Math 569 Final Project Report

Xi Sun, Rena Haswah

May 3, 2019

Percentage of Contribution

- Xi Sun: 50%
- Email: xsun46@hawk.iit.edu
- Rena Haswah: 50%
- Email: rhaswah@hawk.iit.edu

Abstract

In this project, we are going to predict if an observation is a pulsar star or not, based on two groups: profile group and DM-SNR group. After gaining insight on the frequency of response variable, pulsar stars occurence, a stratified sampling method is applied. Since the response variable is binary, a logistic regression is performed. The dataset is split into training and test sets, in order to see how well the model performs. After fitting the model, it is evident that two factors are not significant for the classification of a pulsar star. Backward selection is applied to find the "best" model which uses only a subset of the features. Since the dataset included features of two types, profile and DM-SNR, two separate logistic regression models were made to show the effect of each variable type on prediction of pulsar star occurence.

Introduction

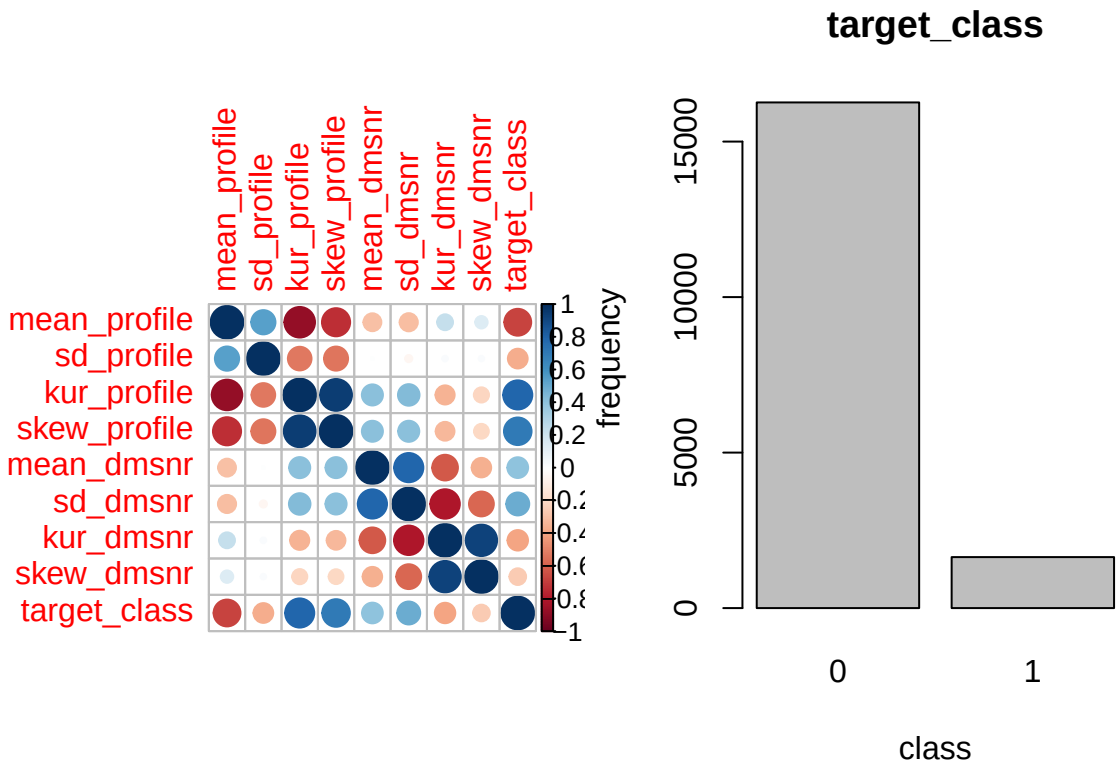
In order to investigate the universe, one is to actually get into space, and do some actual estimate or collection, the other thing is just do some analysis based on the data we have collected(at this moment, human beings cannot live in the universe for more than 15 seconds without any defences). In modern days, data science technology is being increased rapidly, and it is one of the most effecient way to let people understand something outside the circle. However, since the universe is huge, we cannot analyze everything at the same time. Therefore, we are going to estimate the pulsar star based on all the data we have. In the dataset, there are two groups: profile group and DM-SNR group, each group has it's own mean, standard deviation, kurtosis, and skewness. We are going to investigate which group affect our result the most, and how that differs compare to the "best" model(assuming the variables from the best model contains at least one variable from each group). And the challenge might be a poor prediction, in other words, all the information we have in the dataset cannot explain our target variable very well. Although we cannot solve that(at least in this datset/project), we can try different method or do some appropriate transformations which could lead to a more accurate result.

Problem Statement

The dataset is pulsar_stars.csv from kaggle dataset(see appendix for the link). A pulsar star is a highly magnetized rotating neutron star that emits a beam of electromagnetic radiation. We will be using the given dataset to predict the occurence of a pulsar star using information given about profile and DM-SNR. The profile categories describe the frequency of light emission that the stars release across space when they burst, which is their main classifying quality. The DM-SNR features describe delta modulation and signal to noise ratio.

The goal of this project is to use the information provided to predict occurence of a pulsar star, where occurence will be modeled as 1 if it is a pulsar star and 0 if not.

