The Predictive Power of Essays on OkCupid

Tzu-Han Zoe Cheng
A53238806
Cognitive Science
University of California, San Diego
San Diego, CA, USA
tzcheng@ucsd.edu

Sachin Deshpande
A14498623
Computer Science & Engineering
University of California, San Diego
San Diego, CA, USA
scdeshpa@ucsd.edu

Anshuman Dewangan
A59001372
Computer Science & Engineering
University of California, San Diego
San Diego, CA, USA
adewanga@ucsd.edu

Sing Wong
A14764052
Computer Science & Engineering
University of California, San Diego
San Diego, CA, USA
siw011@ucsd.edu

Abstract

During the era of COVID-19, dating has become even more difficult for people. Even when people could meet in person, there are often issues with dating profiles being incomplete or people “catfishing” others by creating profiles that differ from their true persona. Our project is aimed at predicting certain user features (age, gender, whether or not they are a programmer, etc.) given part of their dating profile in an effort to solve these issues. In this paper, we will first summarize an OkCupid (a dating application) dataset and describe our goals to predict a user’s gender, age, body type, and whether or not they are a programmer. Next, we will review contemporary, cutting-edge models that have been developed, several of which are the focus of research papers. Then, we analyze our attempts, where the specifics and motivations for particular models can be found, and give appropriate error metrics describing how our models perform. At the end is a collection of references for the aforementioned research as well as additional exploratory analysis conducted beyond the predictive tasks.

1. Dataset

We explored a dataset consisting of user profile data for 59,946 users on OkCupid (a free online dating website) from the San Francisco Bay Area in June 2012.[1]

1.1 Background

The dataset was provided by Albert Kim from Middlebury College and Adriana Escobedo-Land from Reed College in their paper *OkCupid Data for Introductory Statistics and Data Science Courses* pushed in the *Journal of Statistics Education* Volume 23, Number 2 (2015). The data was scraped from public profiles using a Python script.

1.2 Data Overview

The data consists of 59,946 entries of users with 31 columns representing demographic information and text responses to essay questions posed to all OkCupid users.

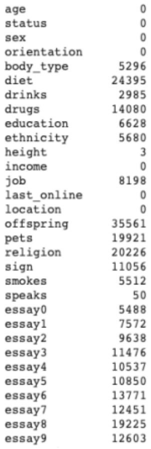
1.2.1 Demographic Variables. The demographic variables are:

1. **Sex**: “male” or “female”
2. **Age**: in years
3. **Height**: in inches
4. **Last Online**: day, month, year, hour, and minute
5. **Status**: “single,” “seeing someone,” etc.
6. **Orientation**: “straight,” “gay,” “bisexual”
7. **Ethnicity**: “white,” “Asian,” “Hispanic,” etc.
8. **Diet**: “anything,” “vegetarian,” “vegan,” etc.
9. **Body Type**: “average,” “athletic,” “thin,” etc.
10. **Location**: city within 25 mi of San Francisco
11. **Education**: graduated from, working on, dropped out of high school, college, PhD, etc.
12. **Job**: “student,” “artistic/music/writing,” “science/tech/engineering,” etc.
13. **Income**: dropdown of choices in $
14. **Offspring**: has kids, wants kids, doesn’t want kids, etc.
15. **Pets**: has dogs/cats, likes dogs/cats, dislikes dogs/cats
16. **Religion**: “Christianity,” “Hinduism,” “agnosticism,” etc.
17. **Sign**: “Pisces,” “Taurus,” “Gemini,” etc.
18. **Drinks**: “often,” “socially,” “not at all,” etc.
19. **Drugs**: “often,” “sometimes,” “never”
20. **Smokes**: “yes,” “sometimes,” “no,” etc.
21. **Speaks**: English (fluently, okay, poorly), Spanish, etc.

1.2.2 Essays. The essays are:

1. **Essay 0**: My self summary
2. **Essay 1**: What I’m doing with my life
3. **Essay 2**: I’m really good at
4. **Essay 3**: The first thing people usually notice about me

5. **Essay 4**: Favorite books, movies, show, music, and food
6. **Essay 5**: The six things I could never do without
7. **Essay 6**: I spend a lot of time thinking about
8. **Essay 7**: On a typical Friday night I am
9. **Essay 8**: The most private thing I am willing to admit
10. **Essay 9**: You should message me if



```
age                0
status             0
sex                0
orientation        0
body_type       5296
diet           24395
drinks          2985
drugs          14080
education       6628
ethnicity       5680
height             3
income             0
job             8198
last_online        0
location           0
offspring      35561
pets           19921
religion       20226
sign           11056
smokes          5512
speaks            50
essay0          5488
essay1          7572
essay2          9638
essay3         11476
essay4         10537
essay5         10850
essay6         13771
essay7         12451
essay8         19225
essay9         12603
```

Figure 1: Distribution of Null Values in Dataset

1.3 Analysis of Variables

We will now proceed with an exploratory analysis of variables in the dataset. In order to keep the most pertinent information in the main body of the paper, here we explore only the variables relevant to our predictive tasks. In Appendix A, we include our exploration of the rest of the variables in the dataset.

1.3.1 Sex. As shown in Figure 2, there are 35,829 (59.8%) users who identify as Male and 24,117 (40.2%) users who identify as Female. This corroborates with the suggestion that dating apps are male-dominated, but this dataset seems less imbalanced than other dating apps (especially newer apps).

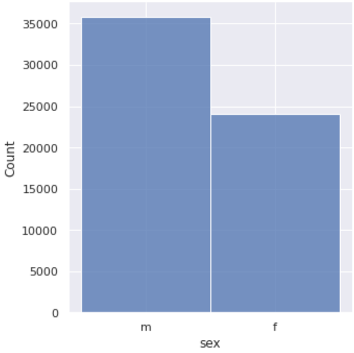
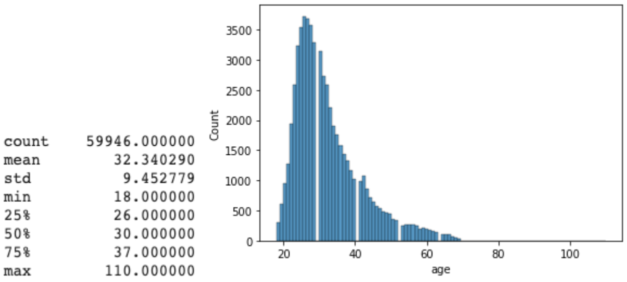


Figure 2: Distribution of Sex

1.3.2 Age. As shown in Figure 3, ages range from 18 (the legal age to join a dating app in the US) to 110, with a mean of 32 and skewed right. The median age is 30.



```
count    59946.000000
mean        32.340290
std          9.452779
min         18.000000
25%         26.000000
50%         30.000000
75%         37.000000
max        110.000000
```

Figure 3: Distribution of Age

There are two users who reported ages above 100; one seems to be a fake account and the other seems to have erroneously entered their age.

1.3.3 Body Type. As shown in Figure 4, most users identify as "Average," followed by "Fit" and "Athletic."

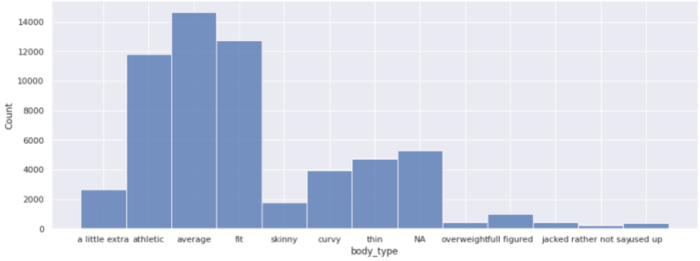


Figure 4: Distribution of Body Type

We imagine that of those who choose not to respond (ie. "NA"), there would be a higher representation of "undesirable" body types.

