# Classification in Sloan Digital Sky Survey RD14

Langwen Guan
05/12/2020

**Abstract**

The position and distance data of various celestial bodies recorded by the Sloan Digital Sky Survey opened the way for humans to study the structure of the universe. We can use the celestial bodies based on the collected information for discriminant analysis. In this article, firstly, visualize the collected 10,000 celestial body data, then use XGBOOST and Linear Discriminant Analysis (LDA) to classify and predict the observation to be a star, a galaxy or a quasar. Through the comparison of prediction accuracy, I find that XGBOOST has higher classification accuracy. Further research found that the use of balance Samples can effectively improve the prediction accuracy defects of unbalanced samples on LDA but not obvious on XGBOOST.

Key word: Sloan Digital Sky Survey XGBOOST LDA Discriminant analysis Classification Accuracy

## Introduction

Because I am interested in science and see a program about Sloan Digital Sky Survey RD14 in the Kaggle website. In order to improve my data science skills and knowledge about the Sloan Digital Sky Survey RD14, so I decided to use the knowledge I learned to analyze Sloan Digital Sky Survey by the given data. Before analyzing that, I will introduce the Sloan Digital Sky Survey briefly.

The Sloan Digital Sky Survey uses a wide-field telescope with a diameter of 2.5 meters, and the photometry system is equipped with five filters located in the u, g, r, i, and z bands to shoot celestial bodies. These photos are processed to generate a list of celestial bodies, including various parameters of the observed celestial bodies, such as whether they are pointy or extended. If it is the latter, the celestial body may be a galaxy and their brightness on the CCD. This is related to its magnitude in different bands. The contribution of the Sloan Digital Sky Survey: It records the data of nearly two million celestial bodies. The position and distance data of these celestial bodies has opened the way for people to study the large-scale structure of the universe.

In this report, this dataset consists of 10000 records of observations of space taken by the Sloan Digital Sky Survey. Every observation is described by 17 feature columns and 1 target column which identifies the observation to be a star, a galaxy or a quasar.

My curiosity leads me to ask and solve these aspects of questions: What the data visualization will be like? Can I use the given data to classify which one is a star, a galaxy or a quasar? Which method should I use to classify it to achieve the desired recognition accuracy? The classification accuracy is reduced due to the imbalance of the sample, can this problem be improved after changing to a balanced sample?

First, visualize these 10,000 pieces of data and make simple descriptive statistics on their distribution. Then use XGBOOST model and decision tree model to classify and predict the data and compare the prediction accuracy of the two method. Finally, I draw the following conclusions that XGBOOST is more accuracy than Decision Tree and using balanced data sets can lead to improved.

Due to time and paper limitations, I was unable to use Net and Decision Tree to fit tests for my model. More specific questions in the survey are required to be adjust to enhance this study. These deficiencies will be my improvement for future analysis.

## Data

The data I used is 10,000 observations of space taken by the SDSS given by the Kaggle website[1], and every observation is described by 17 feature columns and 1 target column which identifies the observation to be a star, a galaxy or a quasar. Besides, Table 1 is an explanation of the independent variables involved in the model for easy understanding (the more information can be found in the link 1).

The dataset in the Kaggle website given provides lots of information about space to explore. And the class column is the perfect goal for classification test.

Table 1: Explanation of Variable Abbreviation

| Variable Abbreviation | Explanation | Variable Abbreviation | Explanation |
|---|---|---|---|
| objid | Object Identifier | rereun | Rerun Number |
| ra | J2000 Right Ascension (r-band) | camcol | Camera column |
| dec | J2000 Declination (r-band) | field | Field number |
| u | better of DeV/Exp magnitude fit | specobjid | Object Identifier |
| g | better of DeV/Exp magnitude fit | redshift | Final Redshift |
| r | better of DeV/Exp magnitude fit | plate | plate number |
| i | better of DeV/Exp magnitude fit | mjd | MJD of observation |
| z | better of DeV/Exp magnitude fit | fiberid | fiber ID |
| run | Run Number | | |

