P8106_midterm_at3535

Amanda Tsai

3/16/2021

Introduction

Our data set was extracted from the 1994 Census bureau database and includes adult census information on income and variables that may affect a person's yearly earnings. Income is something very relevant and important to all households and also the government. We are interested in learning about which variables are statistically important in determining whether a person makes over $50k per year and which models have the best performance in predicting a person's income level.

Our data set have 14 attributes, including age, work class, education level, number of years of education, marital status, occupation, relationship status, race, sex, capital gain, capital loss, hours worked per week, native country, and final weight, an estimate number derived such that people with similar demographic characteristics have similar weights. There is also a class label ("income") indicating whether the individual's income is over 50K or not. The continuous variables are age,final weight, education number, capital gain, capital loss and hours per week. The others are categorical variables. This data set initially had 32 thousand rows.

Data Cleaning For categorical variables, to avoid not having enough entries for specific factor levels, resulting in those levels missing in partitions and cross validation (i.e. one level for a variable is only in the training set), we decided to group factor levels together to form bigger groups.

For work class, we combined all government jobs into one category,and all self-employed jobs into another. The total number of entries for "Never-worked" and "Without-pay" levels was 21, which is much less than 1% of the total data, so we decided to remove them from the data set.

Similarly, we grouped the occupations into 6 categories, according to the 2018 US Census Occupation code list. The categories are MBSA (management, business, science and arts), Service, Sales, NCM (Natural Resources, Construction and Maintenance), PTM(Production, Transportation and Material Moving), and Military. The total number of entries under Military were less than 1% of the whole data so we removed those entries as well.

We also grouped all marital statuses that started with "Married" into one "Married" category and all countries that were not "united-states" into one "others" category.

We removed the "education" variable as it is related to education years but doesn't give as specific information as the latter.

For capital gain and capital loss, we noticed that roughly half of the entries were 0 and half were non-zeroes. Therefore we thought that it would be better to consider these two as binary variables("0" and "non-zero") instead of as continuous variables.

A summary of the cleaned data set is shown.

```
##       age          workclass         fnlwgt        education_num
##  Min.   :17.00   Gov    : 4342   Min.   :  13769   Min.   : 1.00
```

Directly including software output tables is discouraged

```
##  1st Qu.:28.00   Private:22696   1st Qu.: 117808   1st Qu.: 9.00
##  Median :37.00   SelfEmp: 3657   Median : 178530   Median :10.00
##  Mean   :38.44                   Mean   : 189845   Mean   :10.13
##  3rd Qu.:47.00                   3rd Qu.: 237319   3rd Qu.:13.00
##  Max.   :90.00                   Max.   :1484705   Max.   :16.00
##       marital_status    occupation          relationship
##  Divorced    : 4258   MBSA   :12901   Husband      :12698
##  Married     :14737   NCM    : 7087   NotInFamily  : 7861
##  NeverMarried: 9902   PTM    : 1596   OtherRelative:  916
##  Separated   :  959   Sales  : 3650   OwnChild     : 4519
##  Widowed     :  839   Service: 5461   Unmarried    : 3269
##                                       Wife         : 1432
##                race           sex         capital_gain      capital_loss
##  AmerIndianEskimo:  285   Female: 9925   Min.   :0.00000   Min.   :0.00000
##  AsianPacIslander:  973   Male  :20770   1st Qu.:0.00000   1st Qu.:0.00000
##  Black           : 2907                  Median :0.00000   Median :0.00000
##  Other           :  248                  Mean   :0.08428   Mean   :0.04756
##  White           :26282                  3rd Qu.:0.00000   3rd Qu.:0.00000
##                                          Max.   :1.00000   Max.   :1.00000
##  hours_per_week       native_country          income
##  Min.   : 1.00   Other       : 3213   LessThan50K:23046
##  1st Qu.:40.00   UnitedStates:27482   Over50K    : 7649
##  Median :40.00
##  Mean   :40.95
##  3rd Qu.:45.00
##  Max.   :99.00
```

Exploratory Analysis and Visualization

Training and Test data

Graphing the relationship between different variable classes and income level:

For the categorical variables with multiple classes, we can see that in workclass, those working for private companies have a higher proportion of incomes less than 50K. In marital status, those divorced or never married have a much higher proportion of incomes less than 50K. Within occupation, those working in the services and PTM sectors have more observations having lower incomes. Within race, those who are white have more people with incomes higher than 50K.