1.3.4 Speaks. Users can identify speaking multiple languages. As shown in Figure 5, most users speak one language, with some users speaking up to five languages.

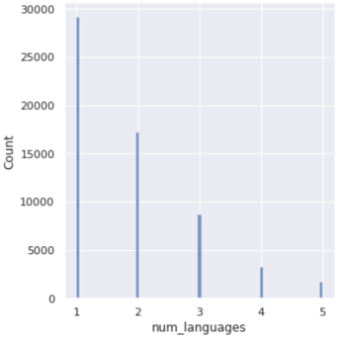
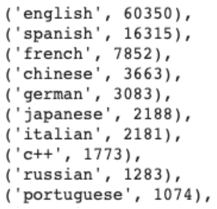


Figure 5: Distribution of Number of Languages Spoken

As shown in Figure 6, “English” was the predominant language, followed by “Spanish” and “French;” California high schools have a foreign language requirement, with Spanish and French being the primary languages taught.

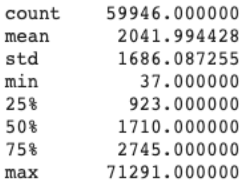


```
('english', 60350),
('spanish', 16315),
('french', 7852),
('chinese', 3663),
('german', 3083),
('japanese', 2188),
('italian', 2181),
('c++', 1773),
('russian', 1283),
('portuguese', 1074),
```

Figure 6: Distribution of Languages Spoken

Interestingly, C++ was a top 10 language, alluding to Silicon Valley geek culture.

1.3. Essays. While each essay had different content and lengths depending on the question it was answering, Figure 7 shows the distribution of lengths of the combined essays for each user.



```
count     59946.000000
mean       2041.994428
std        1686.087255
min          37.000000
25%         923.000000
50%        1710.000000
75%        2745.000000
max       71291.000000
```

Figure 7: Distribution of Length of Combined Essays

The user with the longest combined essay is either a fake account or crazy—it’s hard to tell.

1.4 Variable Interactions

We will now explore interesting interactions between variables.

1.4.1 Gender Differences. First, we look at the distribution of age across gender. As shown in Figure 8, there is no appreciable difference between the distribution of ages between Males and Female.

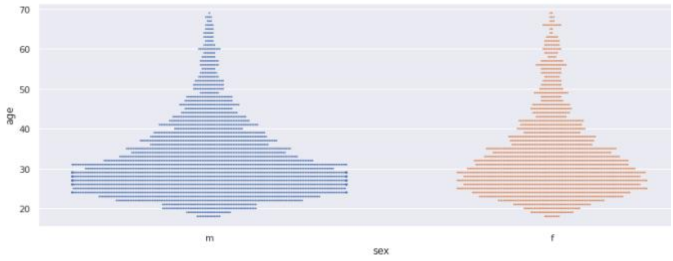


Figure 8: Distribution of Age across Genders

Next, we explore if gender affects the likelihood of having children. In Figure 9, we notice there are slightly more women with children than men with children. This resonates with national stereotypes that women are more likely to be single parents than men.

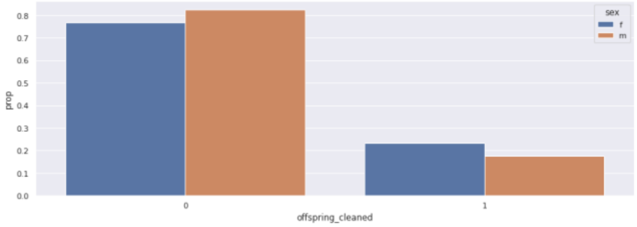


Figure 9: Distribution of Offspring across Genders

Figure 10 shows that men are more likely to hold jobs in “Science,” “Banking,” “Executive,” “Entertainment,” “Computer,” and “Construction.” The gender differences in “Science” and “Computer” are especially big. Women are more likely to hold jobs in “Medicine” and “Education” by large margins.

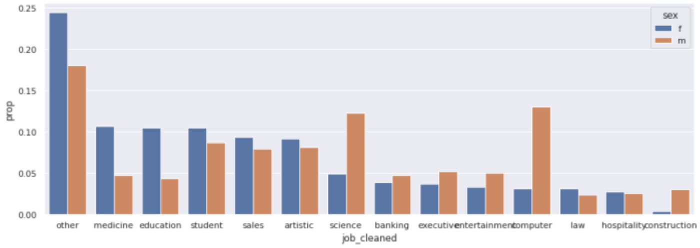


Figure 10: Distribution of Jobs across Genders

As shown in Figure 11, not only do men earn approximately 30% more than women on average, but they also earn higher incomes for every job function except for “Hospitality.” The glass ceiling is real, folks!

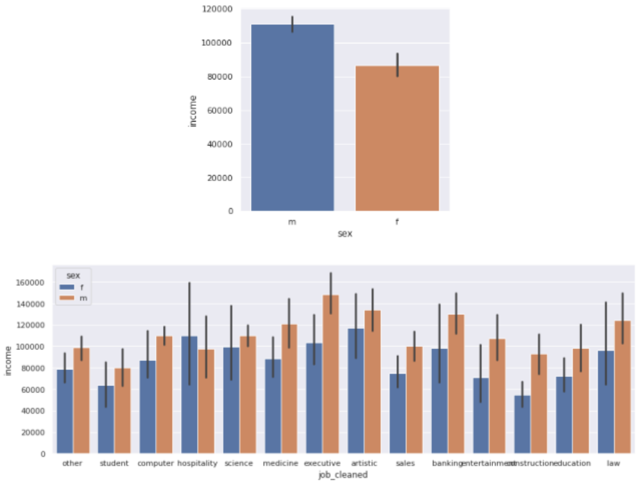


Figure 11: Distribution of Income across Genders

1.4.2 Body Type & Diet. Men are more likely to describe themselves as “Fit” or “Athletic” than women. Women are more likely to describe themselves as “Curvy,” “Thin,” or “Full Figured” than men, following gender stereotypes for desirable body types.

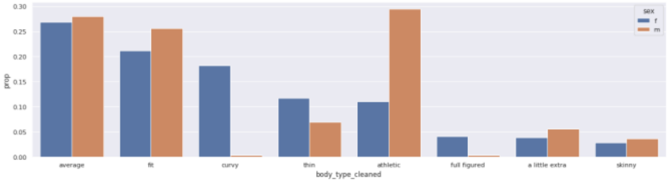


Figure 12: Distribution of Body Type across Gender

Users who eat “anything” are more likely to be “average” body type, while “vegetarians” and “vegans” are more likely to be “fit,” “skinny,” or “thin.”

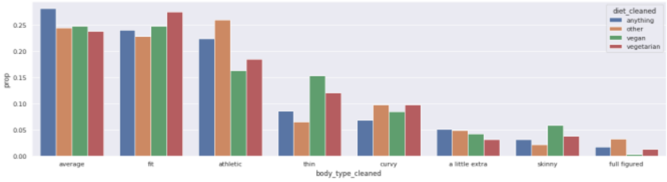


Figure 13: Distribution of Body Type across Diet

More educated users are more likely to be “fit,” “athletic,” or “thin,” whereas less educated users are more likely to be “average,” “a little extra,” or “skinny.” Presumably, higher educated users know how to take care of their bodies better.