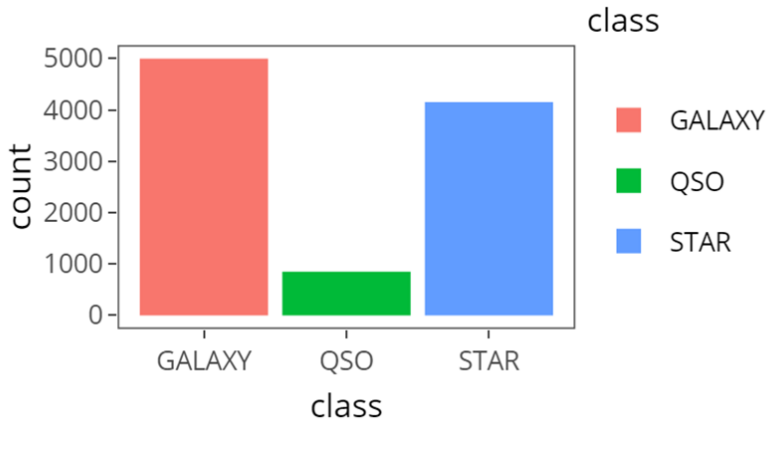
Based on the above question in the introduction section, next I will explore and process the data:

In terms of data exploration, my work mainly focuses on the visualization of data, and initially screens the features by plotting the density maps of the corresponding features for different values of the target columns. Finally, we filter out g, r, i, z, and redshift as the final features into the modeling step. At the same time, we have noticed that the sample has a certain degree of imbalance which can be shown in figure 1, so we have also dealt with this imbalance.

## Data visualization [2][3][4][5]

First, let's plot a bar chart to observe the respective Values of the three types of celestial bodies in the original observation data.

Figure 1: Values of Three Types of Celestial Bodies

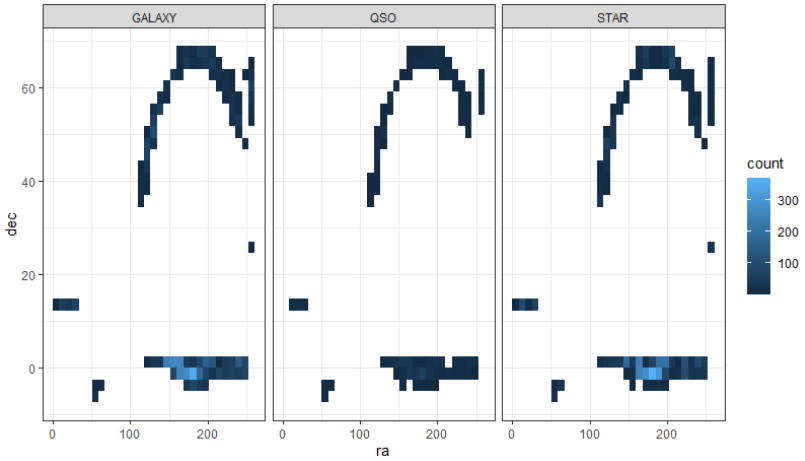



The figure 1 shows that the number of Galaxy in the data is the largest, with 4998 samples; the next is stars, with a number of 4152; and the number of QSQ is the smallest, with only 850 samples. It can be

found that the sample has a certain degree of imbalance in the target column, and the QSO category sample is very small.

Next draw a scatter plot of ra and dec.

