Project Report: Predicting the Churn Rate Amongst Welfare Recipients

Meshal Alkhowaiter

12/14/2019

Contents

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http:

//rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Introduction:

What is the aim of the project?

The purpose of this project is to utilize machine learning models we have covered in this class, along with a linear model: Logisitc Regression, in a classification context. Specifically, I aim to predict a binary-outcome, the Churn Rate, at the individual level for observsations that Leave or Stay in a welfare prgoram.

The main objectives of this project are:

1- Comparing the performance of various non-parametric machine learning models such as SVM, RF, and KNN, in predicting our outcome of interest: Churn Rate amongst welfare recipients. 2- Comparing the performance of the top machine learning model, with a parametric model: Logistic Regression. 3- Discussing the results from a policy-maker's perspective in the context of Saudi Arabia's labor market. Specifically, I am primarily interested in developing a model that is capable of correctly classifying those who have left welfare, before their 5-year wage subsidy period is completed. This is because I believe that gaining an understanding of the characterisitcs of such individuals can offer insight to policymakers on future welfare programs' design.

Project Roadmap:

1- I first began with Layer 0: split the data; however, deciding how to split my data (i.e. randomly or by date, and if so by which date) turned out to be a challenging part of the project.

2- Exploring my data and analyzing relationships; whereby I looked at the distribution of wages, age, gender, and educational qualification across different schools and regions in my main dataset: Teachers' Subsidy Program. In this step, I also decided how to create my outcome variable: left_welfare, and the underlying assumptions behind my two binary outcomes. Additionally, I explored through visualizations some common hypotheses in Saudi Arabia's labor market context. For instance, whether men on average, are more mobile in the labor market than females, and thus are likely to receive higher wages than females with the same job titles and educational qualifications.

3- Data wrangling, where I created new variables based on existing ones, experimented with different functional forms of my variables (i.e. whether to Log, Square, or Cube a given continuous variable).

4- Running both my parametric and non-parametric models on my training datasets. Assessing the performance across linear and non-linear models, as well as the performance within my machine learning non-parametric models. Analyzing the results and discussing the limitations of each model.

5- Running my top performing model on my test_data, and discussing the findings and how the model performed with data it has not seen before.

Problem Statement and Background:

High-level statement:

Using a machine learning models along with a Logistic Regression model, to predict Churn rate or the probability that a given welfare recipient will drop out of a welfare program, before fully exhausting their wage subsidy period. ## Background: The Teachers' Wage Subsidy schema requires subsidized teachers to find a job at a private school that is willing to hire them and that has agreed to participate in the program. The program offers a 50% wage subsidy that is capped at 2500 Riyals per month, and runs for five years. Additionally, the program obliges participating private schools to pay a minimum of 5600 Riyals per month, but schools have the option to pay their subsidized teachers more if they decide to. ## Literature Review: There is an extensive body of literature on the utilization of non-parametric machine learning models in predicting customer attrition or churn rate, and this application of ML has been gaining popularity since the mid-00s. The Xie et al. 2009 paper uses various Random Forest techniques such as improved balanced random forests (IBRF) and decision-trees (DT) to predict churn rate amongst customers in a Chinese bank. Specifically the authors develop a (IBRF) model that predicts with 93% accuracy which customers will leave the bank and close their account. The (DT) model performed poorly compared to (IBRF) with a 62% accuracy. The Benlen et al. 2014 paper builds on prior literature and uses two SVM model techniques, Logistic Regression, to predict churn rate amongst customers in a Chinese bank. The authors found that Logistic Regression performed better with a 64% accuracy, than an SVM Linear model, which had a 57% accuracy. A growing, albeit still limited body of literature exists on applying machine learning models in the context of government and non-profit programs. The Ozekes et.al 2014 paper is one of the first to apply ML techniques such as KNN and RF to predict dropout rate, amongst students in an online education system in Turkey. The authors found that KNN outperformed RF in accuracy; however, the KNN model also had a high Sensitivity or True Positives rate, but extremely low Specificity. The authors argue that this limitation is due to most people actually droping out of the program. ## Data: ### Where does the data come from? I am using administrative individual-level data that I have acquired for my thesis, from the Ministry of Labor and Social Development in Saudi Arabia. The dataset is for the Teachers' Wage Subsidy program, which is a program that was introduced in 2012 to absorb and employ the large number of unemployed teachers at the time. Additionally, I am using the national standardized exam scores at the school-level. ### What is the unit of observation? Initially, I intended to use schools as my unit of analysis; however, for resasons I will explain below, each observation in my analysis is an individual. The rationale behind using individuals instead of schools, as my unit of analysis, is that due to different yet correct Arabic writing styles, the Ministry of Labor and the Ministry of Education had both adopted different ways of writing the same school_name, thus resulting in zero matching based on school_name between the two datasets. This matter is further illustrated in latter sections. ### What are the variables of interest? The main dataset I am using came in a raw administrative-data format, so crucial variables of interest that are necessary for my analsyis were not included, so I had to create them as explained in the upcoming section. However, my variables of interest encompass ecoonomic measures such as monthly_wage and job title, education measures such as educational attainment and college major (i.e. Bachelor's degree in Physics, etc), job title, geographic measures on city and region, demographic information such as gender and age. ### What steps did you take to wrangle the data? This part includes creating key variables for my analysis that did not originally exist in my dataset. These include such as age_at_enrollment, my Y variable: left_welfare, a categorical variable for monthly_wage, experimenting with the functional form of monthly_wage (i.e. standardized and whether the variable is logged, sqaured, etc), categorical variable for 13 regions and 79 cities, dummy variablem for time (i.e. how many years has the person been in the program). I also wanted to check the number of new subsidized teachers (i.e. beneficiary_name variable) per year, for the five-year period of the program. Specifically, the first year starts on 2012-09-01, and ends on 2013-08-31, the second year starts on 2013-09-01, and ends on 2014-08-31, and so on. This particular where I created a year variable, along with a year dummy variable was extremely challenging. ### Definition of Project Success: I would like to develop Logistic Regression model that predicts with reasonable accuracy, an individual's likelihood to leave the program before the five-year wage subsidy period is exhausted. The policy relevance of this is to formulate an understanding of what factors and characteristics cause a welfare recipient to stay or drop out of the program, in Saudi Arabia. For instance, one insightful finding may be that a school's geographic location, is a better predictor of job retention for a subsidized teacher, than factors such as wage and gender, which policy analysts may perceive as are more relevant in Saudi Arabia's context.