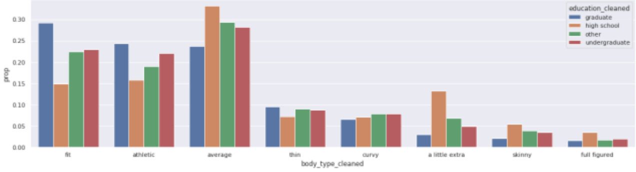


Figure 14: Distribution of Body Type across Education

2. Predictive Task

We decided to define multiple tasks to predict various user attributes from text-based features of the essays, in an effort to see how much insight into demographic information can be gleaned from the text in those essays.

The user attributes we aim to predict are listed below:

- **Gender**: classification, binary, {Male, Female}
- **Age**: regression, real-valued, (1, 110)
- **Body Type**: multi-class classification, categorical, {Fit, Average, Fat, Thin, NA}
- **Speaks C++** (i.e. does the person speak C++ as one of their selected language?): classification, binary, {Yes, No}

We considered predicting other demographic variables available in the dataset, yet rejected most due to the extreme class imbalance or a low expectation of finding anything conclusive.

2.1 Model Evaluation Metrics

For evaluating our model, we will use:

- **Accuracy**: for categorical predictions where our labels are somewhat balanced (Gender, Body Type)
- **MSE**: for quantitative predictions (Age)
- **BER**: for categorical variables in addition to accuracy for when our labels are not balanced (speaking C++)

Accuracy and MSE are standard metrics used in classification and regression problems, respectively. BER helps determine model improvement when our labels are extremely imbalanced, as in the case of predicting whether a user speaks C++.

2.2 Baselines for Evaluation

Our trivial baselines are as follows:

- **Gender**: Predict Male, as it is the predominant gender. This yields about a 60% accuracy rate.
- **Age**: Predict the i) mean or ii) median age of all seen people. Either of these yields an MSE of roughly 90.
- **Body Type**: Predict ‘average’, as it is the most common body type. This yields an accuracy of 24%.
- **Speaking C++**: Predict ‘No’, as most people aren't programmers. This yields an accuracy of 97%, but a BER of .5.

These baselines were left extremely trivial as to not make any assumptions regarding how text could be predictors of these variables. Our goal is to beat these baselines significantly.

2.3 Assessing Model Validity

We will construct a regularization pipeline by randomizing the entries of the dataset and splitting it into train (~50% of data), validation (~17% of data), and test (~33% of data) sets. We will train our models on the training set, tune hyperparameters using the validation set, and report final evaluation metrics using the test set.

For all of our models, we will evaluate their appropriate error metric on the test set, as described above. Models with lower error rates on the test set will be assumed to be better than those with higher error rates, as usual.

2.4 Model Specifications

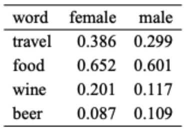
More details about these models are given in Sections 4-5. There, we talk about how features/text are processed, more specifics about the model(s) we chose, what results we were able to obtain, and other specifics omitted here.

3. Literature review

3.1 Source Paper

Before we enumerate our model specifications, we will detail techniques from related literature. The source paper in which the dataset originated from conducted an exploratory analysis of the data. The paper looked at five different analyses: comparison of heights across gender, comparison of sexual orientation across gender, comparison of text from essays across genders, and prediction of height using gender.

In the analysis most relevant to our objective, the comparison of text from essays, the authors explored the top 25 words used by each gender, after removing stopwords and punctuation. The paper also calculated the proportion certain words appeared for essays across genders with some interesting findings as shown in Figure 15:



word	female	male
travel	0.386	0.299
food	0.652	0.601
wine	0.201	0.117
beer	0.087	0.109

Figure 15: Proportion of Words in Essays across Gender

No specific methods from the paper were directly relevant in our predictive tasks.

3.2 Literature Related to OkCupid Data

Data from dating sites is frequently studied to understand concepts such as relationship behavior[2], social self-presentation[3], or even bias and discrimination[4].

A similar dataset of 68,371 OkCupid profiles published in 2016 made the news as a transgression against data ethics. The argument was that just because the data can be scraped publicly, doesn't mean it is ethical to aggregate and distribute widely. Ethics aside, the paper conducts an exploratory analysis of user location, age, and other variables. Instead of 10 essays, OkCupid changed its format to have users respond to any number of 46 True/False or short answer questions. The paper studied the relationship between cognitive ability ("Do you believe the world is flat?"), religion ("Is duty to God the most important thing in your life?"), politics, and Zodiac sign in combination to specific responses to those questions. Where there was overlap, the exploratory analysis had similar results to ours.

Several papers built upon these publicly available OkCupid datasets. Shishido, Narasimhan, and Haller looked at the relationship between age, drug usage, and the text from the 10 essays. They preprocessed the text by removing punctuation and converting to lowercase. They also combined some categories for the demographic variables which served as inspiration for our exploratory analysis above. For methods, they looked at tf-idf, log-odds-ratio, non-negative matrix factorization, and permutation testing. While the gender analysis was more exploratory, the prediction of drug usage resulted in a 72.7% accuracy.[6]

Pandit did an analysis of strategic self-representation using education, race, height, and fitness level. They used topic modeling (latent Dirichlet allocation and non-negative matrix factorization), vector space models (Word2Vec), cluster analysis (DBSCAN), and structural topical models.[3]

Finally, it's worth noting that Christian Rudder, CEO of OkCupid, wrote a book, *Dataclysm*, on insights gleaned from social network data along the themes of education, race, and politics.[7]

3.3 Literature Related to Predicting Demographics from Text

We also looked at papers that predicted demographics from text outside of the context of dating apps.

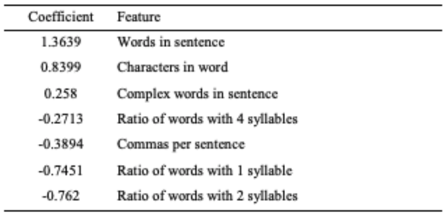
3.3.1 Gender Prediction. One paper generated a variety of text-related features outside of unigrams/bigrams, including parts of speech, production rules, and interactions among all variables. They also tested a variety of models including SVM, Naïve Bayes, random forest, and logistic regression. Their best model used SVM with context-free-grammar features for an accuracy of 82.81%, and improvement upon their unigram model which had an accuracy of 69.33%.[8] Our unigram/bigram model performed in between the two with an accuracy of 75%.

In a paper on gender classification using deep learning, Bartle and Zheng used a Windowed Recurrent Convolutional Neural Network to achieve 86% accuracy on blog posts using bag-of-words, paragraph2vec, and parts of speech features.[9] Ciccone et. al. achieved 80% accuracy using n-grams, bag-of-words, tf-idf, and linear support vector classification.[10] Burger et. al. achieved 77% accuracy using n-grams and 90% accuracy using other demographic variables available in their dataset.[11] Mukherjee and Liu achieved 88% accuracy using frequency measures, stylistic features, and parts of speech.[12]

In general, accuracies were within the same range as our results. For future directions, we should consider other text-based features such as parts of speech.

