Gender Classification from Tweets

Juan Manubens, Quentin Chalvondemersay

April 30, 2017

Executive Summary

This study attempts to predict user gender from two types of predictors reflecting two different sets of attributes: profile characteristics (12 variables) and language usage patterns (244 word-presence binary variables, representing words that appeared in at least 100 of the tweets of the cleaned sample). After cleaning the data set to remove all accounts that were suspected to be fake (bot accounts), we tested 3 different models: (1) Logistic Regression with variables selected through Elastic Net (2) Random Forest with all variables (3) Random Forest with Elastic Net Variables. All models performed reasonably well, but the best performance came from the Random Forest model, with 70.5% and 70.6% accuracy for the 75-cap and 50-cap sets, respectively. This model considered 22 randomly selected variables to be available for splitting at each tree node, as set by the parameter *mtry*. We set *ntree*, the number of trees to grow, at 1000. The results of our study not only provide strong support for our initial view on the presence of language use differences between genders on Twitter but also corroborate the conclusions of existing studies that link gender and social media activity.

Goal of the Study

The ultimate goal of this study is to develop a statistically significant model that can accurately predict user gender based on exhibited Twitter data. Accomplishing this goal involves cleaning the raw data, identifying the variables that are most relevant in predicting gender, and fitting appropriate classification models. This study is especially relevant in today's world as digital presence and social networking activities online are deeply integrated into most people's lives in the 21st century. Given the sheer volume of activities and traffic on leading social networking platforms such as Facebook and Twitter, we believe that there are valuable insights to be unlocked by analyzing the big data made available through such platforms.

Literary Review

There is precedent of studies exploring different text mining techniques, data cleaning procedures, featurization methodologies and classification models aiming to analyze differences between different populations of people. A good example is Schwartz et al. (2013), who analyzed 700 million words gathered from Facebook messages to identify similarities and differences in language use among different groups of participants [1]. Another study conducted by Coviello et al. (2014) even found suggestive evidence that online social networks such as Facebook contribute to "global emotional synchrony" as emotions spread through such online platforms [2]. As such, user data generated on various online social media services can be used to identify insightful user characteristics.

In this study, we focus on identifying user gender. While the topic of gender identification through social media activity data has not been covered extensively in existing literature, the most predominant literature in this domain is arguably a study by Kosinski et al. (2012). He conducted a statistical study using Facebook likes to show that sensitive user information, including not only gender but also sexual orientation, can be identified with high accuracy [3].

Our study analyzes relatively more granular data including Twitter profile characteristics and language usage patterns in individual tweets. Our preliminary expectation was that using this combination of features would allow us to accurately predict user gender with at least 60% accuracy.

Data

The original Twitter data analyzed in this study comes from Kaggle, an open-source web platform for data analytics. The dataset is comprised of approximately 20,000 observations (Twitter profiles) along 26 features [3]. According to Kaggle posting, the original 26 variables are described as follows:

1) **unit_id**: A unique ID for user
2) **golden**: Whether the user was included in the gold standard for the model; TRUE or FALSE
3) **unit_state**: State of the observation; one of finalized (for contributor-judged) or golden (for gold standard observations)
4) **trusted_judgments**: Number of trusted judgments (int); always 3 for non-golden, and what may be a unique ID for gold standard observations
5) **last_judgment_at**: Date and time of last contributor judgment; blank for gold standard observations
6) **gender**: One of male, female, or brand (for non-human profiles)
7) **gender:confidence**: A float representing confidence in the provided gender
8) **profile_yn**: "no" here seems to mean that the profile was meant to be part of the dataset but was not available when contributors went to judge it
9) **profile_yn:confidence**: Confidence in the existence/non-existence of the profile
10) **created**: Date and time when the profile was created
11) **description**: The user's profile description
12) **fav_number**: Number of tweets the user has favorited
13) **gender_gold**: If the profile is golden, what is the gender?
14) **link_color**: The link color on the profile, as a hex value
15) **name**: The user's name
16) **profile_yn_gold**: Whether the profile y/n value is golden
17) **profileimage**: A link to the profile image
18) **retweet_count**: Number of times the user has retweeted (or possibly, been retweeted)
19) **sidebar_color**: Color of the profile sidebar, as a hex value
20) **text**: Text of a random one of the user's tweets
21) **tweet_coord**: If the user has location turned on, the coordinates as a string with the format "[latitude, longitude]"
22) **tweet_count**: Number of tweets that the user has posted
23) **tweet_created**: When the random tweet (in the text column) was created
24) **tweet_id**: The tweet ID of the random tweet
25) **tweet_location**: Location of the tweet; seems to not be particularly normalized
26) **user_timezone**: The timezone of the user

Data Pre-processing

Although the dataset provides a breadth of potentially useful data, it required some basic cleaning in order to be usable in the statistical analysis process. Our current version of R Studio did not support the packages we needed for this phase, so we decided to use Python for this phase. All the code can be found in the bottom of the R Markdown file, and attached in the project zip file.