Models

May consider MARS

I decided to use the GLM, LDA, QDA and GAM models. For all models, all the variables present in the original data set, except the "education" variable, were used as predictors. The response variable is income, which is binary.

GLM

(1)Model Assumptions:

The logistic regression assumes all observations to be independent of each other and there to be little to no multicollinearity among the independent variables. It also assumes the independent variables are linearly related to the log of odds and a large sample size will improve the accuracy of prediction.

(2)Findings:

From the GLM model, we can see that most of the predictors have p-values less than 0.05 and are statistically significant. However, race variables do not appear to be statistically significant.

From the coefficient estimates, we can see that relatively larger changes in the log odds of income level are associated with a change in marital status from the referenced "Divorced" to "Married" (an increase), occupation change from "MBSA" to "Service" (decrease), relationship status change from "Husband" to "Not-in-family"(decrease), and sex from "Female" to "Male" (increase).

(3) Limitations:

The main limitation is that the GLM model assumes that the relationship between the response variable and the predictors is linear. Also, when there is a large number of variables in the model, the model may lead to over-fitting on the training set and lead to a low prediction accuracy on the test data.

Another limitation is that it may not be able to capture the nonlinear trend and interaction terms

LDA and QDA

(1) Model Assumptions:

The LDA model also assumes that all the observations are independent of each other. It assumes all the independent variables are normal for each level of the grouping variable. It also assumes that there is equality of covariance among predictor variables across all levels of the response variable. Higher correlation between variables will lead to a lower accuracy in the predicting performance. The number of predictors has to be less than the sample size.

The QDA model also assumes that predictor variables are drawn from a normal distribution. QDA also has the same assumptions as the LDA model but doesn't assume there is equal covaraince among predictor variables.

(2) Limitations:

LDA also works for non-normal data by Fisher's argument

The main limitation of the LDA model is its assumption that all the independent variables are normally distributed and share a common variance. This will lead to a bad fit and high bias if some of the variables are skewed. This might affect the model performance for this data set as many variables, such as race and sex, are dominantly one level (e.g. race is mainly "white" and sex is mainly "male")

A limitation of the QDA model is its lack of ability to reduce dimensions. This would pose a problem if there are a large number of parameters but a small sample size.

