Predicting Age of User from Product and Review
CSE 190 Spring 2015

Joseph Luttrell
University of California, San Diego
jluttrell@ucsd.edu

Spenser Cornett
University of California, San Diego
scornett@ucsd.edu

ABSTRACT-
Knowing an online user's age can go a long way in deciding which products to recommend and what advertising would be most effective. However, this information is not always available. Sometime the user keeps birthday information private, doesn't provide it, or even lies about their age. In this paper we use beer reviews to predict a person's age based on features of the product and features of the review. It was hypothesized that certain age groups would buy certain products and that the ratings, text, date, and time of a review could be used to determine a user's age. We built multiple models using linear regression to explore this task. Our results prove that using this type of analysis could give insight to a person's age.

INTRODUCTION
The online marketplace accounts for a large percentage of consumer purchases. These websites use all types of marketing techniques based on user demographics and activity. However, this information is not always readily available to the companies. One aspect that can play a large role in marketing to a user is their age. Age can be a good indicator of what products to (or not to) recommend, what advertisements would be most (or least) effective, and whether or not a particular user belongs on a certain site. In this paper we use reviews of beer to try and predict how many months old a person is based on the review and the product alone.

Many websites don't require a user to supply their birthday when signing up and even those that do may not make that information public. Using machine learning we aim to try and fill in this gap. Our goal was to use only features of the review and the product bought to predict the user's age. The only user feature we use in predicting age is their gender. The main reason for this is to see if we could accurately predict a user's age from a "cold start" meaning that we've never seen the user before. We used linear regression to predict their age and mean absolute error (MAE) to measure accuracy. The baseline used was just guessing the age based on the global average age. We then trained two models: one using just product features and one using just review features. Our final model incorporated both product and review features.

DATASET
i. *Overview*
We pulled our data from just over 1.5 million beer reviews from beeradvocate.com collected by Standford's Large Network Dataset Collection [1]. The data spans over 10 years up to November 2011. Each beer review contains the following fields:

- Beer
  - Name, Style, brewer ID, alcohol by volume (ABV)*
- Review
  - Ratings (overall, appearance, palate, taste, aroma), time and date, text
- User
  - Profile name, birthday*, gender*

*Not all reviews contained this field

The dataset included 33,388 users, 56,857 different beers, and 104 different beer styles.

ii. *Preprocessing*
We only wanted reviews from users who also provided their age so that we could effectively train and test our models. This brought the dataset down to 327,596 reviews from 4,910 users. After exploring our dataset, we found that there were a few users who were over 100 years old. Specifically we found users whose reported ages were 111 and 118 years old. We figured that these were most likely falsely reported outliers so we decided to only use reviews from users between the ages of 18 and 80. A user must be 18 years old to be able to sign up on Beer Advocate. We also decided that gender and ABV could possibly be good features in determining age so we only used reviews that included the gender of the user and the ABV of the beer, which only slightly decreased our useable data. Finally, we found 42 reviews within this subset that had no review text so we removed these as well since the review itself is an important feature in our model. The final dataset used for training and testing our models

included 304,582 reviews from 4,778 users. The reviews covered 24,042 beers that were from 104 different types of beer.

iii. *Exploratory Analysis*

We explored certain features of the data to assess their usefulness in our model. We were particularly interested in temporal dynamics, correlation between ratings and age, correlation between age and styles of beer, and the correlation between reviewer gender and age on average. The average age(years) of users in the dataset was 30.3979 with a standard deviation of 8.171 years and the median age was 29. The data was very heavily concentrated with relatively younger users which we suspect may have made our prediction task tougher since the data was not evenly distributed across age groups. The following graphs show our preliminary findings on the useful features described above. Figure 8 illustrates the concentration of younger users in the dataset, clearly user activity itself would not have been a valuable feature due to the lack of correlation.