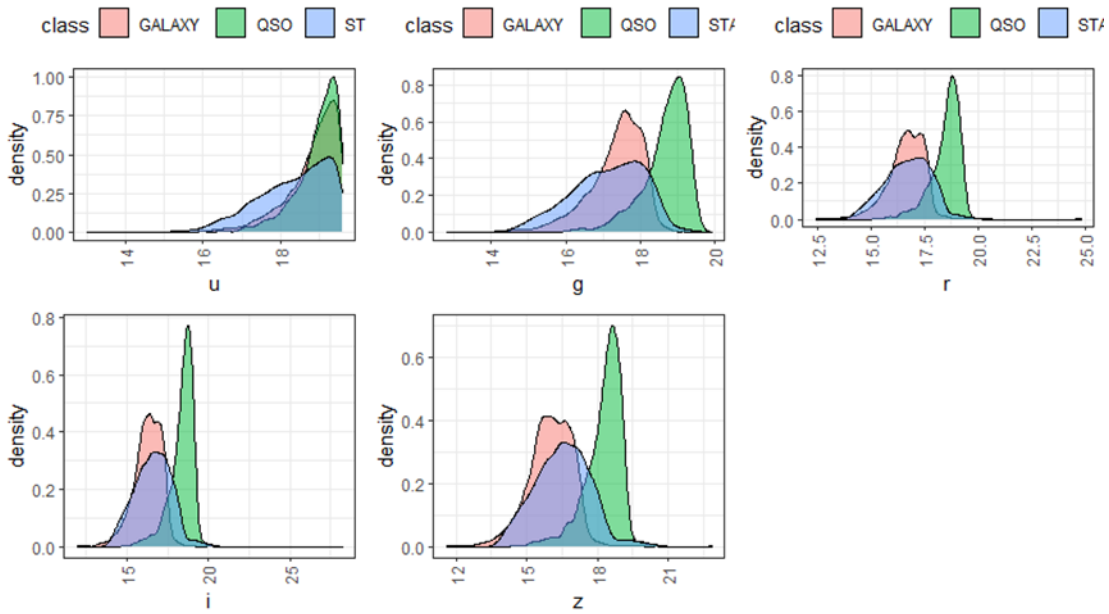
Figure 2: Distribution of Three Types of Celestial Bodies




We can find that the distribution of the values of the three categories on these two features is very Similar from the figure 2.

Further, I plotted the nuclear density curve of the five bands, u, g, r, i, z.

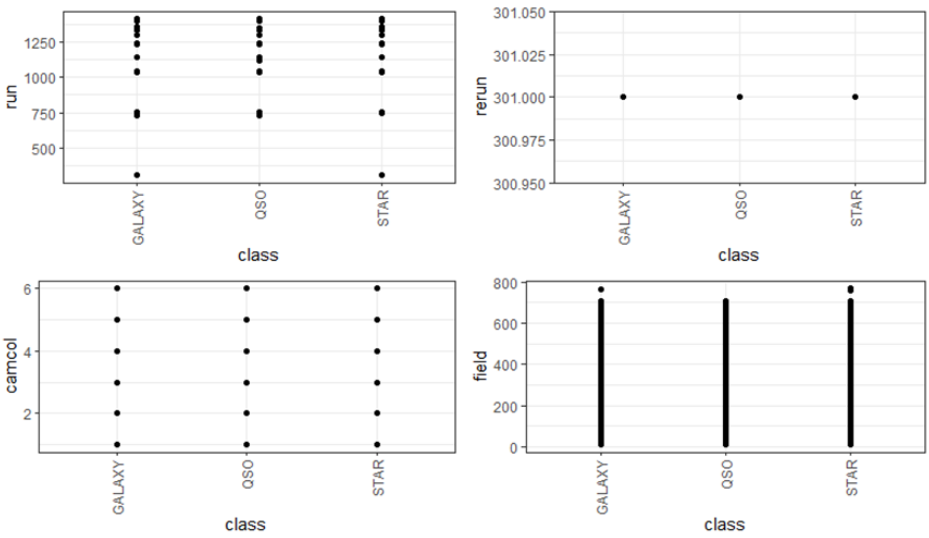
Figure 3: the nuclear density curve of the five bands, u, g, r, i, z




We found that there exists significant difference between QSO and the other two categories for g, r, I, and z responses in figure 3, and GALAXY and STAR are less distinct in these responses.