GAM

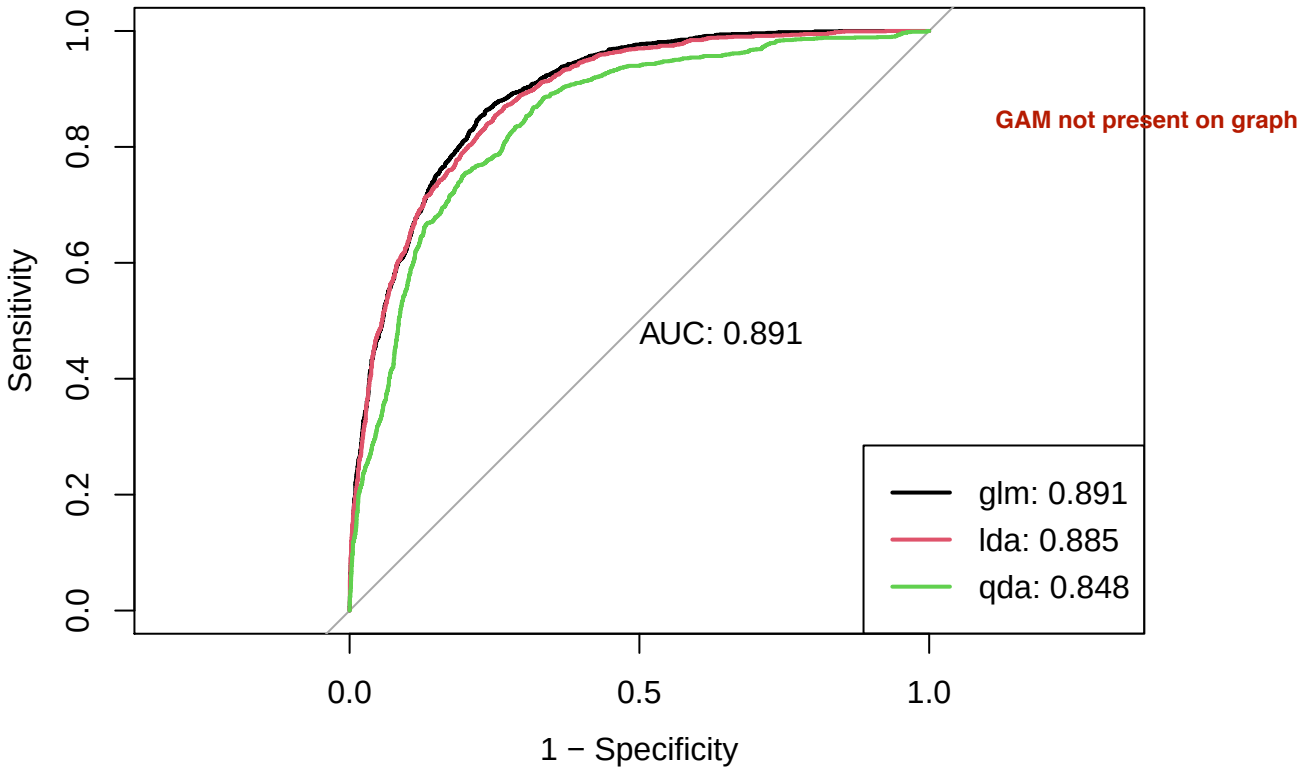
(1) Model Assumptions:

The GAM model assumes that the model function is additive and the components are smooth. The observations are statistically independent and the variables are not highly correlated with each other are other assumptions for this model. Specific observations will not influence the fit of the model.

(2) Limitations:

One of the main limitations of the GAM is its computational complexity. Also, like many nonparametric methods, GAM models have a high propensity to overfit. The choosing of smoothing parameters may affect the bias and variance tradeoff as well as cause either the fitted curves to be wiggly and hard to interpret or too smooth and missing out on important patterns.

ROC curves



Findings:

From the ROC graph, we can see that the GLM model has the highest AUC value and thus the best performance on the test data. However, the LDA and QDA models also have decent performance and can classify income with high accuracy.

Conclusions

For our data set, all of the predictor variables were statistically significant, with the exception of the race variable and the marital status variable that were not "married" or "never married".

The GLM model had the best test performance of all the models, which shows that the underlying relationship between income and the predictor variables used is possibly close to linear. GAM is more flexible than GLM and can reveal and estimate non-linear effects of the predictors on the response but did not perform better than the GLM model. **Model selection should be based on cross-validation rather than test performance.**

Among the models, LDA is less flexible than QDA, but is better at variable reduction. QDA is supposedly better than LDA with large data sets and does not suffer from the assumption that all predictor variables share a common variance. Some of the variables in our data set are skewed and the data set is large so I had expected QDA to be a better model fit.However, in our model-fitting, LDA actually has a better test performance than QDA.

I had expected that race would play an important role in income levels, as race is often tied to socio-economic status. However, from the GLM model it seems that it is not for our data set. This could be due to the low number of observations from non-white races.

Some insights we can make from our findings is that income is more heavily related to sex, education, and family structure.