USAGE BEHAVIOR

SRINIVASAN THANGAMANI

Contents

1. Abstract

In this report, a novel prediction methodology for predicting the end user's age, by taking into account the user behavior and environments, has been proposed. The core idea of this proposed methodology is to extract key features and develop a machine learning algorithm to predict the age.

To achieve this goal, traditional machine learning models and deep learning models to predict the individual user age and age buckets have been developed.

Seven different models to predict the age of the user were developed. Out of these seven models, deep learning-based model (regression) has outperformed the rest of the models with a MAPE of 0.19 (Accuracy: ~ 80%).

2. Data Description

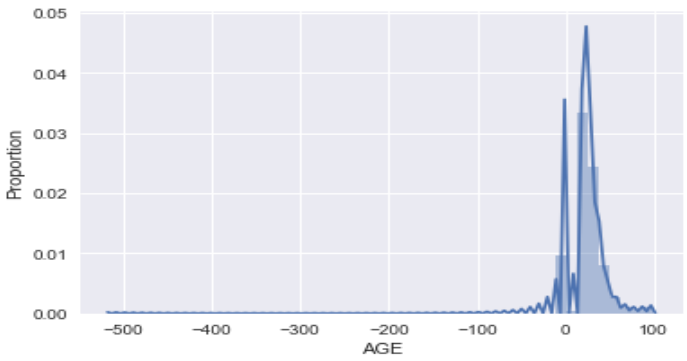
Number of Records	242828
Number of Independent Features	91
Dependent Variable (Age Derived from YOB)	1
Number of Categorical Variables	3 (platform, device category, gender)

The given data sources, hourly brq data and user app usage details, were integrated. Upon integration, it was inferred that most of the users do not have the 'hourly brq' data. Since hourly brq data is generated based on user action, we cannot apply any imputation technique to mimic those missing values. For further processing, only the users have both hourly brq and app usage details have been considered.

3. Data Understanding, Analysis and Transformation

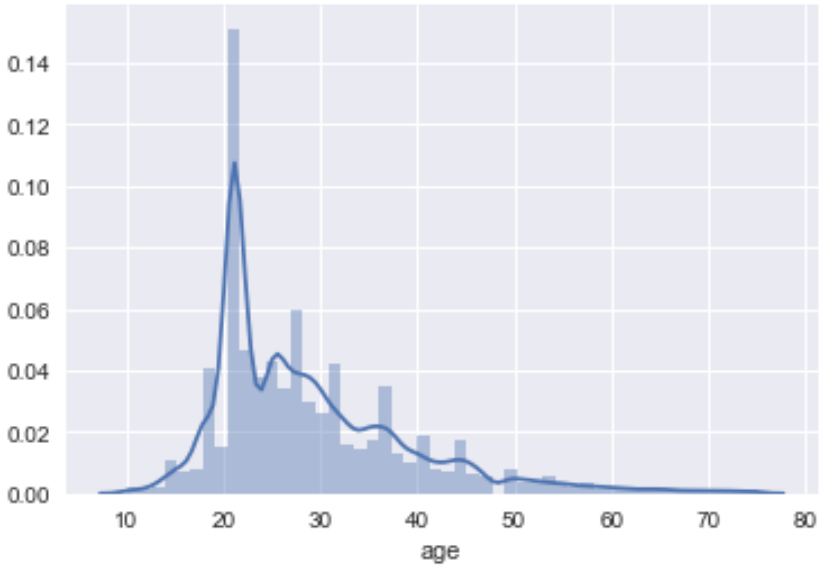
As a part of exploratory data analysis, the following steps were carried out,

1. Inspected all the columns for missing values. No missing values were found
2. Calculated the target variable 'age' from the YOB column



3. From the graph above, it is evident that there are some ages which are less than zero. The users with negative age was removed for further analysis

4. After removing negative ages, plotted the distribution of dependent variable ‘age’. From the graph below, it is evident that the data set is unbalanced



5. Encoded the following categorical columns
    a. device_category
    b. gender
6. Removed the following columns by means of the statistical method and manual inspection
    a. total_conn_brq – Highly correlated with the ‘brq’ column
    b. first_seen and last_seen – Difference of these dates is captured in ‘num_days’ column
    c. Platform – Contains same value for all the rows
7. For classification, the target is recoded as age range, the age buckets are
    a. 18 – 24 : 1
    b. 25 – 34 : 2
    c. 35 – 44 : 3
    d. 45 – 54 : 4
    e. 55 + : 5