Then, continue to draw the scatter plots of run, rerun, camcol, and field

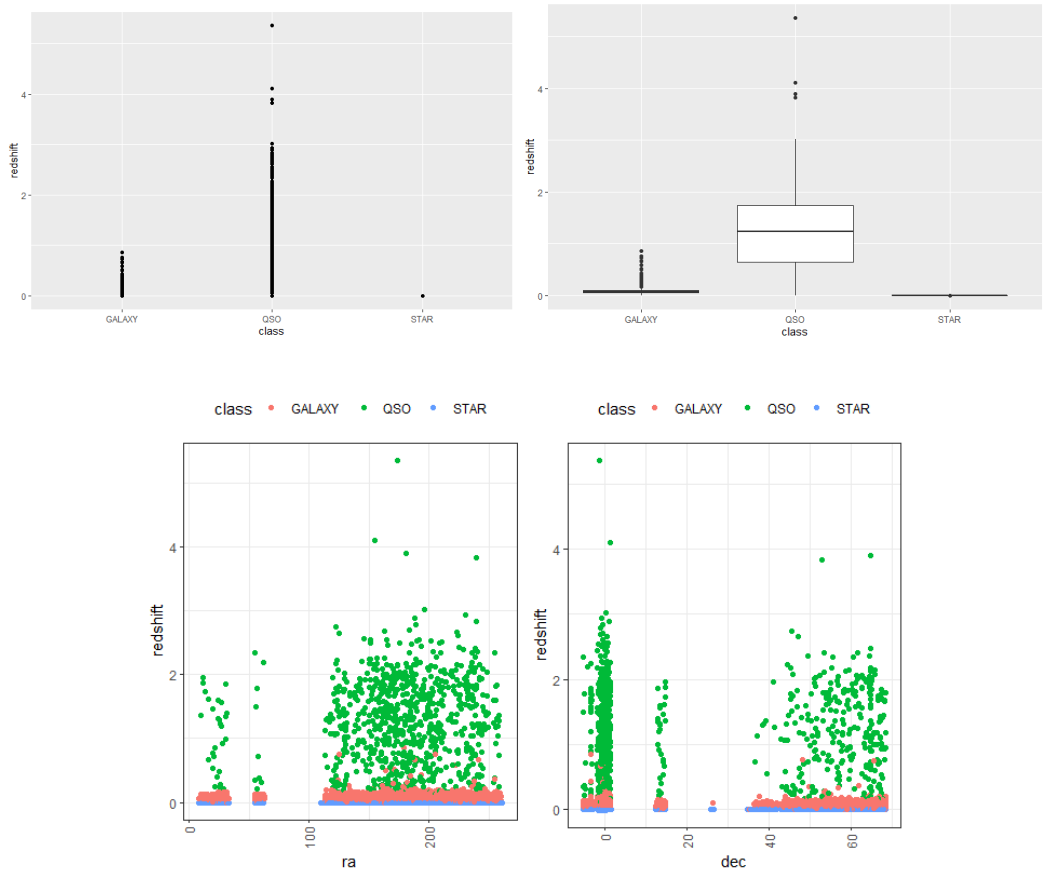
Figure 4: scatter plots of run, rerun, camcol, and field




It can be seen from the scatter plot that there is no significant difference among these three categories in the figure 4.

Next draw a redshift and a scatter plot for each category.

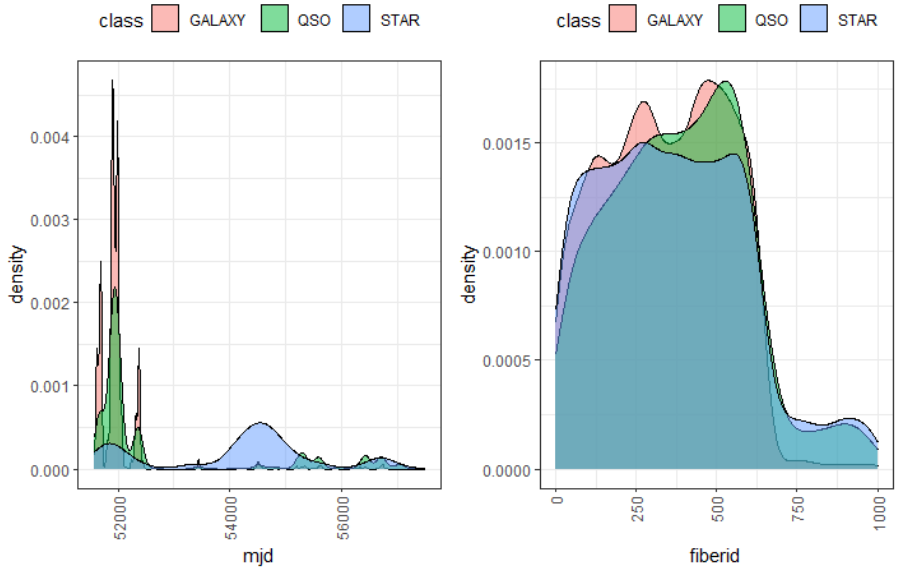
Figure 5: a redshift and a scatter plot for each category