In order to present a simple understanding to support our idea as wells as about the dataset, the correlation plot and the response plot are shown below:



According to the above correlation plot, let's first ignore the "target_class" varaible(since it is our response), we can see that for all the predictors, they are somehow being "grouped", with two groups: profile group and DM-SNR group. For the profile group, we can see that the skewness and kurtosis are strongly positively correlated, and in the DM-SNR group, the skewness and kurtosis are strongly positively correlated as well. Another thing is that in the profile group, kurtosis and skewness are both negatively correlated with the mean, while in the DM-SNR group, the kuotosis is negatively correlated with it's standard deviation. Finally, another interesting thing is that, those two groups between each other are actually not correlated that much, and even the standard deviation from the profile group has almost no correlation(or correlation ≈ 0) with any other variable in the DM-SNR group.

Next, let's look at our "target_class" variable in the correlation plot, there are three variables being captured, mean_profile, with negative correlation; kurtosis_profile and skew_profile, with positive correlation, that is actually very interesting, since all those three variables are in the same group(profile group), so we are going to try to predict our response separately afterwards in order to figure out more pattern behind that. Well, at least for now, let's have a look and understanding our response.

Also, as we can see from the histogram, there are not too much 1's contained in our data, and the ratio is approximately 10:1 refer to 0 and 1 in our response, so one thing we need to be careful is making sure that we have contained the class 1 in both training and test set when doing the prediction.

Methodology

Since the Kaggle problem states that "The data set shared here contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators", which also could be visualized in the histogram that the ratio between 0 and 1 are approximately 10:1. Therefore, we decided to use stratified sampling method. And the statified sampling is a random sampling

method based on group. Here, we will use our response to be the target group. The model is then trained on 80% of the stratified data, and the remaining 20% is used to test the performance, which means that for the "target_class" variable, there will be 80% 0 and 80% 1 in our training data and rest 20% will be our test data. This guarantees that we have included both 0 and 1 in both trainning and test data.

Since the response contains only 0 and 1, we decided to fit a logistic regression for constructing the model. Since in theory, logistic regression model one of the best machine learning method when dealing with binary responses. In other words, if the probability of 0 occurance is grater than 0.5, then it will be classified as 0, and classified as 1 otherwise.

Also, in practice(based on the experience but will check in this project), the best model is usually not the model containing all the features, so the backward selection is applied in order to find the "best" model. Since for backward selection, during each time of the procedure, a non-significant variable will be dropped based on the AIC criteria, which prevents the error caused by dependencies between the regressors.

Last but not least, significance level of 0.95 is being used for doing analysis on the coefficients after building the logistic regression model. Since 0.95-significance level is generally being used in hypothesis test and we usually want our type I error $\alpha = 0.05$.

Analysis and Result

In order to give a better understanding, we first fit a logistic regression with all the features in the trainning data, and the summary statistic will be called to analyze the background of the model.

```
##
## Call:
## glm(formula = target_class ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3895  -0.1646  -0.0996  -0.0559   3.5401
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.759807   1.075561  -8.144 3.81e-16 ***
## mean_profile  0.029566   0.006594   4.484 7.33e-06 ***
## sd_profile   -0.030937   0.011285  -2.742  0.00612 **
## kur_profile   6.595870   0.332334  19.847  < 2e-16 ***
## skew_profile -0.620359   0.042913 -14.456  < 2e-16 ***
## mean_dmsnr   -0.029480   0.003660  -8.055 7.93e-16 ***
## sd_dmsnr      0.049647   0.008109   6.123 9.20e-10 ***
## kur_dmsnr    -0.001456   0.094697  -0.015  0.98773
## skew_dmsnr   -0.003585   0.003346  -1.071  0.28408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8766.6  on 14317  degrees of freedom
## Residual deviance: 2113.1  on 14309  degrees of freedom
## AIC: 2131.1
##
## Number of Fisher Scoring iterations: 8
```

According to the summary table, first, we can see that only "kur_dmsnr" and "skew_dmsnr" are not significant with significance level $\alpha = 0.05$, and all other variables are significant, with the AIC value being

2131.1. Since we are going to use the backward selection method for model selection with the AIC criteria, so recall the AIC formula:

$$\text{AIC} = n\log(\text{SSE}) + 2k$$

where n is the number of observations and k is the number of predictors.

Based on the significance from the summary table, we can actually make a claim that **the profile group is more robust than the DM-SMR group** for doing the prediction on pulsar stars. This claim will be varified by the end of this report.

```
##     true
## pred    0    1
##    0 3235   48
##    1   17  280
```

```
## [1] 0.9818436
```

According to the confusion matrix and accuracy rate, we can see that more than 98% of the data are classified correctly. In general, if this high accuracy model occurs, we can actually conclude that this dataset is very clean, so that the prediction is performing very well. Also, we are expecting this prediction to perform very well since investigating the universe is not an easy work, and we want to spend the least amount of time and do the most efficient work.

