Classification of Income Data using a Logistic Model

Savannah Farney
Purdue University Fort Wayne

Purpose

There are many different factors that have been said to influence a person's income, education, gender, profession, etc. With this project we are interested in how well we can predict a person's income level with the information that's given in a data set. The goal of the project is to better understand the how to predict the category of a variable and see which variables are necessary for the prediction and which could be left out for the model. Throughout this project the goal is to find the best model for predicting whether a person made more than 50 thousand dollars in 1994 or less than 50 thousand dollars.

Data Used

The data was collected by Ronny Kohavi and Barry Becker using the 1994 Census bureau database and was found originally on Kaggle for people to try and correctly predict an observations income. The data includes both categorical and continuous variables. The categorical variables are: work class (ex. Private, state, self, …), education (ex. Master, 12th, …), marital status, occupation, relationship (ex. Unmarried, husband, wife…), race, sex, native country, and income (>50K or <=50K). The continuous variables are: age, final weight, education, capital gain and loss, and hours per week. Final weight is a weighted tally of socio-economic characteristics of the population. Capital gain is the money made from investing in an asset after it is sold while capital loss is the money lost from investing in an asset. Some cleaning of the data was needed. Some of the data was marked with question marks and needed to be changed to NA, the variable names of education-num, marital-status, and native-country was renamed to make referring to them easier in coding. Also, the class data needed to be changed from >50K or <= 50K to 0 and 1 due to the type of modeling that was decided on.

Materials

This project utilizes R and R studio as well as the logistic modeling packages that it contains. Since the data set contains a majority of categorical variables, a linear regression model would not work. With the categorical variables contained in the data set and the goal of predicting a categorial variable, it was decided to use a logistic model and then utilize the log of odds to classify the observations into two groups, >50K and <=50K.

In order to create a model, the original data was split into two sets, a training set and a testing set. In this way, models could be created and then tested to see if the models would behave the same way on a new data set that was not used to create the models. When creating the two data sets, the original data was split in a way that 75% of the observations in each of the data sets had an income level of less than $50,000. This was done with the thought that it would help keep the data sets similar so to not throw off the classification when going from the training data to the testing data. Once the data sets were created, the next step was to select the models.

One of the models that was decided on was the full model utilizing all of the predictors. The other models were selected using the model selection with the best subset with sequential replacement. With this model selection, the best combinations of predictors were found for each subset of predictors. While there were 14 variables to begin with, there ended up being 96 predictor variables possible with the logistic model because the logistic regression model treats each of the categories in the categorical variables as its own predictor variable. This caused some issues that will be mentioned later. In order to pick the best models, there were three different values that were examined. The adjusted R^2, CP, and BIC (Bayesian information criterion) values. The R^2 value is the proportion of variation in the depended variable (income) that is explained by the predictors. The adjusted R^2 is adjusted for large amounts of predictors. The CP and BIC values are both values that look at how the error of the model increases with more predictors, a larger BIC/CP means a larger test error, so the model with a smaller BIC or CP is better. From these, three models were decided plus the model with all of the predictors for a total of four models. Once the best number of predictors were determined, it was time to create the models. Since most of the predictors came from a few categorical variables, it was interesting to create the models, because when creating the models, the function takes in the variables in the data set and the models from the best selection usually had at least one predictor from each of the variables in the data set. In order to create these models, the categories under each variable that were not needed in the model were combined into one category labeled 'other'. Choosing to relabel the data in this way meant that a model that was relabeled for 62 predictors actually contained more than 62, because of the other categories under several of the original categorial variables. Handling the issue in this manner does have its limitation which will be mentioned later on.

Once the new data sets were formed each of the four models were created. From these models, the log of odds ratios were found. An odds ratio is the probability of success over the probability of failure. In this case a success is labeled as having an income of more than $50,000 and a failure is having $50,000 or less in income. Since the odds ratio can range from 0 to infinity, we look at the log of the odds ratio, which makes the numbers range from 0 to 1. This makes it easier to determine a cutoff value in order to classify the observations into the two categories. All of the observations with a log odd value more than or equal to the cutoff value will be labeled as a success and all of the others are labeled as failures. Several cutoff values were tried in order to see if something besides 0.5 would be best. Once this is complete the models are evaluated and compared by finding out how well the models classified the income of each observation vs their actual income. To do this we look at the misclassification of each model for each of the cutoff values that were tested. The model misclassifies an observations income if the actual income was less than 50K and it was classified as a success according to the log of the odds ratio, or if the observations income was more than or equal to 50K and the model classified it as a failure. In the next section we will see the results from this project.