3.3.2 Age Prediction. Abdallah et. al. defined their age prediction as a classification problem, with a positive class as age > 35. They achieved an 82.3% accuracy with context-free-grammar features.[8] Nguyen et. al. looked at blogs, telephone conversations, and online forum posts to predict age as a linear regression problem using unigrams, parts of speech, and linguistic inquiry and word count (LIWC) to achieve an R^2 value of 0.551.[13]

Pentel took a different approach from traditional bag-of-words, parts-of-speech, etc. and generated a lot of text-based features as shown in Figure 16:



Coefficient	Feature
1.3639	Words in sentence
0.8399	Characters in word
0.258	Complex words in sentence
-0.2713	Ratio of words with 4 syllables
-0.3894	Commas per sentence
-0.7451	Ratio of words with 1 syllable
-0.762	Ratio of words with 2 syllables

Figure 16: Coefficients of Text-Based Features in Pentel (2015)

Structured as a classification problem between “teenagers” and “adults,” Pentle achieved an F-score of 0.93 using logistic regression.[14]

3.4 Literature Related to Natural Language Processing

Many of the state-of-the-art natural language processing models leverage deep learning, which is beyond the scope of this course. GPT-3 by OpenAI has revolutionized the field of NLP using deep learning.[15]

4. Our Models

Now that we have reviewed some of the relevant literature, we will discuss our model specifications and results for the predictive tasks described in Section 2.

4.1 Common Processing

All of the models mentioned below combine all 10 essays into one. Different combinations of pre-processing steps were tested for impact on model performance, including converting to all lower case, removing punctuation, removing stopwords, stemming words, and using different dictionary sizes; however, conducting all of these pre-processing steps generally improved performance for all tasks.

Features were extracted from the combined essays and models try to predict the target variable that the user provides. We considered multiple models in the sklearn package, such as SVM; however, the simpler models frequently had the best performance as more complicated models tended to overfit. All models used accuracy, BER, or MSE as performance metrics unless otherwise mentioned.

4.2 Predicting Gender

The final model used logistic regression to predict genders from text data. The model takes in the top 1000 unigrams and 1000 bigrams and 1 constant term (2001 features for each user). The goal is to predict genders (i.e. binary variable coded 1 as female and 0 as male) from the text in their essays; namely, it is a classification problem. Therefore, logistic regression would outperform linear regression.

The methods we tried on text mining were bag-of-words and tf-idf models. We tried different ways of feature engineering, including preprocessing such as removing stemming, stopwords, punctuation and combining unigrams and bigrams or leaving them separated. Preprocessing and combining both unigrams and bigrams significantly enhanced the model’s performance such that the accuracy of predicting genders increased from chance level 60% to ~75%.

The proportion of gender classes is relatively balanced (60% male and 40% female); thus, changing the class_weight in the logistic regression model to be balanced or not did not significantly influence the accuracy and BER such that the change was within in 1 %, respectively.

4.3 Predicting Age

For this task, we tried to predict a real-valued output, a user’s age, so we used a linear regressor. The model was optimized by including/excluding various features from the dataset (whether or not the user drinks alcohol/has pets/income bracket/etc.), and trying to process the text in user-written essays in their profile. For classifying text, we tried to search for several keywords to indicate oldness, like ‘old’, ‘thirties’, and their synonyms and antonyms. This model ran into a few issues with training, namely overfitting. The distribution of ages is somewhat Gaussian, and the population of people over 50 using the application was relatively small, so it was hard to extrapolate what features would generalize well to unseen profiles.

Some methods we tried to analyze the text were the bag-of-words and tf-idf models along with variations of unigrams/bigrams and removing stopwords/punctuation. However, the performance of these features were roughly the same as chance. When analyzing the data, we found that many people used words in unpredictable ways, ie. saying that they felt “old” when they were actually under the mean age (several males and females considered themselves to be old at the age of 25, which was well under the average age of

32). So, while text analysis was useful for other tasks, it was not particularly useful for predicting age.

We tried two models for predicting this: the normal linear regressor that was shown in class, as well as a Lasso linear model, also from the sklearn library. The former had an advantage in its simplicity as it didn't overfit as easily. However, since many of these features were one-hot vectors, the Lasso regressor had an advantage as it seeks to set as many weights to be as close to 0 as possible and is more optimized for the features we collected. The MSE of simply predicting the mean age was roughly 89.3535, while simply predicting the median baseline had a MSE of roughly 94.8304. The MSE of the normal linear regressor on training data was ~85.23542, yet on validation sets rose to 94.3714, which was only slightly better than the median baseline. The Lasso regressor was able to beat both of these, as its MSE on validation sets was roughly 86.8812 (after finding an optimal alpha value of ~72, through a pipeline for testing this hyperparameter).

4.4 Predicting Body Type

Originally there are multiple different labels users can select to best describe their body type; however, most of them are redundant or very similar. So, the labels are combined into 'fit,' 'average,' 'thin,' 'fat' and 'N/A.'

The model used in the final product is a neural-net model. The model takes in 1000 unigrams and 1000 bigrams and one constant term and returns the confidence level for each class, where 0 = "fit", 1 = "average", 2 = "fat", 3 = "thin", 4 = "N/A".

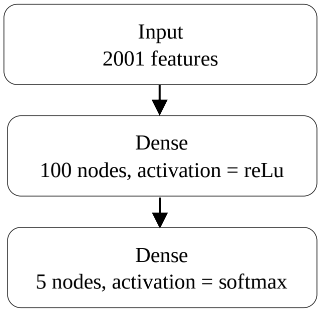


Figure 17: Neural Net Model for Predicting Body Type

The model basically functions as a multi-class classifier with a bit more predictive power due to the hidden nodes in between. The model is also easy to implement with the help of TensorFlow. When the model returns it returns a confidence level of each class, therefore an additional argmax have to be performed for the model's predicted class.

We also tried removing the 5th class, the 'N/A' selection, but it doesn't seem to be any positive benefit to the accuracy. Other variations of the model were tried out, but gave no additional predictive power.

4.5 Predicting whether a user Speaks C++

Predicting whether a person speaks C++ or not is a binary classification with True as the person have selected that they speak C++. The final model used is a very simple logistic regressor, but through some interesting feature tuning, it was able to get some great result. Instead of benchmarking on accuracy, BER is used as a better performance metric due to the huge class imbalance issue (3 to 97); this issue will be discussed more in depth in the conclusion.

All 0th to 9th essays are combined into one single long essay with "\n" separating them, the punctuation is stripped, all words are stemmed, stopwords are removed, and the top 1000 of unigram words' tfidf and the top 1000 bigram words' tfidf are used. Features are then trained in a class balanced logistic regressor, then optimized on different regularization factor. The best performing model yields a 0.29 balanced error rate (BER) and a 0.85 accuracy, while an always false trivial predictor will have a 0.5 BER, but a 0.97 accuracy. Considering some users will know programing but not have known C++, we consider it to be a pretty significant indicator that programmers usually mention that they know programming in one way or another.

Some interesting variations tried are that instead of getting tfidf from all essays, tfidf from only people with C++ marked as selected language sampled from. It performed slightly worse compared to sampling from all essays (which didn't make it into the final model), but it cuts down on prediction time significantly.