Next, the "best" model needs to be ffound by using backward selection based on AIC criteria mentioned above. And the "best" model with it's AIC are shown below.

```
## glm(formula = target_class ~ mean_profile + sd_profile + kur_profile +
##     skew_profile + mean_dmsnr + sd_dmsnr + skew_dmsnr, family = binomial,
##     data = train)
```

```
## [1] 2129.114
```

According to the model from backward selection and compare with the full model, we can see that the only variable being removed from the full model is "kur_dmsnr", so a new logistic regression model without the removed variable is constructed, and the confusion matrix and accuracy rate is checked.

```
##     true
## pred    0    1
##    0 3235   48
##    1   17  280
```

```
## [1] 0.9818436
```

According to the result above, it is amazing that both teh confusion matrix and the accuracy rate are exactly the same as the full model. Now, in theory, both model could be used for comparision with future models. In practiCe, however, the fewer the variable in a model, the less complexity the algorithms have.

Now, the model from profile group and DM-SNR group separately needs to be constructed and investigated. For the profile group, a logistic regression model is constructed with the summary table provided below.

```
##
## Call:
## glm(formula = target_class ~ mean_profile + sd_profile + kur_profile +
##     skew_profile, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3560  -0.1795  -0.1095  -0.0633   3.4499
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -9.341739   0.780815 -11.964  < 2e-16 ***
## mean_profile  0.044697   0.006204   7.205 5.82e-13 ***
## sd_profile   -0.045218   0.010637  -4.251 2.13e-05 ***
## kur_profile   7.533811   0.318303  23.669  < 2e-16 ***
## skew_profile -0.693472   0.047747 -14.524  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8766.6  on 14317  degrees of freedom
## Residual deviance: 2299.6  on 14313  degrees of freedom
## AIC: 2309.6
##
## Number of Fisher Scoring iterations: 8
```

As we can see from the summary table, all the variables including the intercepts are highly significant even with the significance level 0.001, and the AIC for this model is 2309.6 which is very close to the AIC for the best model which is 2129.114. According to the AIC formula, if we fix k and n, the higher the AIC value, the larger teh sum of squared error. Thereofore, sine the AIC from the profile group is very close to the AIC from the best model, so their error sould be close to each other as well, and their prediction accuracy should also be very close.

```
##     true
## pred    0    1
##    0 3236   53
##    1   16  275
```

```
## [1] 0.9807263
```

According to the confusion matrix, it is evident that the model does a very good job on classifying a pulsar star with the given profile inputs. The model's high accuracy may be explained by the fact that the dataset is already very clean and contains only the most important features. All four of the profile features were found to be statistically significant based on their p-values. This also supports the idea in the last paragraph that the prediction accuracy from the profile group is very close to the prediction accuracy from the best model. Now, do the same logistic regression for only DM-SNR features.

```
##
## Call:
## glm(formula = target_class ~ mean_dmsnr + sd_dmsnr + kur_dmsnr +
##     skew_dmsnr, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3814  -0.3088  -0.1915  -0.1138   4.0581
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.684157   0.366019   1.869  0.06160 .
## mean_dmsnr  -0.014363   0.001869  -7.687 1.51e-14 ***
## sd_dmsnr     0.010061   0.003445   2.920  0.00350 **
## kur_dmsnr   -0.642268   0.068509  -9.375  < 2e-16 ***
## skew_dmsnr   0.010662   0.003326   3.206  0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8766.6  on 14317  degrees of freedom
## Residual deviance: 5984.6  on 14313  degrees of freedom
## AIC: 5994.6
##
## Number of Fisher Scoring iterations: 8
```

Here, the "intercept" was the only one that marked with a lower significant code, since it is not one of the variable in our dataset, so we will not going to investigate that(and actually usually the intercept is not necessarily to be investigated). Next, look at the AIC value, we have AIC=5994.6, which is a lot higher than the previous one, and according to the AIC formula from above, we can conclude that the error from the DM-SNR group should performs higher than the profile group, so now we have the confusion matrix and accuracy rate as below and the accuracy rate should be relatively lower in this case.

```
##     true
## pred    0    1
##    0 3135  316
##    1  117   12
```

```
## [1] 0.8790503
```

As we can see, the accuracy is still very good for DM-SNR variables. However, by comparing with the profile group, the accuracy has become relatively lower. Actually this also supports the claim at the beginning of the analysis part that the profile group is more robust than the DM-SNR group when predicting pulsar stars. Therefore, we can say in case that more profile information is observed or we cannot get enough information for the DM-SNR, it is sufficient to predict the pulsar stars based on only the information form the profile group.

Conclusion

To conclude, logistic regression with the variables from profile group performs better than the varaibles in the DM-SNR group. Although in both cases, the accuracy is very high(87% accuracy from DM-SNR group is not low at all in general), since we have an even higher accuracy from the profile group, we will consider properties from the profile group more. And this would hold well with what is used in industry standards for analysis of these pulsar stars. The positive correlation simply verifies that this information is good to record for use of classifying random light beams in space as coming from a pulsar star or not. The binary classification lead to use of logistic regression, which gave a very high accuracy.

Appendix and Reference

dataset: https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star

The elements of Statistical Learning, second edition, by Trevor Hastie, Robert Tibshirani, Jerome Friedman.