Facebook User Profiling

Surabhi Agrawal
University of Washington Tacoma
agraws@uw.edu

Swati Garg
University of Washington Tacoma
swatig1@uw.edu

Kebra Thompson
University of Washington Tacoma
kebrat@uw.edu

ABSTRACT
Gaining insight about user details is an important part of recommender and other similar systems. In this paper we describe the methodologies used with several types of Facebook user information to create a machine learning model which will predict age, gender, and personality traits. We will detail the supervised learning techniques that were employed in the model and show that our results are better than the baseline model.

Keywords
Machine learning; supervised learning; binary classification, multi-class classification; regression; Facebook profiling;

1. INTRODUCTION
Individual people and companies are very interested in mining the vast amounts of data available on the web. They use this data to predict characteristics of internet users. For example, many online sites utilize recommender systems to predict products or pages that users will like. The predictions are often based on personal traits of users like their age group, gender or personality. In this paper, we explain the data and methods used to predict user traits from user profiles. Specifically, the goal of this project is to create a model that receives as input the profile picture, page likes, and status updates of a Facebook user and predicts gender, age, and big5 personality traits of the user from that data. We treat predicting age and gender as classification tasks and predicting personality scores as a regression task. This is not a new concept – in fact it is currently an exciting area of research. See [1] for personality detection from status updates, [2] for age and gender estimation from images, and [3] for personality prediction from Facebook page likes. Our models predict results that in most cases are better than the baseline predictions. We obtained the best results on the binary gender classification task (85% accuracy). The most difficult results to improve were some of the personality traits and the age prediction (61% accuracy).

2. METHODOLOGY
We began by working on the three separate data sources individually. That is, one member implemented model training based on images, another on the text files, and another on the relation data. In the end, we combined the results of methods which gave the best results.

2.1 Images
For images, the best results came from using the oxford feature set. The oxford set is a list of characteristics based on the profile photos in the data set. They are obtained as follows: for those users who have a recognizable face in the photo, the photo is cropped to just the face and many measurements and traits are listed. This is a Microsoft project and anyone can sign up for an account and use the API to submit photos and have the characteristics returned. Once one group in the class was able to get that for some of the users, the oxford data was provided to all of us. For example, two of the features are eyebrowRightOuter_x and facialHair_mustache. More information on the oxford data from Microsoft can be found in [5]. Not all instances in the training or test sets have a profile picture from which a single face can be extracted. For the users who do not, the most common results for age and gender and the average score for the personality traits is returned. This need to make a guess on a significant number of the users is the down-side to using the oxford features or other features based on the images. Many people use profile pictures of their children, their dogs, or of a group. This clearly makes the use of features determined by the photos less accurate.

To train a model for age and gender, two techniques were used. The first was k nearest neighbor from the sci-kit learn package [6]. A parameter of five nearest neighbors was used. This resulted in scores that were slightly above the baseline for gender and slightly below the baseline for age. We had anticipated better results with knn. Next, a random forest ensemble was

used. This also resulted in age prediction results below baseline but the results for gender were higher than the baseline. Random forest is an averaging ensemble method where many decision trees are created and the average of their individual results is calculated and returned. In our case, the random forest had a parameter of ten, meaning ten trees were created each time. The random forest model was one of those used in our final design for gender. It did not make the final cut for age.

Finally, the oxford features were also used in a linear regression model in order to predict the personality traits. A linear regression model is where the data points are plotted and the equation of the line that best fits the data is found. This line is then used in order to predict values for other instances. The linear regression model resulted in scores that were slightly better than the baseline for a couple of the characteristics (openness, conscientiousness, and neurotic) and were about the same as the baseline on the other two. The results stated here were obtained by using a portion of the training data for training and the rest for testing. Linear regression on oxford features was one of the models used in our final compilation for all five personality predictions.

2.2 Page Likes

The most successful results for page likes were obtained using two methods: the user-page-user approach and logistic regression.

