# Final Project
PSTAT 131

Julian Marks
Timothy Nguyen

December 14, 2017

## 1 Abstract

The data set utilized contains user profile information from 59,946 San Francisco OKCupid users from June 2012. The data includes general identification information, lifestyle variables, and text responses to 10 essay questions. We utilize the text responses of the users and create a model to predict the education level of a particular user based on their choice of language. We utilize support vector machines, random forests and logistic regression, in addition to creating a corpus to analyze the term frequency-inverse document frequency of the words used in the user essays. The mined text can then be used to identify those words that carry the most predictive power.

## 2 Introduction

With the online populace rapidly growing in prominence and importance, electronic companies need innovations to keep their product on the foreground of expansion in order to keep up with the consequential effect of growing integration of technology and the internet in day to day life. OKCupid has not only created a product that brings dating to the electronic realm, a place where the younger generation find themselves more comfortable and accessible, but a place that stockpiles information on these people. This provides an avenue for data miners, including themselves, to effectively study the habits of many people using statistical methods. For example, the team behind OKCupid was able to find a trend regarding the sort of messages and response a user receives based on their race; not to mention the intensive algorithmic construction that goes behind the matchmaking procedures for the primary site function. (Del Rey 2012)

The importance of data mining in a rapidly growing technocentric society is plainly visible. As machine learning algorithm development increases its potential, the ability for programmers to recognize and analyze attributes via programs will excel and prove vital when attempting to assess the behavior of others. This is especially true in the process of text mining where the most primary form of communication can be assessed, and with

extensive training computers can be taught to recognize and realize as much of what is communicated through language as humans.

The data used is gathered entirely from OKCupid users in San Francisco, and it is based on personal response and evaluation. As such there could be bias when creating a profile, however this is still an effective data set; because of the nature of the text response questions and the desire to match two people for such intense cooperation, the text responses should act as an accurate avatar of that individual. Focus on users from San Francisco may not have as good predictive accuracy in other areas but within the bounds of San Francisco modelling and predicting this information will suffice. Using the data provided in the essay questions, we will address whether word choice is a significant factor in determining the level of education of any particular user. The words in all of the essays for each person were bagged, and their education level was separated into a number of bins due to the large variance in responses from the education variable.

# 3 Data and Methods

The variables of importance in this data set and for our question lie within the text responses of the users. Since we are processing natural language these are the only variables we need to consider. By considering the importance of particular words, disregarding syntax, creating a bag of words for each user made up of their essay questions made it easy to create a corpus. In addition to bagging the responses, the education levels were binned into 6 categories: high school, pursuing an undergraduate degree, received an undergraduate degree, pursuing a graduate degree, received a graduate degree, and space camp. High school is for all those who are pursuing a two-year degree or has lesser education, and pursuing undergrad degree encompasses those who have graduated two-year college and those who are pursuing a bachelor degree. Many users listed an option for space camp, but with no information on what this means there is no way of using it in our analysis. Hence we decided to remove all users with education listed as empty or with space camp specification, since there is no way to improve the model with observations that don't have a response variable.

The primary task is creating the corpus of all text responses from each individual's documents. This corpus contains information on the importance of each word, its text frequency-inverse document frequency. This measures the frequency of a word between documents, and within documents, and each word gets ranked based on its frequency. Words that appear in many documents are not as important as they are likely to be common and necessary language, but words that appear many times within a single document could be vital in distinguishing an aspect of that document if that word is unique to that document. Hence those words with high tf-idf are significant in reporting information about the document they are found in. This is taken with a grain of salt, since those "important" words could be made up words or proper names. To combat this sort of analysis, we decided to make sure that only words that appeared in multiple documents were included in the corpus. This makes sure that most of what is included in the corpus is informative, natural language.

During our preliminary model building it was found that there was a huge imbalance of data. In our response variable, after binning, those who received an undergraduate degree more than double each other category. Thus when fitting models with all the data, the model is trained to receive more samples from a particular class and as result to predict a more likely result to fall into that class. This can lead to high predictive training accuracy, and possibly test accuracy when determining the education level of any new user, but for our analysis we wish only to judge their education based on their language and this class imbalance will lead to inaccurate modelling. To combat this imbalance, we put undersampling into effect. This technique requires reducing the number of over represented classes so that all classes have equal representation when we create models.

We decided to create a number of different models to run our tests. These included a support vector machine, logistic regression and a random forest of decision trees. These are all good choices for modelling the data set, and each comes with its own perks and downfalls. The support vector machine is good due to the large dataset and high frequency of observations. In this manner, the model doesn't need to utilize all of the information to create a model but only those words that reveal the most information. Logistic regression is an excellent classification tool, as a high density of words to education level will give the model a strong predictive base. Unfortunately due to the strong correlation of language between all users and the large number of classifications (5), logistic regression has trouble making solid, distinguishing predictions. Another modelling technique utilized is the random forest, which yielded the best results. With random forests, subsets of important words can be identified, or might show up more often throughout the trees making classification effective in those trees.

Using the LiblineaR package, we test 8 different modelling types of support vector machines and logistic regression schemes. They all vary in regularization scheme and loss functions, and they all have an associated cost which determines the trade-off between regularization and correct classification. All 8 types of models are compared through 3 different costs to find the model with the highest predictive accuracy.

# 4 Results

# 5 Conclusion

Our best predictive model was able to come up with a correct classification rate of approximately 40%. Based on random guess, which would have correct results only 20% of the time. This model doubles the accuracy of random guesses, and while 40% may not be immensely accurate, it is still far better than the average guess. Some difficulty is the incredible overlap that exists within language, especially English. The most common words used to communicate is known by all who know the basics of the language, and often grammar is a strong indication of knowledge of a language. In the future, a large-scale project would include the analysis of grammar usage, though this would add an immense layer of difficulty in the programming aspect. Unfortunately this could be extremely crucial to predicting levels of education.

# 6 References

2017. "Working with Text Data", *Scikit-learn*.
**URL:** `http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html`

Abramovich, Giselle. 2012. "How OKCupid Built a Data-First Brand", *DigiDay*.
**URL:** `https://digiday.com/marketing/how-okcupid-built-its-brand-on-data/`

"Getting Started with quanteda", *CRAN*.
**URL:** `https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html`

Helleputte, Thibault and Gramme, Pierre. 2013. "Linear Predictive Models Based on the 'LIBLINEAR' C/C++ Library", *CRAN*
**URL:** `https://cran.r-project.org/web/packages/LiblineaR/LiblineaR.pdf`