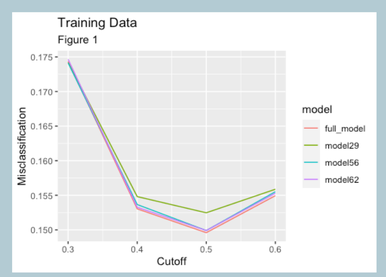
Results

Before creating the models, the best number of predictors for the models needed to be determined. By examining the adjusted R^2, CP, and BIC that were mentioned earlier, it shows that the model from the best subset model selection that had the best adjusted R^2 values was the model with 62-predictors. The best model according to the CP value was 56-predictors and the 29-predictor model had the lowest BIC value. While there were other models that had similar adjusted R^2, CP, and BIC, these were the models with the best values. Because of this, these three models will be used as possible models. The model with all of the predictors was also used, so there were four models to create and compare. It is interesting to note that for the 62, 56, and 29 predictor models, a lot of the native country predictors were not used (none being used for the 29-predictor model), as well as some education predictors, such as 7-8th grade, 12th grade, Doctorate, and professional school, marital status of separated or widowed, and an occupation of armed forces. Please note that this is not the full list and is different for each of the models. This means that when the combination of other variables is used, as predictors these categories are not as significant to help build the model and are therefore left out of them.

After creating the models, the log of odds ratios were found. Using different cutoff values, the data was classified as a success or failure by using the log odds values from the different models. The percent of misclassification of each model was then found and is shown in the chart below. From the chart it can be seen that depending on the cutoff value, the models incorrectly classified the data 14-17% of the time. Since this is the amount of times that the model was incorrect, the model that performs the best would have the lowest percent of error. To better see this, figure 1 and 2 show the misclassification (as a decimal) for each cutoff value separated by the model. From this it can be seen that for each model in both the training data and the testing data the cutoff of 0.5 misclassifies the data the least amount of time for each model. From the graph it can also be seen that the model using all of the predictors has the lowest misclassification of all of the models. In the training data it misclassified the data 14.95% of the time and with the testing data it misclassified it 15.13% of the time. While the training data was better classified, this is expected because the model was created using the training data set. The 62-predictor and 56-predictor models both misclassified the data 14.99% of the time in the training data, but misclassified the testing data differently, with the 62 – predictor model having a misclassification 15.14% of the time and the 56- predictor model 15.18% of the time. By adding the extra 6 predictors, the misclassification the model only had a difference of .04% and to add in all of the predictors, it was only better by .01% compared to the 62-predictor model. Due to there not being a significant difference, some might argue that the full predictor model does not perform with a great enough difference to justify the extra 30+ predictors. With any of these models the income of the observations is correctly classified at least 84.5% of the time.

Cutoff values	Model							
	Training Data				Testing Data			
	62	56	29	Full	62	56	29	Full
0.3	17.46607	17.41302	17.42186	17.43513	17.2656	17.2645	17.3319	17.3118
0.4	15.32647	15.36625	15.48.119	15.30436	15.4621	15.4887	15.6743	15.4621
0.5	14.9905	14.9905	15.24689	14.95955	15.1439	15.1837	15.5019	15.1306
0.6	15.52982	15.55192	15.58729	15.49445	15.555	15.4356	15.6876	15.555

Training Data
Figure 1



Test Data
Figure 2

Conclusion

With this data set, there are some cautions. The data set that was utilized is from 1994. Due to this the results that were found best for modeling a person's income in 1994 might not be the best for modeling newer data. Another limitation is the fact that for the logistic models each of the categories for each of the categorical variables became its own predictor. This is problematic because when it came to building the models, some of the categories of the variables were used as predictors while others were not. Due to the way the logistic model was built using the variables of the data set, in order to include some of the predictors and not others. For example, the variable native country created 40 possible predictor variables. Of these, only a few of them were significant enough to be used in the 62-predictor model. In order to create the model, the categories that were not labeled as a predictor were all relabeled into an 'other' category, which ended up cutting down on the predictors in the model. One of the problems with this method is that the 62-predictor model actually ended up with more than 62 predictors. Because of this, it was different than the 62-predictor model given from the best subset model selection. Also, the significances of some of the predictors were possibly different because of the 'other' category being a predictor in the model. For the future, a thought is to try splitting up the original variables into their categories and making the observations true or false. For example, with native country it would be split by making a variable for the US and then have all of the observations that were labeled as US be 1 (true) and all others 0 (false). This would then continue for all of the possible countries as well as all of the other categorical variables.

These models obtained during the project correctly classified the data at least 84.5% of the time with a cutoff value of 0.5. For the future, it would be recommended to try cut off values of 0.45 or 0.55 to see if they would increase the number of correct classifications. It would also be recommended to try and manipulate the data to avoid the 'other' predictor and see if the results have a significant change.

Work Cited

Learning, UCI Machine. "Adult Census Income." *Kaggle*, 7 Oct. 2016, https://www.kaggle.com/uciml/adult-census-income.