We used the user-page-user approach to predict age, gender and personality scores of users. This technique maintains a mapping of individual pages and users liking that particular page, in the form:

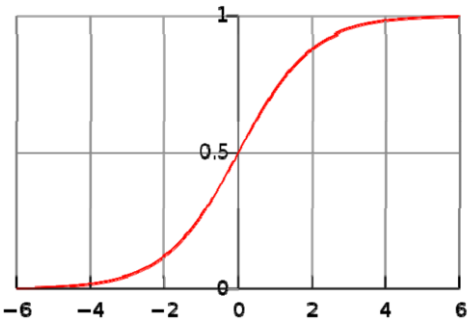
{Page1: [User1, User5], Page2: [User5, User6]……}

This means, that a page with pageid 'Page1' was liked by users with userids 'User1' and 'User5'. When a new user is encountered, and we need to predict the age, gender and personality scores for that user, we convert the page likes data for that user in the form:

UserX : [Page2, Page7, Page9]

meaning, this unseen user with userid 'UserX' liked pages Page2, Page7 and Page9. Now, for prediction, we consider each page the unseen user has liked, and obtain the labels for each user liking the same page. Eg: To predict age, we will obtain the age values of all users liking pages Page2, Page7 and Page9 and take the majority age bucket as the predicted value of age for the unseen user. Similarly, we get the majority gender value and the mean of each personality trait of all the users and predict those values as labels for the new user. We performed feature selection by considering only those pages which had between 0-50 users liking that page. We observed that this range of threshold gave us the most accuracy.

The second technique used was logistic regression. Contrary to what the name suggests, this is a binary classification technique rather than a regression technique. "Regression" comes from fact that it outputs continuous values from 0 to 1 [7]. Thus, we have treated output of the model as the probability of class label to predict age and gender only, in this project. Overall, logistic regression maps a point x in d-dimensional feature space to a value in the range 0 to 1. We represent the hypothesis space of a logistic regressor by a sigmoid function:



When our hypothesis outputs a number, we treat that value as the estimated probability that class label y=1 belongs with input feature vector x. To be more precise, when the probability of y being 1 is greater than 0.5 (in the sigmoid function diagram above) then we can predict y = 1, else we predict y = 0.

For a 2-D feature space, the logistic regressor learns a linear decision boundary for linearly separable data and for higher dimensional space and data which is not linearly separable, it learns a hyperplane which separates our classes.

Thus, for our data, which has a feature space of about 500,000 pages, we predict whether a user is a male (0) or female (1) using logistic regression as described above. If the probability outputted by the model is higher than 0.5, the user is classified as a female (1), else the user is classified as a male (0).

For predicting age, which is a multi-class classification problem, the logistic regressor internally splits the classes into multiple binary classification tasks. The feature age has four possible values, xx-24,

25-34, 35-49 and 50-xx. Thus the logistic regressor will internally build four models and learn the decision boundaries for each:

1) The user belongs to class xx-24 (label 1) or the user does not belong to class xx-24 (label 0)
2) The user belongs to 25-34 (label 1) or not (label 0)
3) The user belongs to 35-49 (label 1) or not (label 0)
4) The user belongs to 50-xx (label 1) or not (label 0)

When an unseen example is encountered, the probabilities for each of the four models are calculated and the class which has the highest probability, is predicted as the label for the unseen example.

In our project, we train the logistic regressor separately for age and gender prediction. The input to the classifiers is of the form:

X = [[0 1 0 0 1 0 ……]
[0 0 0 1 1 0 ……]
[1 1 1 1 1 0 ……]
……]

Yage = [xx-24 50-xx 25-34 ……]

Ygen = [0 0 1 1 1 …….]

where X is the feature vector matrix. Each row in the matrix represents the pages liked by a user. The matrix has a 0 where the user did not like that particular page, and 1 where the user liked that page.

We have about 500,000 pages and 9500 users. Thus, our matrix X has 500,000 columns and 9500 rows. To train our model for age prediction, we will send (to the logistic regressor) the label vector Yage, which gives the age labels for each user, along with the feature vector matrix X. Similarly, for gender prediction, we will send X and Ygen.

We query the model with the information of a new user in the form:

X = [0 1 1 0 0 …]

This represents the pages that user has or has not liked. The age and gender model will individually predict the labels for this user, based on the probabilities calculated in the trained model. For this method we selected only those pages which were liked by 4-200 users. Though feature selection in logistic regression did not make much of a difference in accuracy, we sampled the number of features for speeding up the training of the model

2.3 Status Update of the User

We divided the problem to predict the age, gender and big5 personality traits into classification and regression problems using the status updates of each user. We used the classification model (naïve bayes) for predicting age and gender and regression model (linear regression) for predicting big 5 personality traits (extraversion, agreeableness, openness, conscientiousness and neuroticism).

a. Age and Gender Classification: For predicting the age and gender we used naïve bayes classifier using the python sklearn library.

   Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

   The first step to train an effective model is feature selection. We created the bag of words of the status update of the user using the tokenizer and removed the stop words from this bag of words using the nltk python library. The frequency of each word is calculated. We selected the most common words from this bag of words to train our model

   We passed the training data in the form of {x1,x2,x3,x4…xn|y}

   where x1…xn are the features. In our example it represents the frequency distribution of the bag of words.