It can be seen from the figure 5 that there is a large gap between the redshift distributions of the three categories and QSQ distribution is more discrete.

Further selecting the appropriate variables among other variables for further scatter plotting in figure 6.

Figure 6: the distribution of the mid and fibered variable




It can be seen that the mjd of STAR is somewhat different from the distribution of the other two types, and the classification effect of fiberid is not obvious.

Based on the above analysis, we decided to select the five features g, r, i, z, and redshift into the model section.

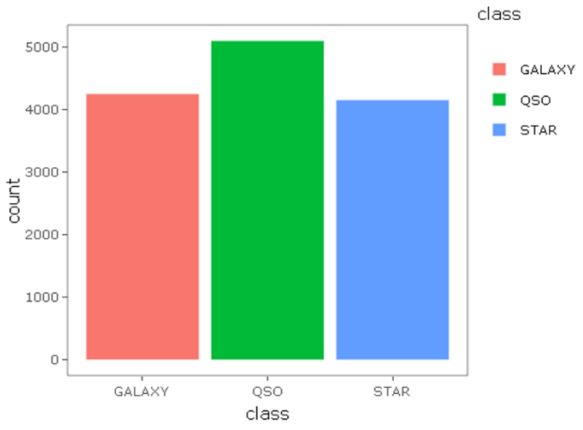
### Sample imbalance exploration

As shown in Figure 1 above, we can see that there is an obvious sample imbalance between the samples among the Star, Galaxy and QSQ. Generally, there are two methods for solving the imbalance: up-sampling and down-sampling[6]. Since the QSO sample is too small, using the down-sampling method will make the sample size used for modeling insufficient and will lose a lot of sample information. Therefore, we use the method of up-sampling in QSO class to balance the sample.

The method we use is the SMOTE(Synthetic Minority Oversampling Technique[7] method, which randomly takes two points from the samples

satisfying k-nearest neighbors, and generates a point on the two-point line as a new generated sample. Then plot the bar of the balanced data set again which was shown in figure 7.

Figure 7: The Balanced Samples Bar of Three Classes




## Model and Results

In this part, in order to deepen the understanding of the model, the model and the result are placed

in the same section instead of being separated.

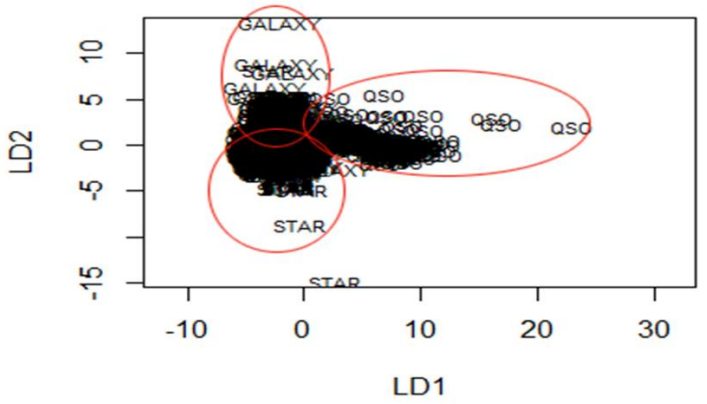
After the first step of data exploration to process, we selected the five characteristics which is g, r, i, z, and redshift to analyze the model. At the same time, the sample was balanced. Next, the paper will show the modeling process. I will use two different models to build model, and in order to explore the impact of sample balance on model results, and I repeat the modeling process for comparison on unbalanced samples. Besides, the training set and the test set are divided into 3:7.