5. Results and Conclusions

5.1 Gender

With the logistic regression model described in Section 4.1, we could successfully predict the gender from text such that the accuracy was 75% and BER was 0.26. This result outperformed the trivial baseline that predicts all users as males (accuracy = 60% and BER = 0.5). This finding suggests that female and male use different words on OkCupid. The top words that predict a female include "good people," "good book," and "get lost," while the top words that predict a male include "also pretty," "comput scienc," and "come back."

Our modeling results suggest that we could use text, including the unigram and bigram, together to make accurate predictions (75%) on gender labels. This is consistent with previous findings and hypotheses that different genders have unique language usages. Starting from Lakoff (1975), studies have continued to investigate the relationship between sociolinguistics and gender.[16] Our study formed this question as a classification task. The main take-away is that we could use the bag-of-word model to predict gender.

Note that our findings are just based on a portion of OkCupid users located at around Bay Area from 2012, which may not be representative to general population nor is up-to-date. Moreover, this sample is likely to be biased since they may be influenced by other users in the same gender, and thus use similar wording/description on their profiles. More importantly, we need to acknowledge that gender is a spectrum, reducing it to be binary labels may be insufficient. More and more dating apps have adopted surveys asking about where the users are on the gender spectrum, which could potentially provide a better dataset for future research on sociolinguistics and gender.

5.2 Age

As mentioned before, the MSE of our baselines of blindly predicting the mean age and median age were 89.3535 and 94.8304, respectively. The MSE of the normal linear regressor on training data was ~85.23542, yet did not generalize well enough to beat the baselines. The Lasso regressor was able to beat both of these, as it had MSE of 86.8812 with an alpha value of ~72. From further analysis of our feature vectors, we found that the most significant predictors of how old people were the length of their reviews (younger people tended to have shorter profiles) and drug use (younger people were more likely to use drugs). Intuitively, these results make sense as these are "young people" things.

We tried various feature representations, as described above. The features we ultimately selected were a user's diet (a more strict diet correlating with older people), whether or not a user used drugs (younger people were more likely to), whether or not a user smoked (younger users tended to smoke less), how often a user consumed alcohol (younger users consumed more), how educated a person was (older people were more academically distinguished, through doctorate programs/etc), whether or not a user had pets (younger people tended to have pets more), and body type (younger people tended to classify themselves as average, thin, or fit, while older people tended to describe themselves as 'a little extra'). We tried incorporating other features, but they weren't predictive (as they had very low weights assigned to them by the linear regressor). The predictiveness of a feature was the magnitude of the its corresponding weight scalar.

The proposed model succeeded, as it was able to perform much better than our baseline mean and median predictions.

5.3 Body Type

After feeding our features though the neural-net, we get about 46% accuracy, which is far better than a 20% accuracy of a trivial model. After inspecting the accuracy of each individual class, 'fit' had a good 57% accuracy, but 'average,' 'fat,' and 'N/A' had an accuracy from 20% to 35% range. The worst of all was the 'thin' class; it had an extreme low accuracy of only 4%. When investigating why the 'thin' class only had an accuracy of 4%, we found that the class wasn't very balanced after combining them into the five mentioned. 'Fit' class is about 40% of the data, 'average' class is about 25% of the data, 'fat' being 10%, 'thin' being 15% and the rest being 'N/A.'

With the class balance in mind, the intuition is that a 'fit' person might mention a bit more about the sports that they like to play or being outdoor type. Some users may claim they are 'fit,' but just talk like an average person, while almost all 'thin' users just talk like an 'average' user; consequently, the model seems to just choose to avoid predicting the thin class. People that are curvier are slightly more identifiable than random. A person that doesn't select their body type has about the same accuracy as a trivial predictor of 20%. These results pretty much agree with our intuition.

5.4 Speaks C++

In the OkCupid data, about 3% of users mark that they speak C++ or Lisp as one of their selected languages in different level of fluency. While it seems like a funny joke, we saw a great opportunity to predict if the user knows how to program just from their self-description.

It turns out that people who code write in a specific manner and it's rather revealing in their essays. However, there are some issues with this feature in a user's profile, as the OkCupid's dropdown selection bar only includes 2 programming languages of C++ and Lisp. This is a limited selection of languages; users might know Python, Java, Golang or other programming languages. So, some user might know programming languages other than C++ or Lisp and mentions it in their essays, but won't have C++ nor Lisp marked. The second issue is that only 3% of the data are a positive case; combining with the first issue, accuracy is no longer a good metric to be benchmarked on, so we mainly optimize balanced error rate (BER) instead of accuracy as a trivial predictor of always false achieves 97% accuracy easily.

Stemming helped aggregate similar word stems that proved to be very predictive. The top five word stems with the highest weights are "comput", "engin", "softwar", "techolog", "program", which are pretty common terms within the software engineering field but rather uncommon outside, similar to other words like "startup," "nerd" and "tech" a bit lower on the list. The strongest bigram word is "softwar, engin" could refer to an occupation of software engineering as their job title somewhere in their essay. So, the final conclusion is that a personal essay can be used to distinguish a person who knows about programming with a good degree of accuracy (low balance error rate).

6. Conclusion & Future Directions

In summary, the performance of our models that apply text mining outperformed the trivial baselines in all four predictive tasks, with more success seen in predicting gender and whether or not a person selects programming languages such as C++ than body type or age. Our finding demonstrates the predictive power of text mining on a variety of demographic variables of OkCupid users.

For future analysis, we are interested in improving our text prediction model. Leveraging techniques from related literature, we consider using additional text features (e.g. parts of speech, log-odds-ratio), topic modeling (e.g. non-negative matrix factorization and/or LDA), and vector space models (e.g. word2vec). We would also explore using deep learning techniques and demographic variables outside of text to improve our models.

We could also take a more exploratory approach to study more in depth about the relationship between our target variable (ie. gender, age, body type, language) and word use and the interactions of the target variable and other variables on word use. More interestingly, future research could extend the findings to broader contexts such as whether the text in tweets, Facebook posts, newsletter, or even diaries.

References

[1] Kim, Albert & Escobedo-Land, Adriana. (2015). OkCupid Data for Introductory Statistics and Data Science Courses. Journal of Statistics Education. 23. 10.1080/10691898.2015.11889737.

[2] Rodríguez-García, M. Á., Valencia-García, R., Colomo-Palacios, R., & Gómez-Berbís, J. M. (2019). BlindDate recommender: A context-aware ontology-based dating recommendation platform. Journal of Information Science, 45(5), 573-591.

[3] Pandit, A. (2020). Strategic Self-Representation by Heterosexual Male Users on American Online Dating Platforms: Converging Towards or Diverging From Emergent Norms?.

[4] Hutson, J. A., Taft, J. G., Barocas, S., & Levy, K. (2018). Debiasing desire: Addressing bias & discrimination on intimate platforms. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-18.

[5] Kirkegaard, E. O., & Bjerrekær, J. D. (2016). The OKCupid dataset: A very large public dataset of dating site users. Open Differential Psychology, 46, 1-10.

[6] Shishido, J., Narasimhan, J., & Haller, M. (2016). Tell Me Something I Don't Know: Analyzing OkCupid Profiles.

[7] Rudder, C. (2014). Dataclysm: Love, Sex, Race, and Identity--What Our Online Lives Tell Us about Our Offline Selves. Crown.

[8] Abdallah, E. E., Alzghoul, J. R., & Alzghool, M. (2020). Age and Gender prediction in Open Domain Text. Procedia Computer Science, 170, 563-570.

[9] Bartle, A., & Zheng, J. (2015). Gender classification with deep learning. In Technical report. The Stanford NLP Group..