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   left_welfare = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   left_welfare = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
## NULL
```

Analysis:

Describe the methods/tools you explored in your project:

I applied parametric and non-parametric models to predict which individuals will leave the program based on roughly 22 individual level variables or characteristics such as the level of education, region, city, date of birth, gender, wage, age_at_enrollment. The results of the models I ran are as follows:

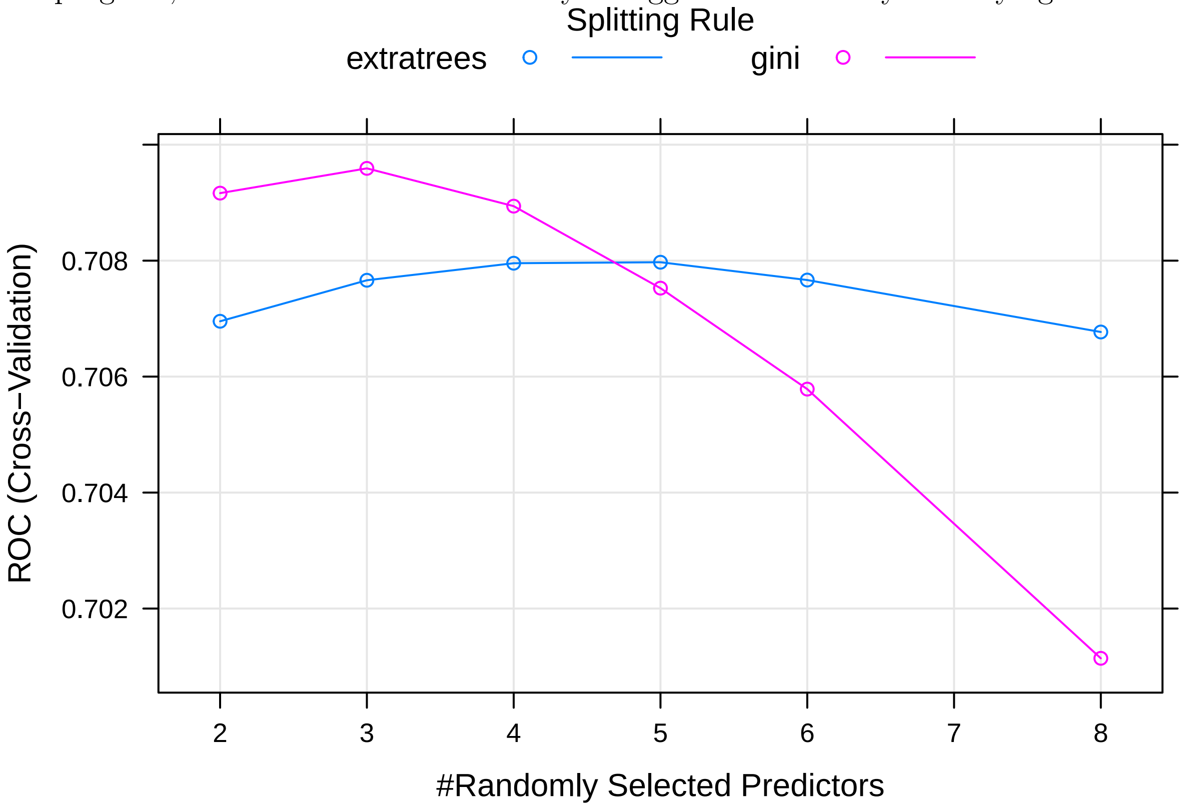
Results:

```
## Growing trees.. Progress: 49%. Estimated remaining time: 32 seconds.
```

Train_data Results and Discussion:

ROC:

The highest ROC here is roughly 70.8%, which is a significant improvement from my prior RF models. ### Sensitivity: True positive or Sensitivity in an RF model, tells us the proportion of people that left welfare that were correctly classified. The proportion correctly predicted by our model: mod_rf3, as leaving welfare, is roughly 41%. I find the results disappointing from the perspective of my project's objectives and from a policy-maker's view interested in gainig a better understanding in what factors contribute to one's decision to leave welfare. However, at the same time, it reinforces the underlying complexity in understanding why people behave in a certain manner and our often times irrational and unstructured decision-making process. ### Specificity: On the other hand, True Negatives or Specificity in our context, tells us about the percentage of individuals that DID NOT leave welfare, that were correctly classified by the RF model. The proportion correctly classified by our model: mod_rf3, as staying in welfare, amongst those who ACTUALLY stayed is 85.6%, which means our model is good in predicting individuals who will stay in the program for the five-year period, until the wage subsidy is fully exhausted. Additionally, this finding is insightful from a policy-maker's perspective because it may inform policymakers about the characteristics of individuals who are likely to stay in the welfare program. Specifically, our model: mod_rf3, may have performed well in predicting those who have stayed in the program because it is a more homologous group with common characteristics, compared to individuals who left the program, where the model relatively struggled in correctly classifying their status.



### Here, I ran a Logistic Regression model, with the same variables and observations that I ran in: train_data_rf. I did this to compare the performance of a linear regression model: Logisitc regression with my top performing Machine Learning model: Random Forest. In general, I would expect a Logistic Regression model to perform worst when there are non-linearities in the data, due to the Bernoulli distribution assumption, while

a major advantage of a machine learning model such as RF over linear models is that it can easily handle non-linearities in the data. Interestingly, when comparing my mod_logit and mod_rf3 models, on the basis of ROC, we find that RF performed slightly better at, ROC=0.7072, and mtry = 4, splitrule = extratrees and min.node.size = 4, compared to an ROC of 0.7070 for Logistic Regression. However, given that my a major objective of my project is correctly predicting individuals that leave welfare, we should also look at the proportion of TP or Sensitivity. Again, in my context, this is the proportion of individuals who actually left welfare, and were correctly classified by the model. Logisitc regression performed marginally better in that area, by correctly classifying the status of 43% of individuals who left the program, compared to a True Positive proportion of roughly 41% under the RF model.