**Model 1: LDA (Linear Discriminant Analysis)[8][9][10][11]**

The idea of linear discriminant analysis is very simple: given a set of training examples, try to project the examples onto a straight line so that the projection points of similar examples are as close as possible, and the projection points of different examples are as far away as possible; When classifying, project it on the same straight line, and then determine the class of the new sample according to the position of the projection point.

The figure 8 plots the distribution of three types of samples on LDA1 and LDA2

figure 8: the distribution of three types of samples on LDA1 and LDA2




From the picture 8, The distribution of three types of samples on LDA1 and LDA2 can be roughly distinguished from the figure.

The table2 Sorted out the results on the initial data set which we use the linear discriminant analysis method to train and classify the model.

Table2: The LDA Results on Raw Train Set

| Raw train set | GALAXY | QSO | STAR |
|---|---|---|---|
| GALAXY | 3179 | 110 | 316 |
| QSO | 2 | 435 | 0 |
| STAR | 336 | 26 | 2596 |

In the raw train set, the accuracy equals (3179+435+2596)/7000=88.71%

Table3: The LDA Results on Raw Test Set

| Raw test set | GALAXY | QSO | STAR |
|---|---|---|---|
| GALAXY | 1334 | 59 | 130 |
| QSO | 2 | 207 | 0 |
| STAR | 145 | 13 | 1110 |

In the raw train set, the accuracy equals (1334+207+1110)/3000=88.37%

Then calculate the result on the balanced data set with LDA method.

Table4: The LDA Results on Balanced Train Set

| Balanced train set | GALAXY | QSO | STAR |
|---|---|---|---|
| GALAXY | 2608 | 296 | 284 |
| QSO | 4 | 3240 | 1 |
| STAR | 344 | 53 | 2621 |

In the balanced train set, the accuracy equals (2608+3240+2621)/9541=89.61%

Table5: The LDA Results on Balanced Train Set

| Balanced test set | GALAXY | QSO | STAR |
|---|---|---|---|
| GALAXY | 1120 | 114 | 129 |
| QSO | 2 | 1371 | 1 |
| STAR | 172 | 26 | 1116 |

In the balanced train set, the accuracy equals (1120+1371+1116)/4051=89.04%

We can draw the conclusion that the balanced operation of the data set can improve the learning ability of the LDA model.

**Model 2: XGBOOST**

Before building model, I will make a brief explanation for the XGBOOST idea.

XGBOOST[12][13][14] is one of the Boosting algorithms. The idea of the Boosting algorithm is to integrate many weak classifiers together to form a strong classifier. Because XGBOOST is a boosted tree model, it integrates many tree models to form a strong classifier. It has unique advantages in the prediction accuracy of the model and is very suitable for big data.

Just like the previous decision tree model, we use XGBOOST model to make classification predictions on the raw data set and balanced samples, and compare the prediction accuracy.

So the table 6-table7show the classification results and accuracy and the confusion matrix [15] on the raw data train and test data. The table8 shows the classification results and accuracy and the confusion matrix on the balanced samples train and test data.