[10] Ciccone, G., Sultan, A., Laporte, L., Egyed-Zsigmond, E., Alhamzeh, A., & Granitzer, M. (2018, September). Stacked gender prediction from tweet texts and images notebook for pan at CLEF 2018

[11] John D. Burger, John Henderson, George Kim, Guido Zarrella. Discriminating Gender on Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011

[12] Arjun Mukherjee, Bing Liu. Improving Gender Classification of Blog Authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[13] Nguyen, D., Smith, N. A., & Rose, C. (2011, June). Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities (pp. 115-123)

[14] Pentel, A. (2015). Automatic Age Detection Using Text Readability Features. In EDM (Workshops).

[15] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

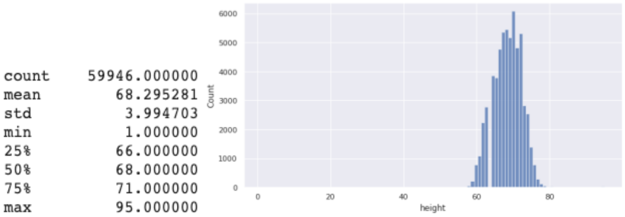
[16] Lakoff, Robin. 1975, 2004. Language and Woman's Place: Text and Commentaries. New York, Oxford University Press.

Appendix A. Dataset Exploration Continued

In order to keep the most pertinent information in the main body of the paper, we explored only the variables relevant to our predictive tasks above. Here, we explore the rest of the variables in the dataset.

A.1 Analysis of Variables Continued

A.1.1 Height. As shown in Figure 17, height is relatively normally distributed with a mean of 68 inches.



```
count    59946.000000
mean        68.295281
std          3.994703
min          1.000000
25%         66.000000
50%         68.000000
75%         71.000000
max         95.000000
```

Figure 17: Distribution of Height

Users who report excessively small heights (ie. 1 inch) or large heights (ie. 95 inches) seem to be legitimate users who do not want to disclose their actual height.

A.1.2 Status. A shown in Figure 18, users are predominantly "Single." Common sense prescribes that users who are not single would not be on a dating app.

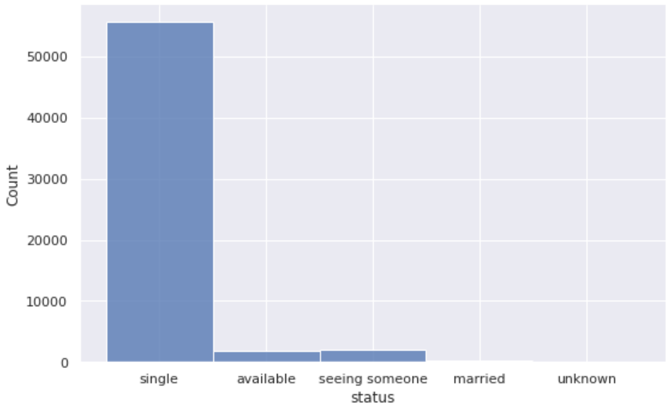


Figure 18: Distribution of Status

A random survey of users that are “Married” or “Seeing Someone” generally either acknowledge that they are not interested in using the dating app anymore or do not acknowledge their committed relationship at all. There was no apparent commonalities between users who identified as “Unknown.”

A.1.3 Orientation. As shown in Figure 19, most users (86%) identify as “Straight.” About half as many users identify as “Bisexual” than “Gay.”

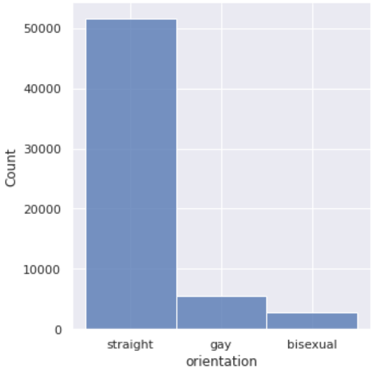
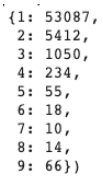


Figure 19: Distribution of Orientation

A.1.4 Ethnicity. Users can identify as multiple ethnicities. As shown in Figure 20, most users identified as one ethnicity, with a steep drop-off at four ethnicities.



```
{1: 53087,
 2: 5412,
 3: 1050,
 4: 234,
 5: 55,
 6: 18,
 7: 10,
 8: 14,
 9: 66})
```

Figure 20: Distribution of Multiple Ethnicities

As shown in Figure 21, “White” was the predominant ethnicity, followed by “Asian” and “Hispanic.” This relatively reflects the demographics of the San Francisco Bay Area.

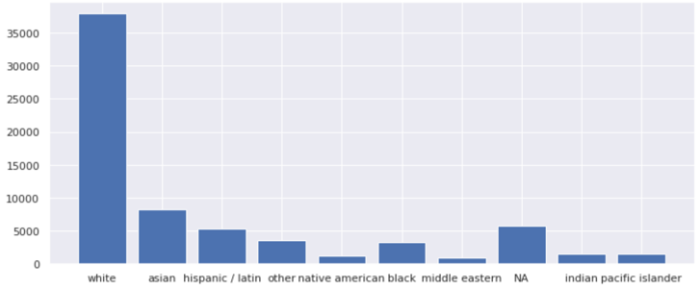


Figure 21: Distribution of Ethnicities

A.1.5 Diet. As shown in Figure 22, most users eat “Anything” followed by “Vegetarian.” Few users are “Vegan,” “Halal,” or “Kosher.”

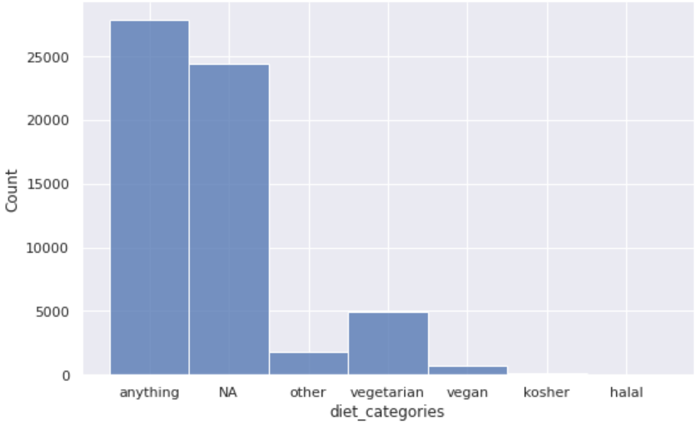


Figure 22: Distribution of Diet

A.1.6 Location. Users are predominantly located in San Francisco (51.8%), with Oakland, Berkeley, San Mateo, and Palo Alto as other top cities. Berkeley (UC Berkeley) and Palo Alto (Stanford) are home to the biggest universities in the region, while San Francisco is full of young professionals. Consequently, it makes sense that these cities would have the highest populations using dating apps.

A.1.7 Education. While the dataset captures more granular detail on if the user is “working on,” has “graduated from,” or “dropped out of” different university programs, we summarized the data in Figure 23 below.