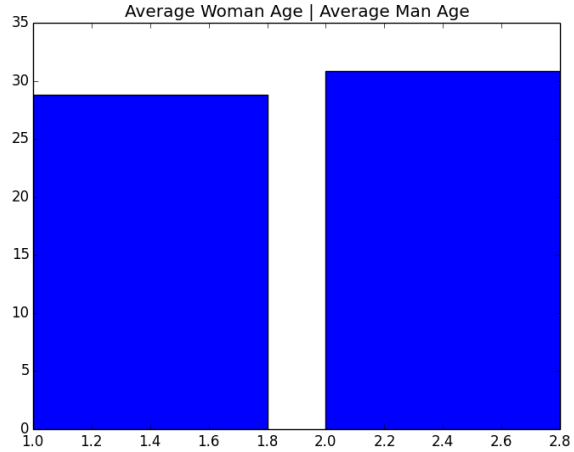

Fig 1

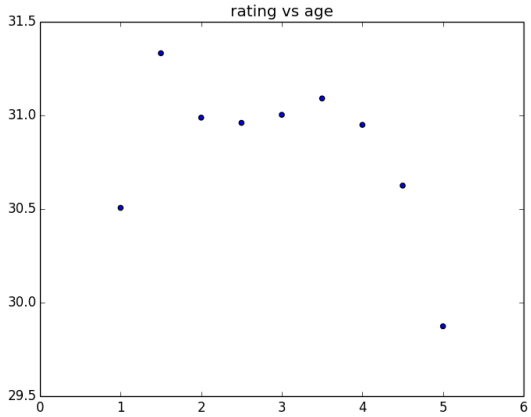

Fig 2

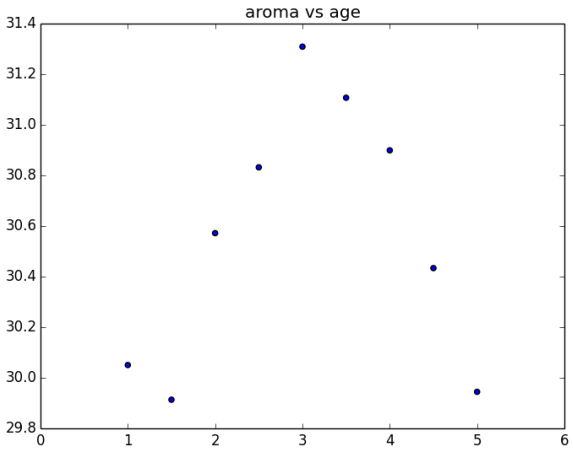

Fig 3

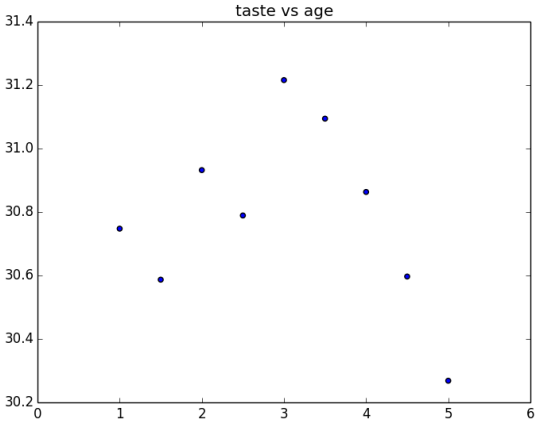

Fig 4

Figure 1 shows that both male and female average ages are similar. Although it may not give much insight to the age directly it could give insight into text analysis since males and females are known to use different language [4]. Figure 2-4 are age trends in rating data. It shows that younger people tend to rate beers higher than older people. Figure 5 shows a nice trend and the correlation between the hour of the review and the average user's age. Figure 6 shows a positive linear trend for what users buy from which brewery. Figure 7 and 8 show an almost positive linear trend in age for the year and day of the week.

