Teradata University Challenge: Executive Summary

By Roy Hu

Abstract

This paper tries to estimate the effectiveness of Hired Heroes USA's volunteer program on raising the labor market outcomes for its participants. I use Rubin's causal model for potential outcomes to measure the causal relationship between participation in the program and hiring outcomes. To address difficulties with the data, I supplement the analysis with insights from machine learning, such as K-means clustering algorithm to extract as much usable signal from noisy text data, as well as the XGBoost (eXtreme Gradient Boosting) algorithm to generate propensity scores for all 64,000 observations in our sparse dataset. By weighting each observation by its propensity, I am able to run a logistic regression of 60 "rich variables" on the observed hiring outcome, and estimate that the causal effect of participation in Hired Heroes' volunteer program raises one's chances of getting hired by 20.4%.

Motivation and Methodology

I investigate whether or not there is a positive effect in the hiring outcomes of veterans when they make use of the volunteer services provided by Hired Heroes USA, a non-profit whose mission is to connect veterans with meaningful career opportunities. One of Hired Heroes' main services is its volunteer program, in which trained volunteers provides a range of job-preparation services for its veterans, including conducting mock interviews, resume compilation, translating military skills into civilian work experience. Hired Heroes wanted to know if their volunteer program was effective in raising hiring outcomes for its veterans who participated.

If participation in the volunteer program were truly exogenous, then we could directly estimate its effect on hiring outcomes; however, it is certainly possible that job-seekers who use the volunteer program have some extra intrinsic and unmeasured motivation, and would have done better than their peers even if the volunteer program hadn't existed. This is sometimes called the "Fundamental Problem of Causal Inference" (Holland 1986) because one can never observe both potential outcomes for the same individual at the same time. Thus, I viewed this as a Program Evaluation problem, and used Rubin's framework to estimate the causal effect of the treatment (participation in the Volunteer program) on hiring outcomes. The major and untestable assumption behind Rubin's causal model is that once we control for a set of covariates, there are no other variables which affect both the treatment and outcome. While there are always arguments for additional confounding variables, this assumption is more plausible if the set of covariates is considered to be "rich." One development in the causal inference literature has been to estimate the propensity score, which measures the likelihood of receiving treatment given a vector of covariates, and weighing the regression of outcomes (hired or not) on the treatment (participation) by the estimated propensity score.

Data Exploration and Challenges

Some data exploration is helpful to understand the problem from a high level. By far, most (78%) participants used the service for mock interviews, while some (9%) used it for resume reviews, and others (13%) used it for other career counselling services. When the data is split by education level, there is a clear monotonic trend upwards in participation: while only 1.3% of all veterans with only a high school degree used the volunteer program, 7.8% of veterans with doctorate degrees did. This reinforces the need to evaluate participation in the volunteer program as a causal inference problem. Like participation in the volunteer program, education is also one of the often problematic variables in the causal inference literature. It can be difficult to estimate the effect of education on, for instance, wages because the data does not capture any intrinsic motivation--the same motivation that makes someone seek out extra volunteer services. Moreover, 55% of men who used volunteer services were hired, while only 32% of men who did not use volunteer services were hired.

Unfortunately, the data for this project suffers from the major problems of sparsity, in which there are many missing values for each variable. Other problematic elements of the data, such as a vector that contained some credit card information, speaks to the need for Hired Hero to improve its data collection and management operations. Most covariates have large sections of missing values, and many of them are almost entirely empty. The danger with running a logit regression on these is that will drop any observation with a missing value, even if the observation is nearly complete.

Data Wrangling, Feature Engineering, and Variable Selection

The majority of the project was spent on this stage. The first step that I took was to try dropped features that were nearly or completely NA. I then removed features in which there was near zero variance, because if a column consisted of just single value, e.g. all ‘Yes’, it would provide no information gain. This greatly reduced the dimensionality of the data set from 653 to 96 features. I also removed a couple covariates that were perfectly correlated. I then downloaded some government statistics for zip codes and Metropolitan Statistical Area and merged it with the data by a column of zipcodes. In particular, we were interested in having an MSA variable to control for the effects of living in a big MSA like New York or LA. An MSA feature is hopefully an improvement upon only having zip codes, since there are so many different zip codes that it might be noisy.

Other features such as ones for salary, education, and military rank, were transformed or merged into a single feature to the best of my ability. For instance, there were several variables for salary, which I coded into a single feature in hourly wages. Education was particularly difficult because some features were expressed in numeric years, while others were in text describing degrees. I created an education variable based on these by searching using regular expressions. There was also a feature in which veterans described their areas of expertise in words and phrases. Text data can be difficult to work with, but I wanted to extract some signal from text data, so I cleaned the text to remove punctuation, numbers, and common words. I then used the K-means algorithm to cluster "Areas of Expertise" into 6 different clusters. K-means then assigns a number of centroids and then iteratively reassigns each centroid each time around to get drag them closer to the mean of the clusters. I selected 6 based on the "elbow rule," indicating the optimal level of clusters versus sum of squared errors from each data point to the centroid. "Areas of Expertise" is a good feature to include in the model because there's a decent amount of information gain when predicting probabilities.