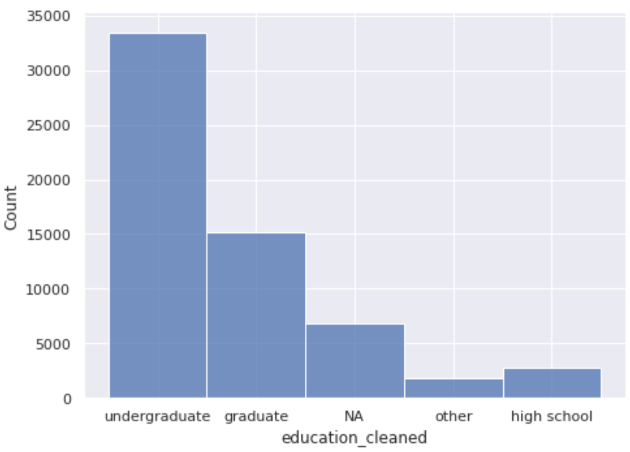


Figure 23: Distribution of Education

Most users' highest degree is at the undergraduate level. Less than half are at the graduate level and very few have only a high school education. This corroborates with the fact that the Bay Area is known to have the most educated populations in the world.

A.1.8 Job. As shown in Figure 24, "Computer," "Science," and "Student" are top jobs, resonating with the highly educated, technology-focused industry in the Bay Area.

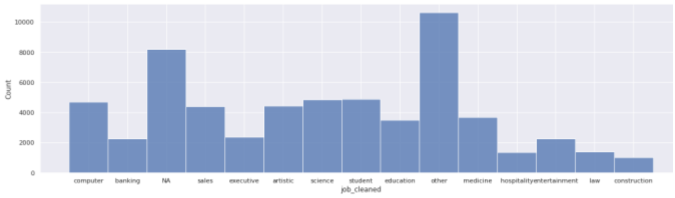
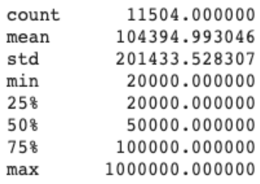


Figure 24: Distribution of Jobs

"Artistic," "Medicine," and "Education" are also highly represented.

A.1.9 Income. As shown in Figure 25, most users earn in the $20K - $100K range. This seems low given the high cost of living in the Bay Area, but may be representative of the high number of students in the dataset.



```
count       11504.000000
mean       104394.993046
std        201433.528307
min         20000.000000
25%         20000.000000
50%         50000.000000
75%        100000.000000
max       1000000.000000
```

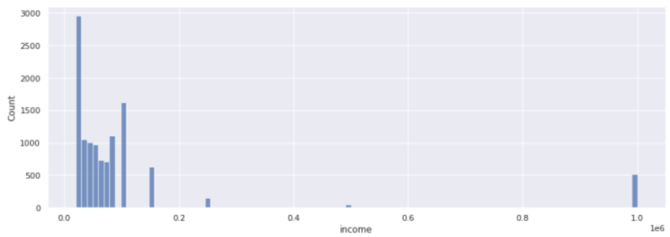


Figure 25: Distribution of Income

521 users reported incomes of $1M. While it seems as though many have duplicitously reported their incomes, many works in "Medicine," "Finance," and technology and may be commanding such high incomes. Surprisingly, a disproportionate amount work in "Artistic" domains – are pop stars and celebrities looking for love in the Bay Area?

A.1.10 Offspring. The dataset provides copious options to denote if a user 1) has or does not have kids and 2) wants, might want, or doesn't want kids. We simplify our analysis to note that most users (59.3%) did not respond to the question, 32.4% do not have kids, and 8.2% have at least one kid.

A.1.11 Pets. The dataset provides copious options to denote if a user 1) has a cat or dog, 2) likes cats, and 3) likes dogs. We simplify our analysis to note that most users (50.2%) do not have a pet, 16.5% have a pet, and the rest (33.2%) did not respond to the question.

A.1.12 Religion. As shown in Figure 26, most users did not respond to the question. Of those that responded, "Agnosticism," "Atheism," "Christianity," and "Catholicism" are the top religious beliefs. This corroborates with the Bay Area's a-religious stance and the dataset's predominantly "White" and "Hispanic" population.

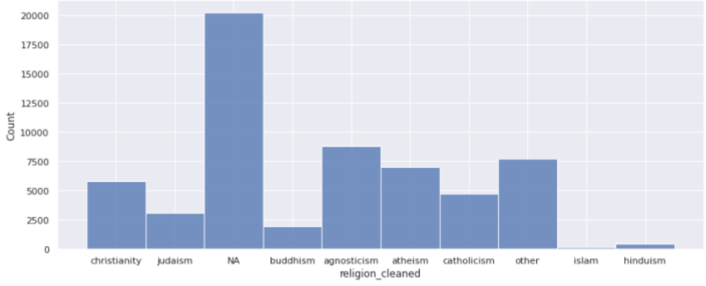


Figure 26: Distribution of Religion

A.1.13 Sign. As shown in Figure 27, astrological signs are roughly evenly distributed; this makes sense as astrological signs are demarcated based upon one’s birthday and birthdays are evenly distributed throughout the year.

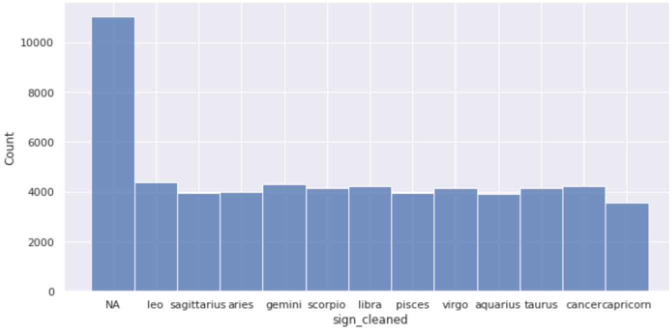


Figure 27: Distribution of Sign

32.2% of users express that astrological signs are “fun to think about,” while 30.0% express that signs “don’t matter.”

A.1.14 Drinks. As shown in Figure 28, most users (70.0%) report drinking socially. Only 15.4% of users report drinking “rarely” or “not at all.”

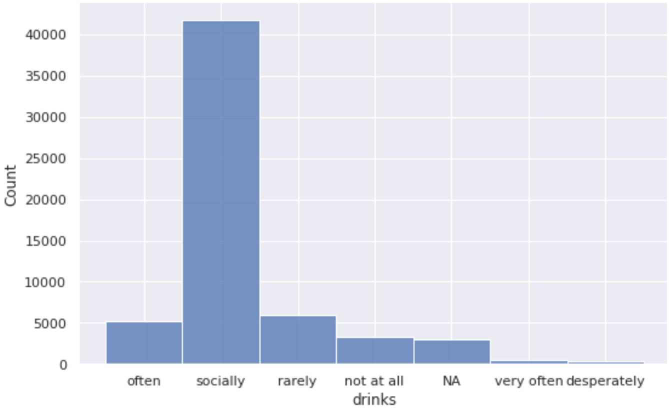


Figure 28: Distribution of Drinking Habits

322 users curiously report drinking “desperately.” Some of these users have alluded to their drinking habits in their essays.

A.1.15 Drugs. As shown in Figure 29, most users (63.0%) report not using drugs. It seems some users included marijuana usage in “drugs,” while others included it in the category “smokes.” That said, marijuana usage was not as big in California in 2012 as it is now.