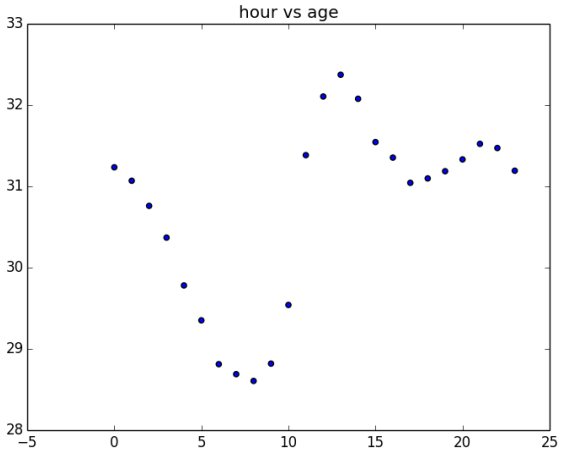

Fig 5

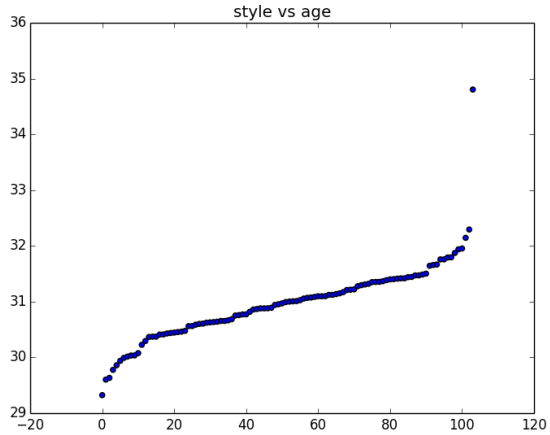

Fig 6

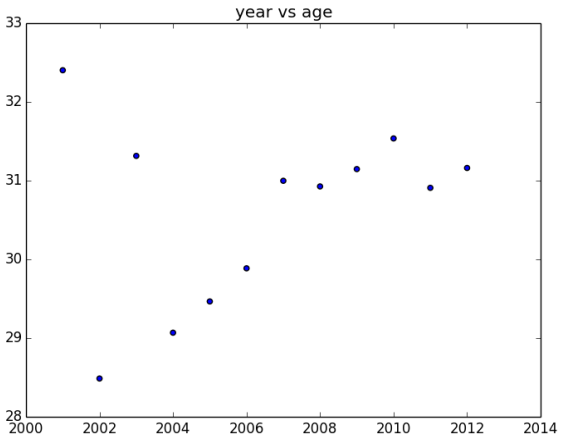

Fig 7

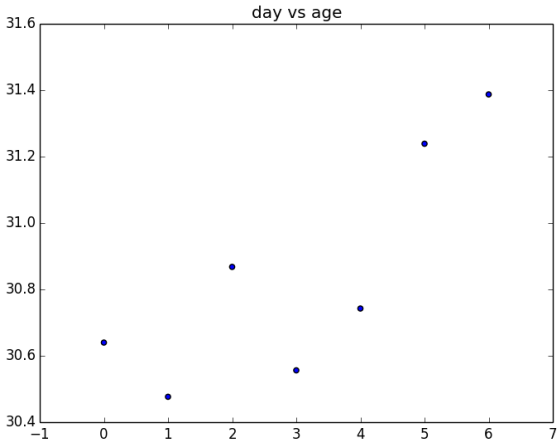

Fig 8

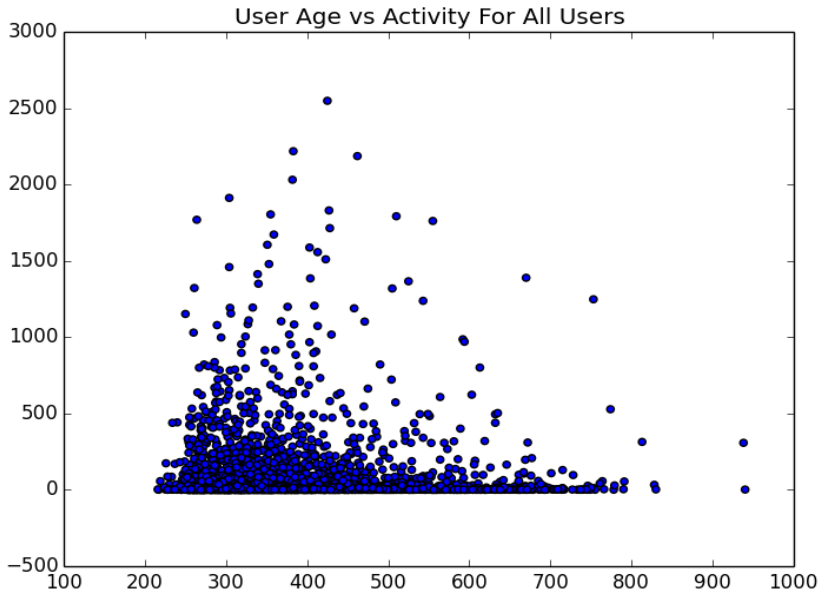

Fig 9

PREDICTIVE TASK

The goal of our project was to predict a user's age based on the characteristics of a single review. We decided to use linear regression on specific features of the review in order to predict the user's age. We used two base models and then combined them both resulting in a third model. The first model used features concerning the beer itself while the second model used features concerning the actual review. Since age is a continuous value, it was an easy choice to use regression for our predictions. Our model should be as follows:

$$f(review) = age_{predict}$$

To evaluate the accuracy of our models we use the mean absolute error (MAE) in which we aim to minimize the equation:

$$MAE(T) = \frac{1}{|T|} \sum_{r(review) \varepsilon T} |f(review) - r(review)|$$

RELATED LITERATURE

This data was used in [2] to model a user's expertise over time and use that model to predict products. The user's age was not directly used as an indicator of expertise but rather their age of development. This proved to improve recommendation performance.