To begin, we extracted the relevant data for only human observations as the dataset originally includes non-human profiles related to brands. Additionally, several feature columns were eliminated early on as they do not add meaningful value to the analysis process (e.g. unique identifiers); these included variables such as **unit_id**, **golden**, **unit_state**, **trusted_judgments**, **last_judgment_at**, **profile_yn**, **gender_gold**, and **profileimage**.

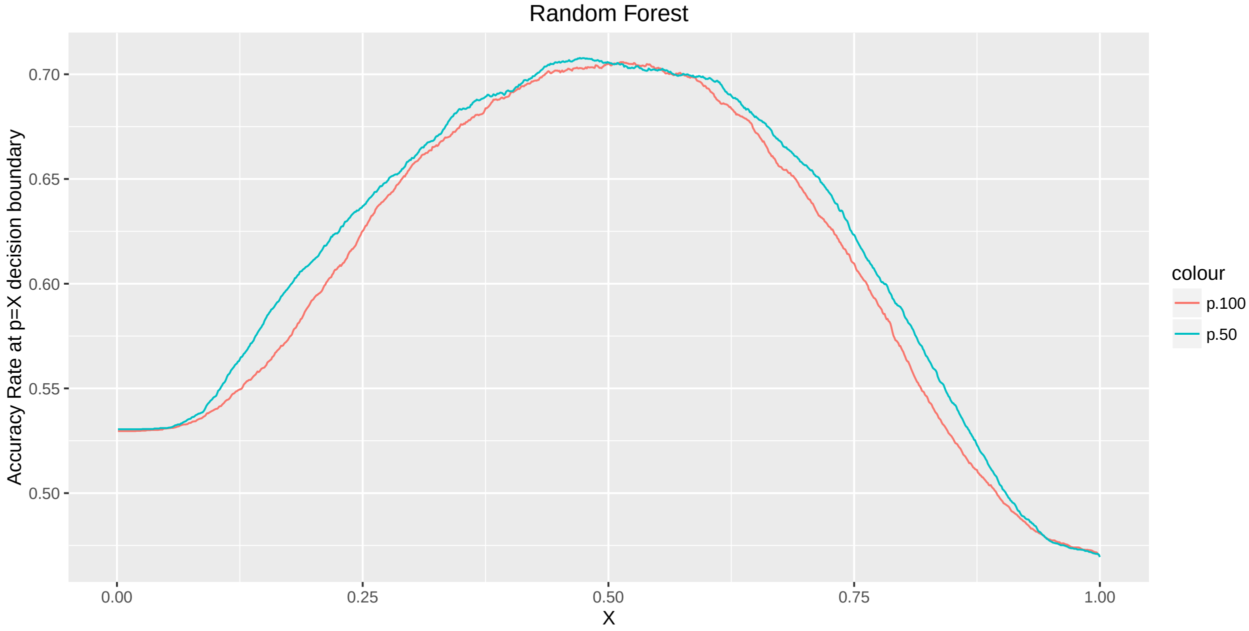
We decided to eliminated several questionable observations based on **gender:confidence**, **profile_yn:confidence**, and the number of tweets per day as we deemed them to be invalid observations likely generated by automated bots. For the purposes of this study, we considered two subset data with tweets per day capped at 50 and 75.