Table 6: Results of Raw Data Test Confusion Matrix

| Data Train | GALAXY | QSO | STAR |
|---|---|---|---|
| GALAXY | 3477 | 28 | 12 |
| QSO | 20 | 550 | 1 |
| STAR | 0 | 3 | 2909 |
| Accuracy | | (3477+550+2909)/7000=0.9909 | |
| Sensitivity | 0.9943 | 0.94664 | 0.9956 |
| Specificity | 0.9886 | 0.99673 | 0.9993 |
| Pos Pred Value | 0.9886 | 0.96322 | 0.999 |
| Neg Pred Value | 0.9943 | 0.99518 | 0.9968 |

Table 7: Results of Raw Data Test Confusion Matrix

| Data Test | GALAXY | QSO | STAR |
|---|---|---|---|
| GALAXY | 1459 | 14 | 8 |
| QSO | 16 | 263 | 0 |
| STAR | 1 | 1 | 1238 |
| Accuracy | | (1459+263+1238)/3000=0.9867 | |
| Sensitivity | 0.9885 | 0.94604 | 0.9936 |
| Specificity | 0.9856 | 0.99412 | 0.9989 |
| Pos Pred Value | 0.9851 | 0.94265 | 0.9984 |

| | | | |
|---|---|---|---|
| Neg Pred Value | 0.9888 | 0.99499 | 0.9955 |

Table 8: Results of Balanced Test and Train Data Confusion Matrix

| **Data Train** | GALAXY | QSO | STAR | **Data Test** | GALAXY | QSO | STAR |
|---|---|---|---|---|---|---|---|
| GALAXY | 2912 | 36 | 8 | GALAXY | 1275 | 16 | 3 |
| QSO | 40 | 3548 | 1 | QSO | 18 | 1493 | 0 |
| STAR | 0 | 0 | 2906 | STAR | 0 | 0 | 1246 |
| Accuracy | (2912+3458+2906)/9451=0.991 | | | Accuracy | (1275+1493+1246)/4051=0.9909 | | |
| Sensitivity | 0.9864 | 0.99 | 0.9969 | Sensitivity | 0.9861 | 0.9894 | 0.9976 |
| Specificity | 0.9932 | 0.993 | 1 | Specificity | 0.9931 | 0.9929 | 1 |
| Pos Pred Value | 0.9851 | 0.9886 | 1 | Pos Pred Value | 0.9853 | 0.9881 | 1 |
| Neg Pred Value | 0.9938 | 0.9939 | 0.9986 | Neg Pred Value | 0.9935 | 0.9937 | 0.9989 |

For ease of understanding, let's explanate the confusion matrix first.

Table 8: Confusion Matrix

| | | Reference | |
|---|---|---|---|
| | | Positive | Negative |
| Prediction | Positive | TP | FP |
| | Negative | FN | TN |

Note:

- Predicted to be positive and actually positive----True Positive (TP)
- Predicted to be positive and actually negative----False Positive (FP)
- Predicted to be negative and actually positive----False Negative (FN)
- Predicted to be positive and actually negative----True Negative (TN)

Sensitive: TP/(TP+FN)

Specificity: TN/(FP+TN)

Pos Pred Value: TP/(TP+FP)

Neg Pred Value: TN/(FN+TN)

Accuracy: (TP+TN)/(TP+TN+FP+FN)

From the above results, we found that the XGBOOST model is slightly better than the LDA model, and the impact of the balanced operation of the data set on the results is not obvious for the difference of accuracy between raw data and balanced set are subtle.

# Discussion

## Analysis

Table 9: The Comparation Between LDA Model and XGBOOST Model

| Model | LDA | | | XGBOOST | | |
|---|---|---|---|---|---|---|
| **Raw Data Set** | GALAXY | QSO | STAR | GALAXY | QSO | STAR |
| GALAXY | 1334 | 2 | 145 | 1459 | 14 | 8 |
| QSO | 59 | 207 | 13 | 16 | 263 | 0 |
| STAR | 130 | 0 | 1110 | 1 | 1 | 1238 |
| Accuracy | 0.884 | | | 0.9867 | | |
| Model | LDA | | | XGBOOST | | |
| **Balanced Data Set** | GALAXY | QSO | STAR | GALAXY | QSO | STAR |

| | | | | | | |
|---|---|---|---|---|---|---|
| GALAXY | 1120 | 2 | 172 | 1275 | 16 | 3 |
| QSO | 114 | 1371 | 26 | 18 | 1493 | 0 |
| STAR | 129 | 1 | 1116 | 0 | 0 | 1246 |
| Accuracy | | 0.89 | | | 0.9909 | |

The above table represents the result of integrating all the models, of which we are mainly concerned with the Accuracy indicator.