Many others have tried to predict age from text. Paper [3] predicts a user's age using linear regression on text of blogs and forums. We used some similar features, namely unigrams and gender. Their average MAE was found to be similar to ours. They concluded that unigrams were a strong indicator of age. They also used part of speech (POS) unigrams and bigrams which they say helped in predicting older people's age. Santosh et al. [4] also tried predicting age and gender from blogs. Their research was mainly based on the topic of the text. Since all of our text comes from beer reviews, it would not quite be as beneficial to do topic analysis. However, they also noted the importance of using n-grams and POS ngrams.

FEATURES

i. *Beer Features*

- Beer Style: there were 104 different beer types present in our dataset. We implemented this feature with a 104-D binary vector where the position corresponding to the beer style was equal to one and all other positions were set to zero. Based on the provided graph, beer style appeared to correlate with age. This fact also led to our addition of the next three features concerning the beer style.
- Average Age for Beer Style: We appended the average age of users who reviewed that specific beer type to the features. If a style is not present in the training set, we use the global average user age.
- Popularity of Beer Style: We appended the number of reviews for the specified beer style in an attempt to quantify the popularity of that beer style.
- Max and Min age of reviewers for beer style: We included the maximum and minimum age of reviewers for the specified beer style in the training set. Unseen styles were given the global min and max of 18 and 78 respectively.
- ABV: we used the ABV of the beer being reviewed as a feature as well in binary form. We have ranges from 0 to 13 ABV in increments of 0.5, then a range from 13 to 20, and finally all positions will be zeros if the ABV is above 20.
- Brewer ID: given the brewer ID, we append the average reviewer age for that brewery to the features. If the brewery did not show up at training time, we use the global average.

ii. *Review Features*

- Unigrams were used in our review features similarly to [3] and [4]. They were found to result in a relatively significant improvement on our MAE.
- Bigrams were also used in our review features similar to the POS bigrams used in [3], which also had a relatively significant improvement on our MAE.
- Length of review without stop words is appended to the review features in our model. It was found to improve our MAE so we kept it in the model.
- We ran sentiment analysis on the review text and used the sentiment and the subjectivity rating as supported in [6].
- Average sentence length and ratio of capital letters were taken from the review text and used as a features. This is supported by [5] "Younger people use more alphabetical lengthening, more capitalization of words, shorter words and sentences, more self-references, more slang words, and more Internet acronyms "
- Temporal Dynamics: The hour of the review is included as a binary feature. The motivation for this is strongly supported by the graph of hour vs average age included in the exploratory analysis. We also used the year of the review and day of the week in our temporal dynamics features because of the positive correlation between these two features and reviewer age as shown in the exploratory analysis.

- Rating values for the different review categories with meaningful correlations to age based of the exploratory analysis were used as features.
- Gender was used to help predict age based off the slight correlation found in the dataset during exploratory analysis and also in accordance with [3].

MODELS

The baseline model we used was the average age of all users in the training set. Our goal was to create models that improved upon this simple baseline.

Our first model used linear regression based solely on the review features:

$$p_{review} = \sum_{u \in unigrams} \Theta_u \cdot \delta_u(u) + \sum_{b \in bigrams} \Theta_b \cdot \delta_b(b) + \Theta_{length} \cdot length + \Theta_{hr} \cdot hr + \Theta_{day} \cdot day + \Theta_{category} \cdot \delta_c(category) + \Theta_{gender} \cdot \delta_g(g) + \sum_{r \in ratings} \Theta_r \cdot r + \Theta_{sent\ length} \cdot length_{sentence} + \Theta_{sentiment} \cdot sentiment + \Theta_{subjectivity} \cdot subjectivity + \Theta_{capitals} \cdot R(capitals)$$

Where $\delta_u(u)$=count of unigram u, $\delta_b(b)$ = count of bigram, $\delta_c(c)$=1 if c == category or 0 otherwise, $\delta_g(g)$=1 if gender == male or 0 if female, and R(capitals) = $\frac{number\ of\ capital\ letters}{letters}$