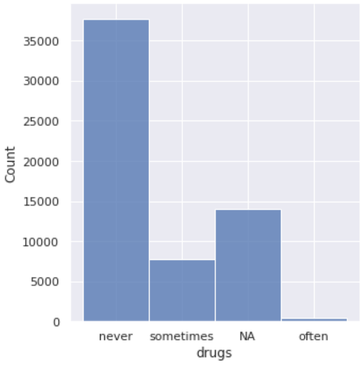


Figure 29: Distribution of Drug Usage

410 users report using drugs “often.” Some of these users have alluded to their drug habits in their essays (e.g. one user admitted to having bad ADD).

A.1.16 Smokes. As shown in Figure 30, most users (73.2%) do not smoke. Some users report smoking only “when drinking” or are “trying to quit.”

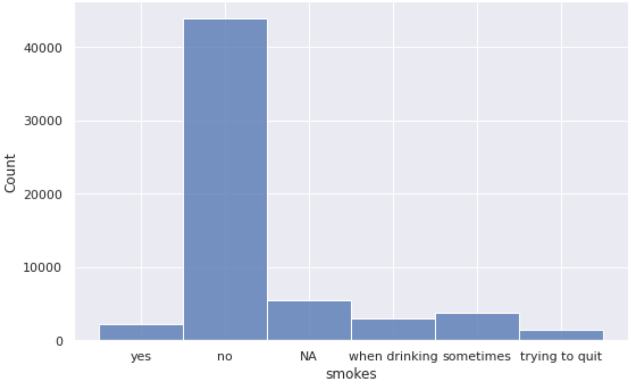
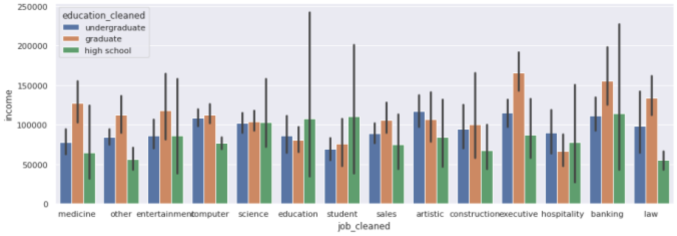


Figure 30: Distribution of Smoking Behaviors

A.2 Variable Interactions Continued

A.2.1 Income & Education. Across most job functions, users with graduate degrees earn more than those with undergraduate degrees, and both earn more than users with high school degrees.



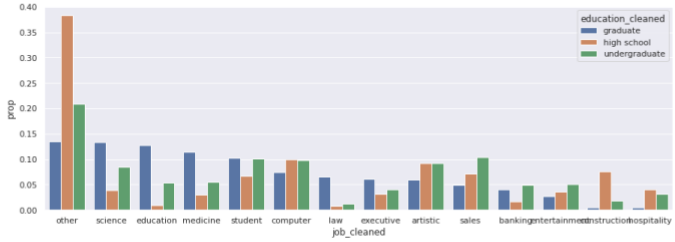


Figure 31: Distribution of Income & Jobs across Education

Users with higher education are more likely to have jobs in "Education," "Medicine," and "Executive." "Construction," "Hospitality," and "Artistic" fields are less likely to have users with traditional education.

A.2.2 San Francisco. Do the stereotypes of high-income earners in San Francisco hold true? Figure X shows that users in San Francisco earn higher incomes across all job functions than their non-SF counterparts.

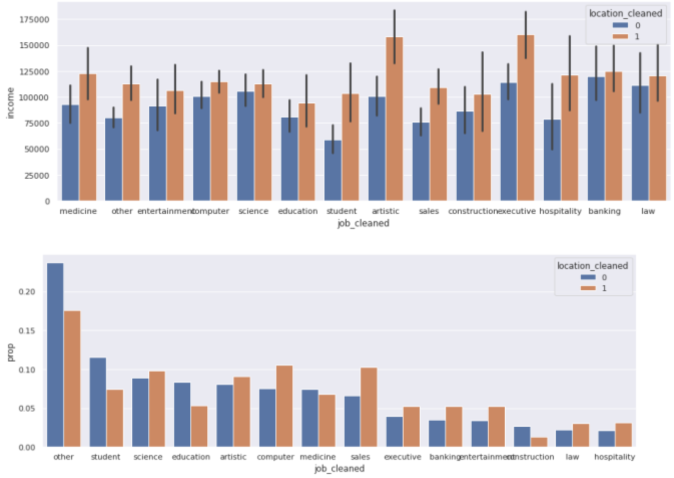


Figure 32: Distribution of Income & Jobs across SF vs. Non-SF Residents

San Francisco residents are more likely to hold jobs in "Science," "Computer," "Sales," "Executive," "Banking," and "Entertainment" than the rest of the Bay Area.

A.2.3 Drinking & Drugs. Users who identify as Christian, Buddhist, or Muslim are less likely to drink than their non-religious counterparts. Christians and Catholics are much less likely to use drugs than their non-religious counterparts.

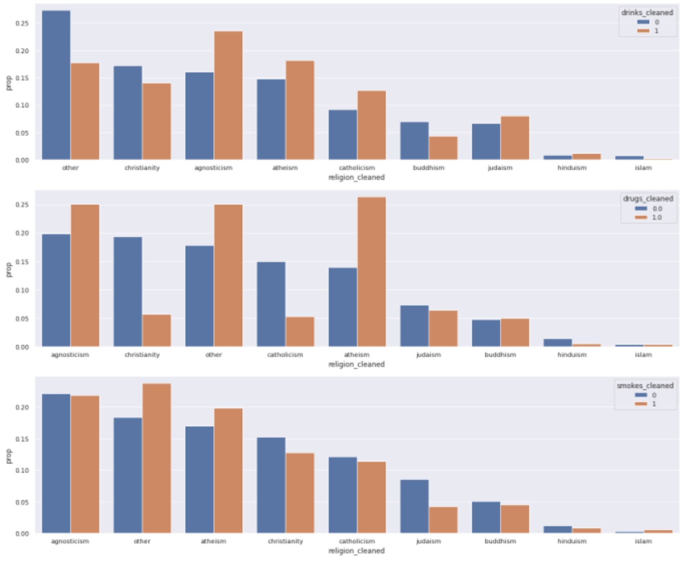


Figure 33: Distribution of Drink & Drug Usage across Religion

Users with only a high school education are less likely to drink, presumably because they are under 21 years of age, the legal drinking age in the US. Graduate students are much less likely to do drugs or smoke than their less educated counterparts.

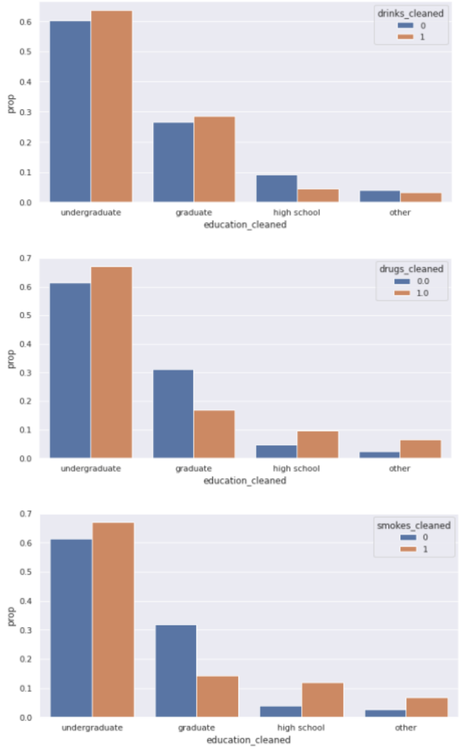


Figure 34: Distribution of Drink & Drug Usage across Education