From the table we can draw the following conclusions:

XGBOOST has higher prediction accuracy than LDA model; but balanced samples can improve the prediction accuracy of LDA model. However, the change in XGBOOST model is not significant.

Simultaneously, About Model

LDA may not be a good model when the boundaries are particularly complex. But Models like XGBOOST are particularly suitable for classification on large data.

Why is there no significant improvement in XGBOOST performance?

The data set has a particularly good feature: Redshift, So the "more complex" XGBOOST did not bring much benefit. The number of features is small, the complexity of the data set is not large enough, and the XGBOOST model cannot reflect its advantages.

Balancing data sets does improve XGBOOST model performance. However, the best adjustment is still unknown.

### Limitations and Future Work

Due to time and paper limitations, this paper was unable to use Net and Decision Tree[16] to fit our model. More specific questions in the survey are required to be adjust to enhance this study. As the paper shown above, the data set itself has a particularly good feature: Redshift, so that Decision Tree may get a very good classification effect by selecting this variable. These deficiencies will be my improvement for future analysis.

Due to the limited knowledge learned, only a confusion matrix is used in the prediction accuracy, and multiple indicators are not used to measure it, so I should find other Statistics to test the model's robustness.

## Appendicies

### Citations


[1]. Data source. LennartGrosser(2018) Sloan Digital Sky Survey DR14 https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey

[2]. Carson Sievert.plotly: Create Interactive Web Graphics via 'plotly.js'. Version:4.9.2.1 https://CRAN.R-project.org/package=plotly

[3]. Hadley Wickham (2020).ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. Version:3.3.2https://CRAN.R-project.org/package=ggplot2

[4]. Claus O. Wilke (2020) Version:1.1.0. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2' https://CRAN.R-project.org/package=cowplot

[5]. Jeffrey B. Arnold [aut, cre] (<https://orcid.org/0000-0001-9953-3904>), Gergely Daroczi [ctb], (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. Version:4.2.0. https://CRAN.R-project.org/package=ggthemes

[6]. Dragoș Dumitrescu, Costin-Anton Boiangiu(2019);A Study of Image Upsampling and Downsampling Filters; DOI https://doi.org/10.3390/computers8020030

[7]. Luis Torgo (2013). DMwR: Functions and data for "Data Mining with

R".Version:0.4.1.https://CRAN.R-project.org/package=DMwR

[8]. ML|Linear Discriminant Analysis(2019). https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

[9]. Rosenstein, Leslie D (2019). Research design and analysis: a primer for the non-statistician / Leslie D. Rosenstein, UT Southwestern Medical Center.

[10]. John Fox [aut, cre], Sanford Weisberg [aut], (2020). car: Companion to Applied Regression package version 3.0-10. https://CRAN.R-project.org/package=car

[11]. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wie(2020) Version:1.7-4 https://CRAN.R-project.org/package=e1071

[12]. Jason Brownlee(2016.08). A Gentle Introduction to XGBoost for Applied Machine Learning https://github.com/dmlc/xgboost

[13]. Tianqi Chen [aut], Tong He [aut, cre] (2020). xgboost: Extreme Gradient Boosting. Version:1.2.0.1. https://CRAN.R-project.org/package=xgboost

[14]. Brian Ripley [aut, cre, cph], Bill Venables [ctb](2020). Functions and datasets to support Venables and Ripley, "Modern Applied Statistics with S" (4th edition, 2002). Version:7.3-53 https://CRAN.R-project.org/package=MASS

[15]. Pablo Diez,(2018)Confusion Matrix in Machine Learning.https://www.sciencedirect.com/topics/engineering/confusion-matrix

[16]. Brian Ripley [aut, cre] (2019). tree: Classification and Regression Trees. Version: 1.0-4https://CRAN.R-project.org/package=tree

## Code

https://github.com/Owen-Guan/slogan-sky-code.