   Y is the target or Llabel. For gender it is either 0 or 1 (0 = Male and 1 = Female). For age it is one of the four age buckets. The model is trained using the naïve bayes classifier. For our model we used multinomial NB as we had multiple labels for the age.

   Once the model was trained, the age and gender for the unseen data was predicted.

   We computed the accuracy of our model by calculating the number of correctly classified predictions.

b. Big5 personality traits:

   Predicting big5 personality traits is the regression problem. We used the linear

regression to predict the personality traits of the user.

Linear regression is one of the algorithms used to predict scalar values given some inputs. Linear regression can be modelled like this:

$$yi=\beta 1xi1+\beta 2xi2+\cdots+\beta pxip+\epsilon i$$

$$yi=\beta 1xi1+\beta 2xi2+\cdots+\beta pxip+\epsilon i$$

It can also be written in a vector form as:

$$Y=X\beta+\epsilon$$

For the prediction of personalities using status updates we were provided with the LIWC and NRC features [described in Appendix A]. We created the dataframe using the pandas in python. This dataframe consists of the personality traits in combination with the LIWC features. We trained the model using this dataframe with linear regression. We used the trained model from the above to predict the big5 personality traits of the hidden data. As this is the regression problem we computed the root mean square error to validate the model.

2.4 Ensemble

Ensemble consists of using various independent classifiers to increase the accuracy of the model. The models whose individual accuracy was best were selected to give the combined higher accuracy. However, they also needed to be above the baseline in order to be chosen. There are various ways to create the ensemble like majority vote, bagging, boosting.

For our project we used majority voting for predicting age and gender. For age we used the majority votes of the following models: Naïve bayes on text, user page user with page likes and logistic regression with page likes.

For gender we used the majority votes from the following models: Naïve bayes on text, user page user on page likes, logistic regression on page likes, and random forest on images

For big5 personality we calculated the average of the following models: linear regression with text, user page user with page likes and linear regression with images.

We observed that after creating the ensemble the accuracy of our prediction models increased drastically.

3. DATASET AND METRICS

3.1 The data

The dataset used came from the myPersonality application. This Facebook app allowed users to take tests and have information about them recorded for use in future data science endeavors. The data was recorded with consent and users had motivation to be honest since their reward was to receive feedback on their results [4]. Many features were collected for each user. These included their page likes, posts, and profile pictures. Participants took surveys multiple times. Their friends took surveys about their personality. There is a great deal of material available through myPersonality. More information on the data and the application in general can be found in [4].

The particular set of data that we were given for training and testing purposes contained 9500 users. For each of these users, denoted with a user identification number, we received a file containing their profile information, the actual gender, age, and personality scores. We also received a text file containing a list of the status updates for each user, a file containing the identification numbers of the Facebook pages liked by each user, and their profile pictures.

In addition to the training data, we were also given a set of 334 users along with their data formatted exactly as with the training data. However, we were not given labels for this set of data. This set, called the public-test-data, was intended to allow us to try our code on data set up exactly the same as with the hidden data used in the evaluation of the model.

The profile data consists of age, gender, and five personality trait scores. The age is one of the following ranges:

xx-24, 25-34, 35-49, or 50-xx

The gender is listed as 0.0 if male and 1.0 if female. The personality traits correspond to the five traits tested in the popular Big Five Personality test. This test reports scores on a scale of 1-5 for each of five traits: openness, conscientiousness, extroversion, agreeableness, and neuroticism.

The hidden data set used to evaluate our model consisted of over 1000 users with profile picture, text file of status updates, and set of page likes. This data was never seen by us.

3.2 Evaluation Measures

We began by running a simple script to use the training data and determine the most common gender and age bracket as well as the average score for each of the five personality traits. These baseline values were obtained by selecting the most common values for age (xx-24) and gender (female) in the training data and by calculating the mean over all the instances for the five personality scores. The trained model should perform better than the baseline.

The evaluation measures employed were root mean square error and accuracy. The accuracy measurement was used for the classification tasks: age and gender. This value is simply the number of correctly classified instances divided by the total number of instances. The root mean square error was used for the regression tasks: the five personality scores. It is calculated by squaring the difference between the predicted score and the actual score for each instance. These squared differences are then averaged and the square root is taken. This value is called the root mean square error.

We tested our data independently, using a portion of the training data. Primarily this was done using an 80/20 split where 80% of the training data was used to train the model and the remaining 20% was used to test the model and calculate the accuracy and the root mean square error. In addition, the data was tested weekly on the hidden data set. The accuracy and root mean square error were calculated for these. Our initial goal was to produce scores that were better than the baseline. After achieving that, we then tried to improve our scores as much as possible.