Featurization

We decided that the most reasonable approach was to consider two types of predictors reflecting two different sets of attributes: profile characteristics (12 variables) and language usage patterns (244 word-presence binary variables, representing words that appeared in at least 100 of the tweets of the cleaned sample). To illustrate a few of these:

1) **has_mention**: When tweeting, a person has the option of tagging another user in the tweet. This is done by writing '@' and the other user's twitter name without any spaces. A twitter name can be a maximum of 15 characters so we will look for ocurrences of '@' followed by a max of 15 characters. To avoid counting ocurrences that fulfill these requirements but are not mentions, we note that twitter handles only allow alphanumeric characters and underscores.
2) **days_active**: days since the account was created until the tweet was created
3) **has_hashtags**: indicates whether or not there are hashtags in the tweet

Findings



	Elastic Net Logit	Random Forest	Random Forest - Elastic Net
Cap at 75	0.697	0.705	0.686
Cap at 50	0.702	0.706	0.687

Table 1: Model Results

All models performed reasonably well, but the best performance came from the Random Forest model, with 70.5% and 70.6% accuracy for the 75-cap and 50-cap sets, respectively. This model considered 22 randomly selected variables to be available for splitting at each tree node, as set by the parameter *mtry*. We set *ntree*, the number of trees to grow, at 1000.

Detailed Analysis

Using the cleaned dataset, we approached the statistical modeling process in two ways: logistic regression and random forest.

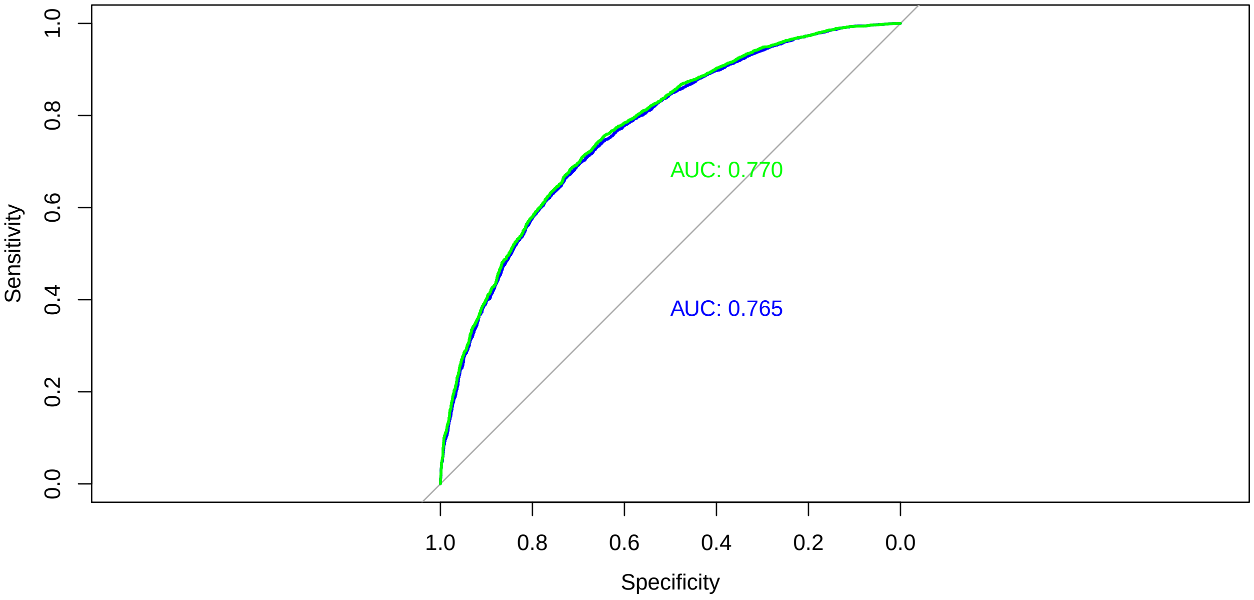
Model 1: Elastic Net Logistic Regression

Variable Regularization

First, the construction of the logistic regression model began with the selection of important variables using elastic net. As previously noted, we applied the modeling process to two subsets of data (tweets per day capped at 50 and 75) to examine the presence of any noteworthy differences. Using an alpha value of 0.99, cross-validation k-fold value of 10, and lambda value of **lambda.1se**, the elastic net variable selection process identified 124 variables for the dataset with tweets per day capped at 50 and 122 variables for the dataset with tweets per day capped at 75. We believe that the use of lambda.1se as the lambda value for the elastic net variable selection process is appropriate as it essentially yields the most parsimony without compromising analytical accuracy.

Logistic Regression Fit with `beta.elastic` Output

Once the relevant variables had been identified, we ran binomial logistic regressions with the selected variables and generated ROC plots for performance comparison (green curve for tweets capped at 50 per day; blue curve for tweets capped at 75 per day).