4. Feature Extraction

Developed following feature selection methodologies, to extract more useful features from the given dataset,

1. Feature selection based on p-Value (select_features module in preprocessing.py) – (82 Features)
2. Feature reduction using PCA (reduce_dimension module in preprocessing.py)
3. Multivariate analysis, RFECV by using random forest regressor – (81 Features)
4. Multivariate analysis, RFECV by using random forest Classifier – (65 Features) (For classification)

5. Modeling and Results

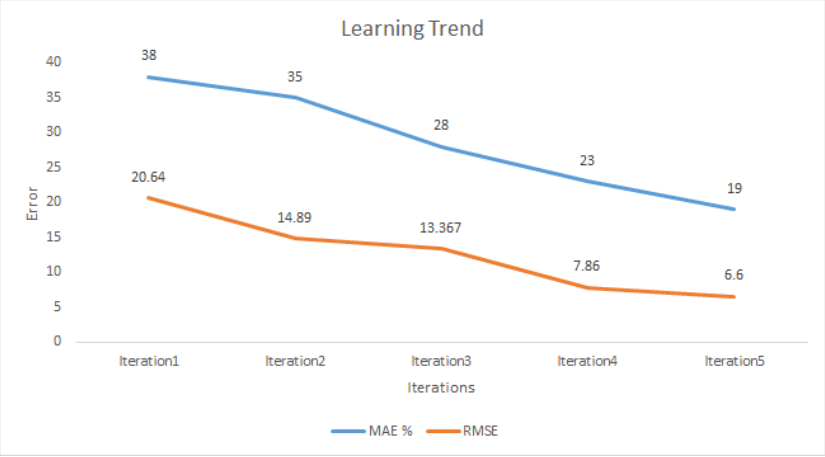
Following models were used for to predict the age

1. Deep Learning sequential model to predict the age (as regression)
    a. Activation function – relu – Rectified Linear Unit
    b. Number of layers: 4 fully connected dense layer
    c. Loss function: 'mape' Mean Absolute Percentage Error
    d. Metrics to optimize: RMSE and R2(More useful in linear relations)
2. Random forest Regression

Iteration results are tabulated below, in the decreasing order of MAE:

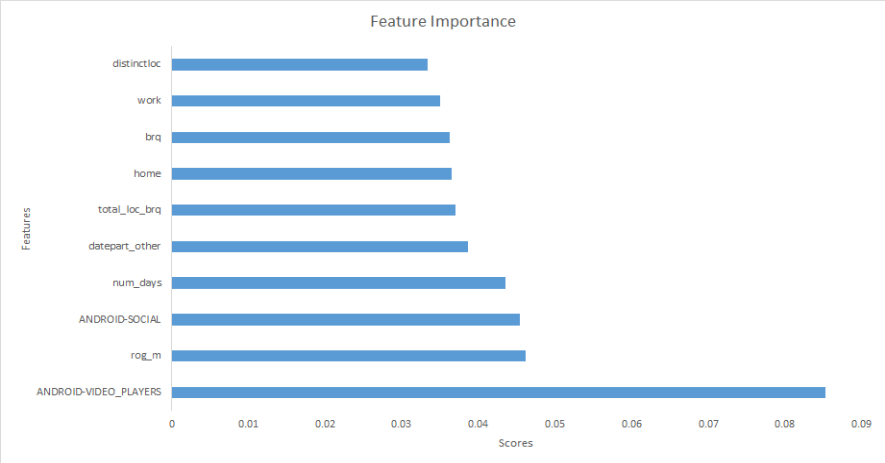
Iteration	Model Name	No of features	Evaluation Metrics
1	Random Forest Regression (4)	89 – All features	MAE: ~38 %, RMSE:20.64
2	Random Forest Regression (4)	82 – Based on pValue	MAE: ~35 %, RMSE:14.89
3	Random Forest Regression (4)	81 – Selected by using multivariate analysis	MAE: ~28 %, RMSE: 13.367
4	Deep Learning sequential Model (1)	89 – All features	MAE: ~23%, RMSE: 7.86
5	Deep Learning sequential Model (1)	73 – Selected by using multivariate analysis	MAE: 19.72, RMSE: 6.6

The graph below shows the learning trend across each iteration, the error rate significantly getting reduced upon selecting meaningful feature by using multivariate feature selection algorithm.