	Age	Gender
Baseline	0.59	0.59
Random Forest Oxford (training data)	0.53	**0.83**
User-Page-User	**0.60**	**0.71**
Logistic Reg Page Likes (training data)	**0.66**	**0.78**
Naïve Bayes Text	**0.61**	**0.71**

Table 1: Results of various individual methods for age and gender

	Open.	Neur.	Extr.	Agree.	Cons.
Baseline	0.65	0.80	0.79	0.66	0.73
Lin Reg Oxford (training data)	**0.60**	**0.70**	**0.78**	**0.66**	**0.77**
User-Page-User	**0.65**	**0.79**	**0.79**	**0.65**	**0.72**
Lin Reg Text (LIWC)	**0.65**	**0.79**	**0.79**	**0.65**	**0.72**
Lin Reg Text (LIWC + NRC)	0.65	0.79	0.79	0.65	0.72

Table 2: Results of various individual methods for personality scores

4. RESULTS

The results that we obtained with our individual models varied dramatically. The results of some of our models for age and gender are shown in table 1. The results of some of our regression models are shown in table 2. In both tables, the values shown in bold type highlight which techniques were used in the final model for each classification or regression problem.

The linear regression model used on both oxford features from the images and on the text LIWC features provided good results. Linear regression is a simple technique that assumes the features and labels are linearly related. Because our results are not terrible when we use this method, we believe this is a reasonable assumption. When training over a large enough sample of instances, the linear regression model is robust to occasional outliers in the data. The regression model using the oxford features was particularly accurate with openness and neuroticism prediction. While these scores are only based on a subset of people who had oxford features available, we believe that these two qualities are more easily predicted with this data because facial features are related to these characteristics. For example, someone who is open might have eyebrows in a raised position or someone who is neurotic may have compressed lips. With text, we chose to use the linear regression model that relied only on the LIWC features instead of the one that used both LIWC and NRC features.

This was because there appeared to be no difference in the final accuracy values even when used on the hidden data.

The random forest model returned excellent values for gender though the result for age was not so successful. Independently, this model predictor for gender was the best score we acquired. This was only on the training data and never on the hidden data by itself, but we believe that our good result is due to a couple of factors. First, gender of course is the easiest prediction task since it is a binary classification problem. It is expected then that our gender results would be better than that for age or regressions. Secondly, random forests take care of feature selection inherently by first choosing a random set of features and then building a decision tree based on the best feature being selected at each node. Those features that are not as important end up not being used to build the tree. Those features that are best at classifying the users rise to the top in multiple trees. Therefore, the outcome of the majority vote taken by the trees in the forest relies on the best predictive features. This is the benefit of using an ensemble method such as random forest rather than an individual decision tree.

The user-page-user approach using page likes gave pretty good results (71% accuracy) for gender prediction, but it did not give impressive results for age prediction (60%). These results are as seen on the hidden data. The personality prediction was not impressive either. Three of the five personality traits were observed to be below the baseline. Feature selection did improve the accuracy to some extent, but not by a huge margin. The user-page-user approach works well since, intuitively, a user who likes a particular page, is considered similar to other people liking the same page. For example, people liking the Harry Potter page on Facebook are likely to be in the 'xx-24' age group, while people liking sports pages are more likely to be males. Similarly, an introvert person is more likely to like pages that are more likely to be liked by other introverts. Thus if we obtain the majority age group or gender, and the mean of personality scores from the pages an unseen user has liked, we are likely to correctly predict the age, gender and personality of the user.

The user-page-user approach particularly did not seem to work for predicting age, since the data that was available to us was skewed toward the majority age bucket xx-24. We had about 5500 users in this age bucket out of a total of 9500 users. So whenever we wanted to predict the age of an unseen user, the probability of this age group being in majority was very high. Thus, our model predicted the xx-24 age group for most of the unseen users. Therefore, overall accuracy was very close to baseline accuracy.

The logistic regression technique gave decent results. The results for this technique were only observed on the 80/20 split on the training data. This technique gave us an impressive accuracy of 78% for gender and 66.4% for age, which was higher than the user-page-user approach. We could have used Naive Bayes for classification but there are differences between the two. Naive Bayes models the densities of classes and selects the class that most likely produces the features. In this case the age bucket xx-24 again would have the highest probability and would be the most frequently predicted class, since in Naive Bayes, we set each feature's weight independently, based on how much it correlates with the label. Weights come out to be the features' log-likelihood ratios for the different classes.