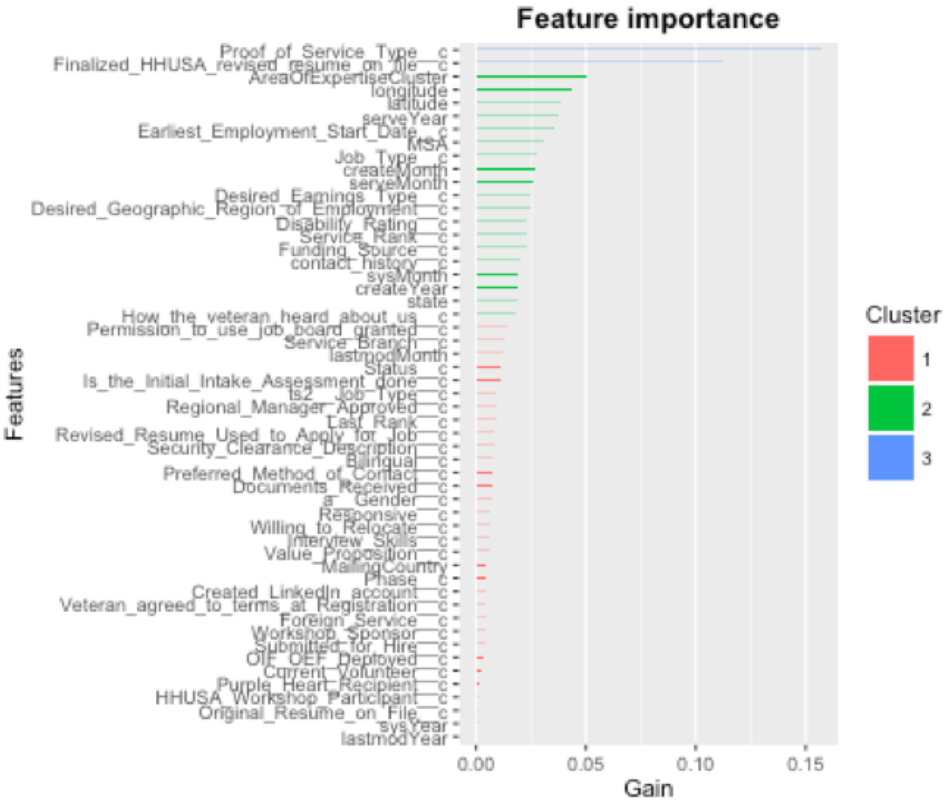
Finally, I wanted to select variables that kept in accordance with econometric sense and domain knowledge, keeping variables that show up in common econometric models. As a matter of judgment call, I removed some variables that I felt were irrelevant, such as the number of guests a veteran brought to a banquet or picnic. I also removed variables that did not provide much marginal contribution to the model's predictive power.

Propensity Scoring

In classic causal inference, one would estimate the propensity of treatment using a logistic regression. However, as mentioned described, it's difficult to generate propensity scores using a traditional logit regression because it drops *any* observation with missing data, even if the observation consists of mostly complete cases. In some model specifications I tried with a logit model, the regression wouldn't even run because it dropped every single observation. First I tried to impute the median value, but for a number of reasons this did not get us good results. There is some developing literature on "multiple imputation" as a better way than median or "knn" imputation, but I did not decide to pursue this path.

To get around the missing data problem, we decided to use the XGBoost algorithm, which has lately been extremely popular in Kaggle data science competitions: almost all recent winners incorporate XGBoost in ensembles, many of whom solely use XGBoost. XGBoost is a boosting algorithm that sequentially combines a set of weak learners into a strong learner. It optimizes on a training loss function (l used log-loss) combined with careful regularization to help avoid overfitting The algorithm is particularly useful here because XGBoost internally handles the sparse data by initially splitting in the default direction "yes," while gradually learning which direction to split based on information gain.

With a set of 56 "rich" covariates to generate a propensity score for each observation, I used cross-validation to find the optimal number of boosting sequences. The point of this cross-validation is that by splitting the data into k-folds, it can try to find the optimal XGB iterations (minimizing log-loss), and serves as a check against "over-fitting." Using this cross-validated model, I was ultimately able to generate probabilities for the full set of 64,000 observations without needing to drop any. Below is a generated output of the most important features according the XGBoost:



One thing I noticed is that the "Areas of Expertise" cluster was the third most important feature for XGBoost to split on, which indicates that the previous k-means algorithm was effective.

Results

XGBoost generated propensity scores for each observation regressed on the indicator variable "Volunteer Services." Now that I had propensity scores for every observation in the data set, I weighted the 64,000 observations by propensity score and was then able to run a logit model on the outcome variable "Hired." According to our model, I found a positive effect on hiring, which was significant at the 1% level. Since the coefficients of a logistic regression are not always straightforward to interpret, I was able to convert this result to a marginal effect, which demonstrated that in the volunteer program has a causal effect of raising one's chances of getting hired by 20.4%.