Our second model uses only features of the beer:

$$p_{beer} = \Theta_{style} \cdot \delta_s(style) + \Theta_{style\ age} \cdot AvgAgeStyle + \Theta_{style\ age} \cdot AvgAgeStyle + \Theta_{style\ popularity} \cdot StylePopularity + MaxAgeofStyle + MinAgeofStyle + \Theta_{ABV} \cdot ABV + \Theta_{Brewer} \cdot Brewer$$

Where $\delta_s(style) = 1$ *if* $s == style$, 0 *otherwise*

The final model was a combination of both models: $p = p_{review} + p_{beer}$

RESULTS

Model	MAE (In Months)	# Features
Baseline	73.542841	1
Product Based	70.631780	7
Review Based	62.833660	14
Combined	62.499620	21

Table 1: Prediction Results

This prediction task proved to be very difficult. Our baseline was hard to beat originally until we figured out which features were the most powerful. We believe the baseline performed so well because the dataset is so heavily concentrated with younger users. The baseline predicts reviewer age with an MAE in months of 73.54 (6.13 years). Using only the product based model we were able to achieve an MAE of 70.63 (5.88 years) which is a 4.1% improvement over baseline. In our beer model we ended up using only features of the beer styles and the brewers but not features of the specific beer because we found that there weren't enough examples of every specific beer to provide useful data but with only 104 styles of beer and 3,513 different brewers, we were able to use those to create more useful features. We were disappointed with the performance of our product model but we expected much higher performance from our review model anyway. Our review based model resulted in an MAE of 62.83 (5.24 years) or roughly a 14.6% improvement over baseline. It took us a while to figure out the right combination of features to really improve the performance of our review model. Bigrams and Unigrams (Top 350) from the review text were found to be very powerful features and made a valuable contribution towards performance. We also found that rounding down to the nearest year provided us with a boost in prediction accuracy, this is most likely due to our use of MAE as our performance standard and also the fact that the dataset was heavily concentrated with younger users.

After creating and evaluating these two models we combined them to form an integrated model. Basically we took all the features from both models and used them all in our feature vectors for training and testing. This combined model outperformed its individual components slightly. The combined model achieved an MAE of 62.5 (5.2 years), a 15.1% improvement over baseline. Clearly features of the review itself are the most effective for prediction. We were able to predict user age within +/- 5.2 years on average, this is roughly a 10-year range, which corresponds to the ranges used for age targeted advertising. Our

model appears to be accurate enough to place users within their correct market on average.

Top 5 Positive Unigrams	Top 5 Negative Unigrams	Top 5 Positive Bigrams	Top Negative Bigrams
23.2491867175 lace	-25.9328724202 retention	18.5655957746 head and	-17.6337322133 followed by
15.6112226707 nose	-19.5502539045 colour	18.334692303 finish is	-16.4611612009 12 oz
14.0466545096 followed	-18.2477823152 amount	14.8784667792 was a	-15.0491248654 the nose
13.6377965354 poured	-15.0528031739 m	13.6126596956 carbonation is	-14.8186045279 up front
11.6423560104 session	-12.6692395023 tap	10.9630246038 oz bottle	-14.0738905767 little bit

Table 2: Top Unigram and Bigram weights

CONCLUSION

In this paper we explored ways to predict a user's age based on beer reviews. It was found that using text analysis had the most significant impact on lowering the MAE. Although product features did not perform as well as anticipated, perhaps with a more sophisticated model and a larger dataset they could improve performance. Future work could also take into account a user's entire review history. Utilizing more reviews and review text would lead to greater insight to a user's age. Other similarity measure techniques such as pearson correlation could also help improve the accuracy. This would entail finding the user most closely related to another known user and using the known user's age as an indicator of the age of the unknown user. Overall we were satisfied with our results of getting within ~5 years on average based solely on one review but would like to explore further into the possibilities of accurate age prediction.

REFERENCES

[1] Jure Leskovec and Andrej Krevl. Stanford Large Dataset Collection. http://snap.stanford.edu/data/. June 2014.
[2] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. CoRR, abs/1303.4402, 2013.
[3] Dong Nguyen, Noah A. Smith, and Carolyn P. Rose. Author Age Prediction from Text using Linear Regression.
[4] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author Profiling: Predicting Age and Gender from Blogs. Notebook for PAN at CLEF 2013.
[5] Dong Nguyen, Dolf Trieschnigg, A. Seza Dogruoz, Rilana Gravel, Mariet Theune, Theo Meder, Franciska de Jong. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment.
[6] Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. Prediction of Age, Sentiment, and Connectivity from Social Media Text.