```
## Run a Logisitc Regression Model,
## Estimate the results
mod_logit <-
  train(left_welfare ~ .,
        data=train_data_rf, # Training data
        method = "glm", # logit function
        metric = "ROC",
        na.action = na.omit,
        trControl = control_conditions
  )
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```
# mod_logit
```

RF Model Improvements:

Based on the above RF model results, paricularly on TP or Sensitivity, it appears that as the Sensitivity rises as I increase my mtry, which is the number of predictors that our model is randomly selecting. Furthermore, I will experiment with a larger node.size, or minimum number of observations in each terminal node, increasing it from 4 to 5. Therefore, by tuning the parameters, I attempt here to increase TP or the proportion of individuals who actually left welfare, that is correctly classified by my RF model.

Updated RF Model Results:

The best performing RF model below, using ROC as my metric, is the one where mtry = 3, splitrule = gini and min.node.size = 5, where ROC= 70.8%. However, given that I am interested in improving my model's Sensitivity performance, it appears that tuning my RF model's parameters above was effective in achieving that goal. Furthermore, my TP or Sensitivity proportion increased from roughly 43% in mod_rf3, to 46.6% in mod_rf4. In other words, my RF model now correctly classifies the status of 46% of individuals who left the program. It is crucial to point; however, that has while my TP proportion has improved, my ROC has declined slightly to 68%, rather than 70.5% under mod_rf3. Additionally, my Sensitivity improvement has also come at the cost of a lower Specificity or True Negative proportion, declining from 84% in mod_rf3, to approximately 80% in mod_rf4. Putting my results from both models: mod_rf3 and mod_rf4, respectively, in explicit statistical terms: One may reasonably argue that after tuning my model's parameters and raising my Sensitivity Or True Positive proportion, I essentially reduced Type one error/False Positive, or the proportion of people incorrectly classified as stayed, when they have actually left the program, which is evinced by a higher TP proportion in mod_rf4, compared to mod_rf3. A consequence or cost of this improvement in Sensitivity, is that I increased Type 2 error/False Negatives, thus my model is incorrectly classifying more individuals as Left, when they have actually stayed in the program. Again, this is also evinced by a lower Specificity in mod_rf4, compared to mod_rf3. For instance, a True Negative/Specificity proportion of 80%, suggests a False Negative proportion of 20% in mod_rf4, which is higher than the False Negative/Type 2 error proportion of roughly 16% we had in model mod_rf3 (Sharma, 2009). In conclusion, I believe a key driver for selecting our tuning parameters and how we evaluate one model to another, should be the policy relevance of our outcome. Specifically, in the context of welfare programs in Saudi Arabia, I think it is useful to understand and be able to correctly classify individuals that have ‘left’ welfare, before the system ‘officially’ mandates them to leave a welfare program. Therefore, I have placed greater emphasis when tuning my models on increasing Sensitivity/True Positives.

###Given that our best performing model was a non-linear model (i.e. Random Forest) model, then we may use metrics such as the Gini Coefficient to evaluate which variables had the most predictive/explanatory power. We will perform this, using a Permutation approach. Additionally, I will be addressing questions such as: ### What features appear most associated with the left_welfare outcome?

```
## Warning in partial.default(mod_rf3, pred.var = c("year_1"), ice = T, center
## = T, : Centering may result in probabilities outside of [0, 1].
```

Discussion:

Now, let’s run our best performing model, RF, on the test_data_rf:

I have encountered the following error while running my RF model on test_data_rf, which I tried to rectify by running the chunk above, but unfortunately it did not work.

What tools/methods did you consider but not use in the final analysis?

I had considered using CART as one of my machine learning techniques and have attempted to run the model several times, but my computer could not handle the complexity of the model. In fact, in the three times I ran the CART model, my R would shutdown w/out my control.Second, I intended to use a dummy variable to categorize geographic regions as urban, suburb, and rural areas. However, the Statistical Agency in Saudi Arabia had not updated and released recent on the categorization of geographic regions since 2010, but given the drastic changes in where people currently live compared to 2010, I I thought that using an old classification would not produce meaningul results. ### How would you expand the analysis if given more time? The following will be conducted for my thesis purposes: First, I will manually match the two datasets that I had initially intended to use, based on school_name. 1- Teachers' Wage Subsidy Program, which I am currently using. 2- The Standardized Exams database Second, I will perform the same analysis that I had performed but on the school, rather than individual-level. Third, I will use the 2020 updated geographic unit classification, which will be released by Saudi Arabia' Statistical Agency in January, 2020. This is because I suspect that one's geographic area, particularly females, plays a key role in whether they leave the wage subsidy program.

References:

He, Benlan, Yong Shi, Qian Wan, and Xi Zhao. "Prediction of customer attrition of commercial banks based on SVM model." Procedia Computer Science 31 (2014): 423-430.

Sharma, Devashish, U. B. Yadav, and Pulak Sharma. "The concept of sensitivity and specificity in relation to two types of errors and its application in medical research." Journal of Reliability and Statistical studies 2, no. 2 (2009): 53-58.

Yukselturk, Erman, Serhat Ozekes, and Yalın Kılıç Türel. "Predicting dropout student: an application of data mining methods in an online education program." European Journal of Open, Distance and e-learning 17, no. 1 (2014): 118-133.

Xie, Yaya, Xiu Li, E. W. T. Ngai, and Weiyun Ying. "Customer churn prediction using improved balanced random forests." Expert Systems with Applications 36, no. 3 (2009): 5445-5449.