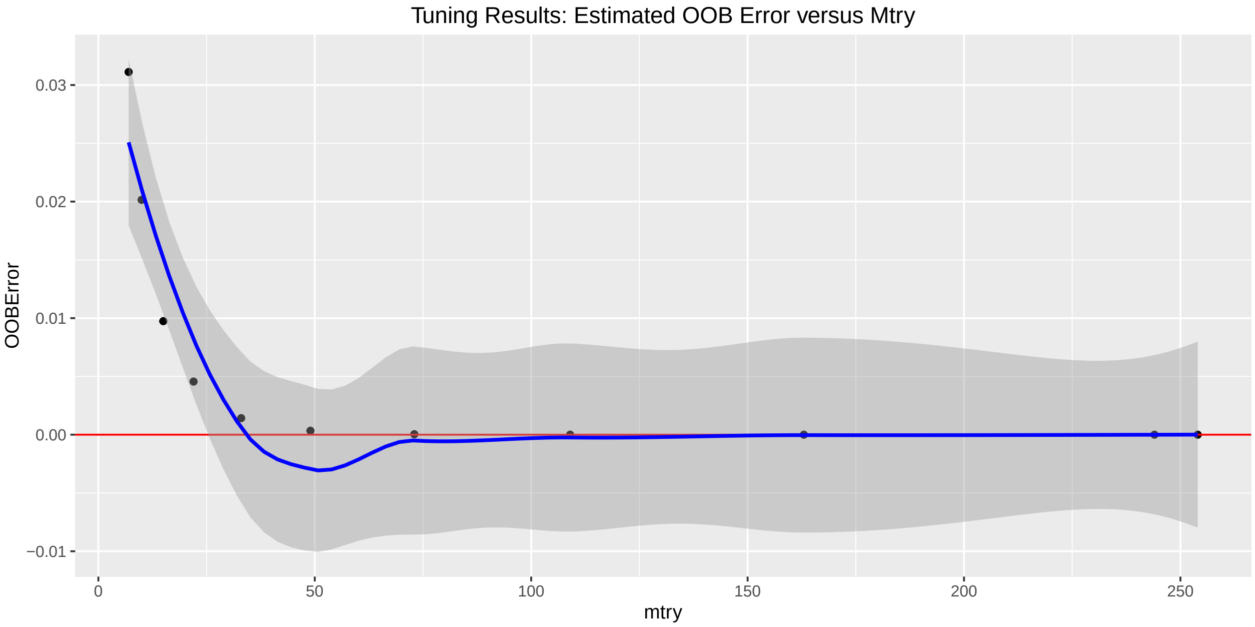


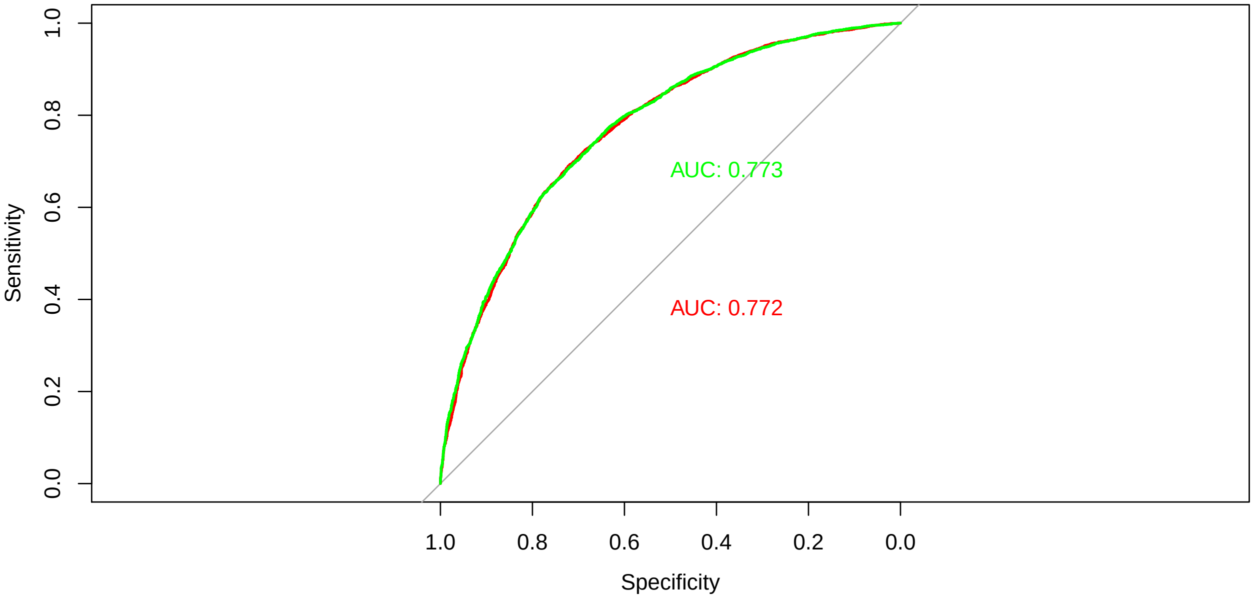
Although the logistic regression using the dataset with daily tweets capped at 50 technically yields a higher AUC value, the output above suggests that there is no substantial difference between using the two datasets.

