Part 1: Prediction Tasks

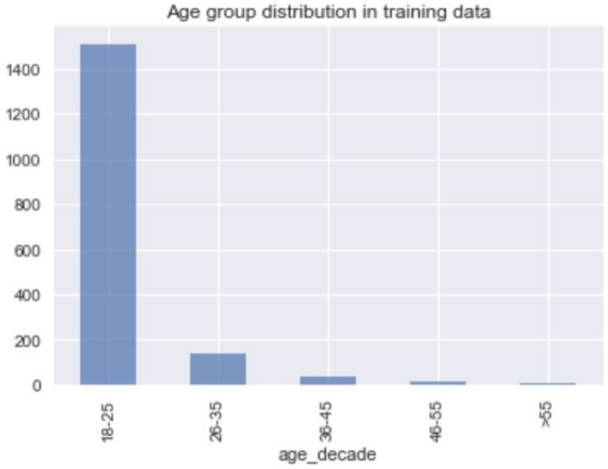
Candidate: Kelly Peng
Date: 12/03/2017

Introduction

The goal is to predict user age based on user twitter profile, friends profile, mentions profile, and user tweets. To solve this problem, age can either be treated as a continuous variable or a categorical variable. I decided to convert age into age groups and treat this problem as classification problem, the reasons are: (1) It's hard to predict the exact age based on user behavior, age group is more reasonable; (2) Predicting age range makes more sense for marketing usage, we usually target marketing at a specific age group instead of an exact age.

Data exploration and integrity

In ages_train.csv, we have 1,711 users, after joining with age_profiles.json (2,410 users), the final training dataset has 1,688 users. The minimum age is 18, the maximum age is 111, 88.45% users age between 18-25. The age is converted into 5 groups. And the number of users in each age group looks like follows:



Age Group	Percentage
18-25	88.45%
26-35	8.10%
36-45	2.17%
46-55	0.76%
>55	0.53%

Figure 1: Age distribution in training data

Features and Key Takeaways

The features that are included in the final model and explanation are as follows:

Feature	Explanation
number of friends, followers, statuses, liked tweets	Younger users tend to use twitter more frequently, thus the higher of these features, the more likely the user is in a younger age group.
Statuses count	The more tweets, the higher chance the user is in a younger group
Has profile description	Younger group (18-25) are more likely to have profile description,
Certain words usage	If user uses words like : 'married','producer','engineer','mother','30','family','woman','work','worki

	ng','writer','gallery','mom','wife','kids','retired','c.e.o','nurse','lady','business','employed', etc, there's high chance the user is older than 25. On the contrary, if there's 'semester', 'school', 'college', 'study', there's higher chance the user is younger than 25.
Source iOS or Android	Younger users are more likely to use mobile devices to tweet. Using iPhone is a stronger indicator than Android to show that the user is in a younger group.
Emoji count in tweets	The higher the number, the higher chance the user is less than 25.
Has emoji in tweets	Younger users are more likely to use emojis in tweets.
Use profile background tile	Younger users are more likely to use profile background tile.

Since friends count and followers count are highly correlated (people with more friends usually have more followers, vice versa), I only included followers count in the final model. From the feature importance plot shown below, the most important features are the number of liked tweets, number of followers/friends, number of tweets, then followed by whether the user uses background tile or not, whether the user uses specific words in profile description.

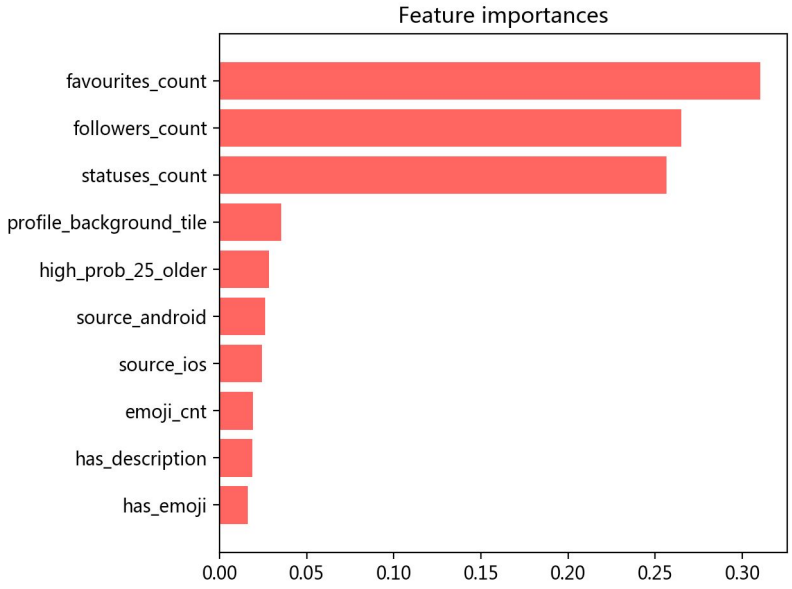


Figure 2. Feature importance plot generated from random forest model

Model Performance

Multinomial logistic regression and random forest classifier are used in model building, random forest classifiers outperforms multinomial logistic regression, because the original dataset has many outliers, random forest is robust to outliers, thus I used random forest as the final model. The average F1 score of final model is 0.83, precision score is 0.80, recall score is 0.87.

In order for this model to be applied to marketing campaign targeting, I need more data to train the model. The dataset I have now are very limited, many potentially useful features cannot be added to the model.

For example, the average age of users' friends, average age of users' mentioned IDs, this data is not available, but the features can be good indicators.

Challenges

1. Imbalanced class and limited data size: only 1,688 users available, with 200 of them not in 18-25 age group.
2. Outliers: For example, users age more than 100, users with extremely large number of friends, followers, statuses, favourites, etc.
3. Switched to python 3 to deal to deal with unicode emojis: started with python 2.7 but switched to python 3 in the middle in order to better deal with emojis.
4. NLP: when I was doing NLP to tweet text, I couldn't find much difference in the most important words in 18-25 and >25 age groups, for example, both groups use https, happy birthday, just, like, etc. In order to better extract information from tweet text, if I have more time I might manually create a stopwords list, and remove these common words from tweets.

Limitations

1. Friends and mentioned ID's age not available;
2. Limited data size;
3. Not enough feature to characterize older age groups.

Future Work

1. More natural language processing work can be done;
2. Sentiment analysis to tweet text.