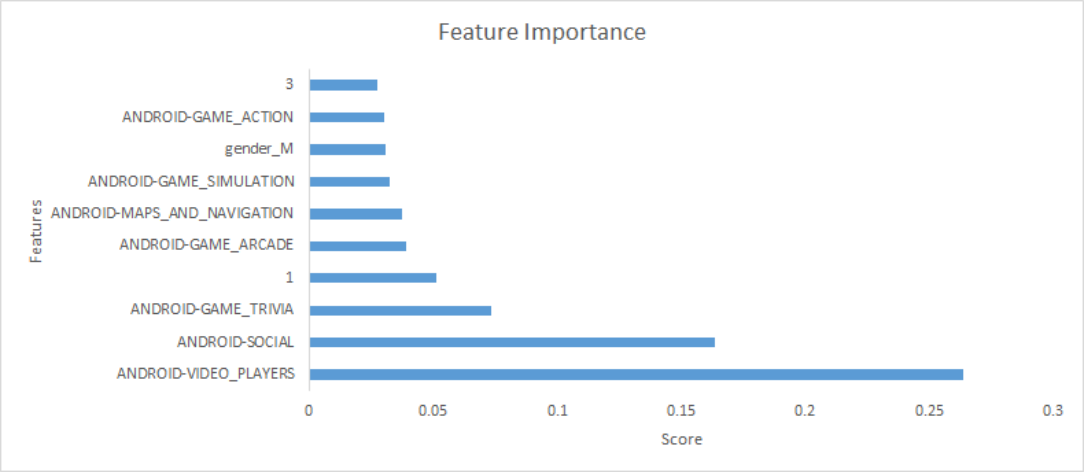


6. Insights and Inferences

Classification:



Regression:



From the charts above, it can be inferred that the video player, social media application, movement of the user, mobile usage at work significantly contribute to predicting the age of the user. These findings can help strategize the customized marketing's to the user.

7. Key findings from models developed classifying the user into different age buckets and Future enhancements

Following models were used for to classify the user into different age buckets

1. Deep Learning sequential model to classify the user into different age buckets
    a. Activation Function: 'relu', 'tanh' and 'sigmoid'
    b. Number of layers: 6 dense layers with dropout in the last before the layer
    c. Loss Function: categorical_crossentropy

    d. Metrics to optimize: fmeasure and accuracy
2. Random forest Classifier

From the below table, it is observed that the Random forest-based classification model has got higher accuracy than the deep learning based one. However, the f1 measure of deep learning NN model is higher than the random forest model. It signifies the robustness of the deep learning model over random forest model.

Iteration	Model Name	No of features	Evaluation Metrics
1	Random Forest Classifier (4)	65 – Selected by using multivariate analysis	Accuracy: 63 % f-measure: 0.4
2	Deep Learning sequential Classification Model (2)	65 – Selected by using multivariate analysis	Accuracy: 40 % f-measure: 0.57

1. Since the dataset is unbalanced, there is a drop in classification accuracy. There is no significant improvement in accuracy after adjusting the class weights and downsampling
2. There is a high possibility that the mobile is registered under a user than the real user, in this case, the recorded age and the usage behavior differs
3. With more data in all age range and appropriate training environment, we should be able to build a more robust prediction model

8. Conclusion

1. Based on the detailed analysis of data provided, a machine learning model to predict the age of the users has been developed
2. Extracted key features and its score
3. By combining both the key features and age we can strategize the customized marketing's to the users

Acknowledgments

1. Demographic Prediction Based on User's Mobile Behaviors, Josh Jia-Ching Ying, Yao-Jen Chang, Chi-Min Huang and Vincent S. Tseng
2. Demographic Prediction of Mobile User from Phone Usage, Shahram Mohrehkesh, Shuiwang Ji, Tamer Nadeem, Michele C. Weigle
3. Demographic Prediction based on Mobile User Data, Podoynitsina L., Romanenko A., Kryzhanovskiy K., Moiseenko A., Samsung R&D Institute Russia, Moscow, Russia