With regards to the threshold value for classification, the gender context of this study is such that only a 50% threshold values makes sense. In other words, it makes no sense to more heavily penalize misclassification as male over misclassification as female (or vice versa).

Model 2: Random Forest - All Variables

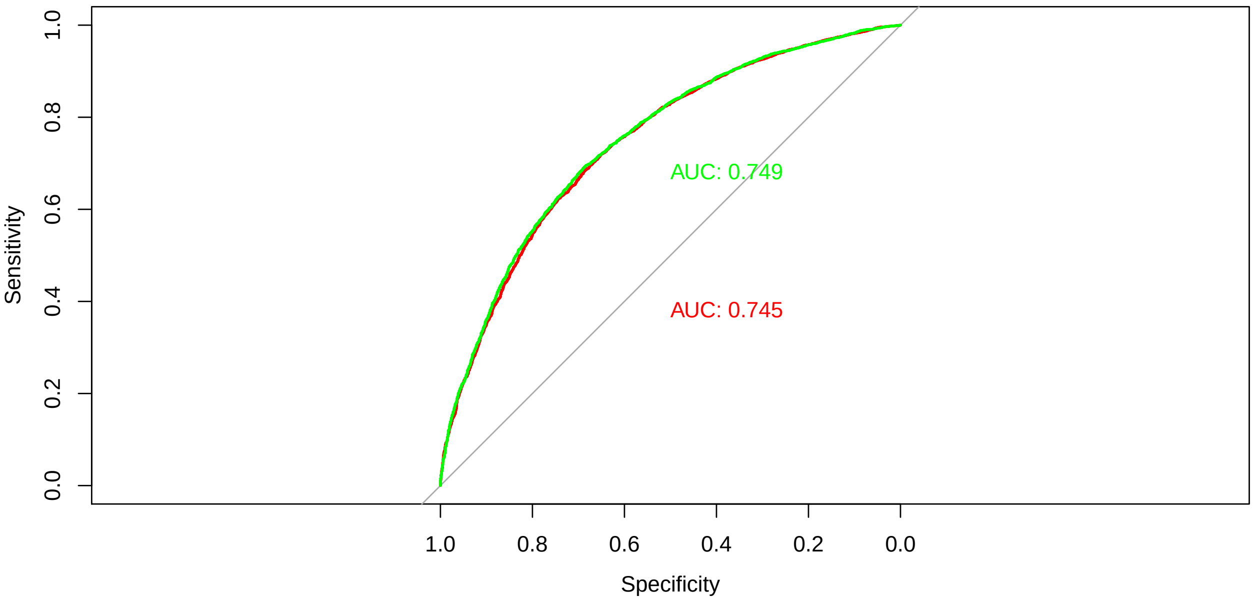
After tuning the algorithm parameters, we random forest models were then built using 1,000 trees and an mtry value of 22. Again, both subsets of data with daily tweets capped at 50 and 100 were considered. As exhibited in the ROC comparison below, the model using data with daily tweets capped at 50 once again produced a slightly higher AUC but the outputs are ultimately identical for practical purposes. It is, however, worth noting that that the random forest approach yielded a slightly better performance compared to the logistic model approach in both data subsets as seen in the higher AUC values. Similar to the case of the logistic model, the performance of the random forest models peaks at around a classification threshold value of 50%.


Tuning Results: Estimated OOB Error versus Mtry
OOBError
0.03
0.02
0.01
0.00
−0.01
0
50
100
150
200
250
mtry


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.773
AUC: 0.772
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

Model 3: Random Forest - Elastic Net Variables

Finally we ran another Random Forest using only the Elastic Net variables to compare performance. This model was the worse performer, suggesting that important variables were ignored.