Logistic regression has a different approach; similar to linear regression, it makes no assumptions about distributions of classes in feature space. It sets all the weights together such that the linear decision function tends to be high for positive classes and low for negative classes. It tries to model class boundary and membership directly, e.g. in >3 feature dimensions case, it would be looking for the hyperplane that best separates the classes. Thus, this method can predict the biased class with much more accuracy than Naive Bayes or the user-page-user approach.

One additional advantage of logistic regression is that it performs internal feature selection to some extent by giving lower weights to features that contribute minimally to the class label. Thus, we observed that feature selection on the user's part made no noticeable difference to the accuracy of the model.

The final result submitted on the VM includes all of the models whose individual accuracies and root mean square errors were best. These individual results were shown in bold face type in tables 1 and 2. Each of the models selected returned a classification prediction for gender and/or age or a regression value in the case of personality prediction. The personality results were averaged to obtain the overall prediction. For gender and age, the majority of the three or four models was used. When this majority model was used on the hidden data, the results were strong, especially

for gender and openness. In fact, four of the five personality traits were improved from the baseline, age was improved slightly, and gender was improved dramatically. These results are shown in tables 3 and 4. The final predictions were then written to an output file in the xml format [Appendix B].

	Age	Gender
Baseline	0.59	0.59
Majority Model	0.61	0.85

Table 3: Results of the combined majority model for age and gender

	Open.	Neur.	Extr.	Agree.	Cons.
Baseline	0.65	0.80	0.79	0.66	0.73
Average Model	0.63	0.79	0.78	0.66	0.72

Table 4: Results from combined average model for Big5 personality scores

5. CONCLUSION AND FUTURE WORK

There is much that still could be done related to this project. There are many feature selection and feature combination techniques that we would like to try. The oxford features could be used for other models; in particular, we would like to try them with logistic regression. We would certainly like to improve our age result. One technique we would like to try for improving our age results would be sampling the data. Since our data was skewed toward the xx-24 age bucket, we could sample the training data in a manner that would normalize the ratio of the age buckets with respect to each other. This would avoid the bias of models like Naive Bayes toward the most frequently occurring class. In addition, we need to go back to researching how to improve personality scores as our best work in that area resulted in fairly small improvements. We would like to lower those errors even further. We also observed that Naïve Bayes classification works better with the binary classification than with the multi-classification. The accuracy of the classifier was higher in the gender prediction than the age prediction.

6. ACKNOWLEDGMENTS

Our thanks to Dr. Martine de Cock and Golnoosh Farnadi for their assistance with the development of this project.

7. REFERENCES

1. Farnadi, G., Zoghbi, S., Moens, M., and De Cock, M. 1993. Recognizing personality traits using Facebook status updates. In *Proceedings of WCPR13 (Workshop on Computational Personality Recognition) at ICWSM13 (7th International AAAI Conference on Weblogs and Social Media)*, pages 14-18, 2013.
2. Eidinger, E., Enbar, R., and Hassner, T. Age and gender estimation of unfiltered faces. In *IEEE Transactions on Information Forensics and Security*, 9, 12, (December 2014), 2170-2179.
3. Kosinski, M., Stillwell, D., and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. Direct submission to *Proceedings of the National Academy of Sciences*, 100, 15, (April 9, 2013), 5802-5805.
4. Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D. (2015) Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. *American Psychologist*. DOI= http://mypersonality.org/wiki/doku.php
5. Microsoft's Project Oxford. https://www.projectoxford.ai/
6. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.Ski-kit learn. http://scikit-learn.org/stable/
7. http://courses.washington.edu/css490/2012.Winter/lecture_slides/05b_logistic_regression.pdf

8. Appendix A: Feature Set Descriptions

8.1 LIWC Features

The Linguistic Inquiry and Word Count tool, known as LIWC. There were 81 features for each document including features related to standard counts (e.g., word count), psychological processes (e.g., the number of anger words such as hate and annoyed in the document), relativity (e.g., the number of verbs in

the future tense), personal concerns (e.g., the number of words that refer to occupation such as job and majors), and linguistic dimensions (e.g., the number of swear words). For a complete overview of the features.

8.2 NRC Features

NRC is a lexicon that contains more than 14,000 distinct English words annotated with 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), and 2 sentiments (negative, positive). For each document we 10 features per document.

9.Appendix B: XML Format

<User id = "f67e60570d27984cd729aa128e6d9b08"
age_group = "25-34"
gender = "Female"
open = "3.89886155017"
conscientious = "3.57752855848"
extrovert = "3.61433069746"
agreeable = "3.70635219047"
neurotic = "2.68082332498"
/>