Concluding Remarks

The results of our study not only provide strong support for our initial view on the presence of language use differences between genders on Twitter but also corroborate the conclusions of existing studies that link gender and social media activity. The predictive accuracy of our models peaks at approximately 70%; although certainly less than ideal, we believe that this figure is acceptable given that the study??s analysis was based on a single random tweet message per user profile. In future research endeavors that follow the statistical modeling processes outlined in this study, we believe that the predictive performance of the models can be significantly improved by compiling and using a greater number of tweet messages per user profile, giving the constructed models more data to work with.

Bibliography

[1] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. https://doi.org/10.1371/journal.pone.0073791

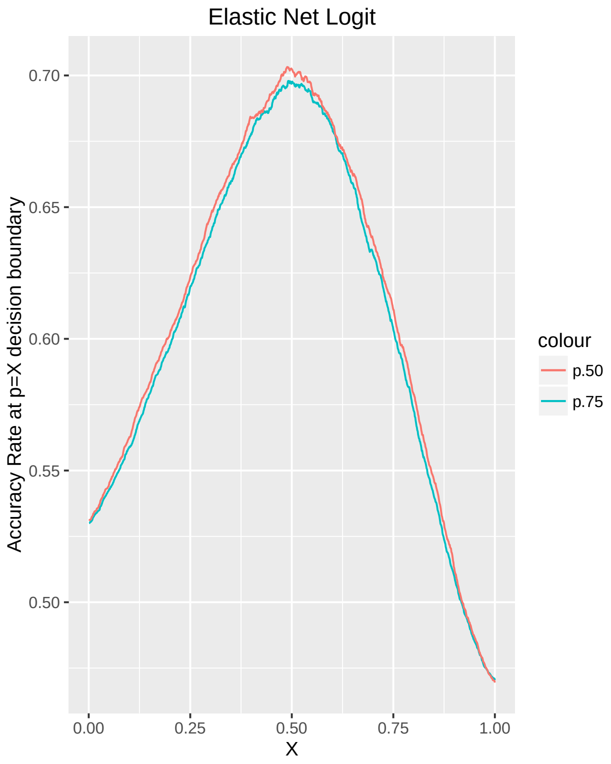
[2] Coviello L, Sohn Y, Kramer ADI, Marlow C, Franceschetti M, Christakis NA, et al. (2014) Detecting Emotional Contagion in Massive Social Networks. PLoS ONE 9(3): e90315. https://doi.org/10.1371/journal.pone.0090315

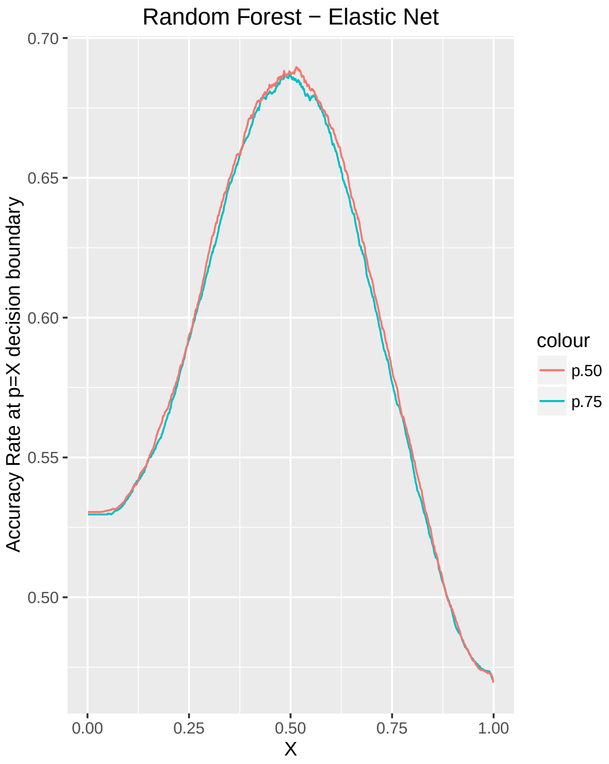
[3] M. Kosinski, D. Stillwell, and T. Graepel, ??Private Traits and Attributes are Predictable From Digital Records of Human Behavior,?? in Proceedings of the national Academy of Sciences of the United States of America, 2012. [Online]. Available: http://www.pnas.org/content/110/15/5802.short.

[4] Kaggle, https://www.kaggle.com/crowdflower/twitter-user-gender-classification

Appendix

Other Model Results





Python Code (See Rmd file)