

Shawnee State University

Digital Commons @ Shawnee State University

Master of Science in Mathematics

College of Arts & Sciences

Spring 2024

Empirically comparing the performance of LDA and QDA when classifying customer sales data with different properties of normality and equality of covariance matrices

Ayuk Egbe Bate-Eya

Follow this and additional works at: https://digitalcommons.shawnee.edu/math_etd



Part of the [Other Mathematics Commons](#)

SHAWNEE STATE UNIVERSITY

Empirically comparing the performance of LDA and QDA when classifying customer sales data with different properties of normality and equality of covariance matrices

A Thesis

By

Bate-Eya Ayuk Egbe

Department of Mathematical Sciences

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science, Mathematics

22nd July 2024

Accepted by the Graduate Department

 7/22/24

Dr. Douglas Darbro, 22nd July 2024

Graduate Director, Date

The thesis entitled '**Empirically comparing the performance of LDA and QDA when classifying customer sales data with different properties of normality and equality of covariance matrices**' presented by **BATE-EYA AYUK EGBE**, a candidate for the degree of **Master of Science in Mathematics**, has been approved and is worthy of acceptance.

22nd July 2024

Date

22nd July 2024

Date



Dr. Douglas Darbro

7/22/24

Graduate Director



BATE-EYA AYUK EGBE

Student

ABSTRACT

Discriminant analysis is a statistical technique used to classify data into different classes. Many studies have compared different methods used to classify data as regards their performance. This study compares Linear Discriminant Analysis and Quadratic Discriminant Analysis under varying conditions of normality and the equality of covariance matrices. More precisely, this study seeks to determine which of the two techniques is better when classifying datasets with different properties of normality and equality of covariance matrices and aims to determine whether normality and equality of covariance matrices influence the prediction performance of each method. This study processes online stores' customer sales data. Though the data processed was randomly generated, it was close to reality, since the data generation took into account different aspects like the mean and standard deviations of purchases of a particular type of product for a given period. By varying such parameters as the mean and the standard deviation, approximate real-world datasets were obtained. These datasets were processed using LDA and QDA for classification and the ROC-AUC score was used as the performance metric for each method. By statistically comparing these metrics, information was obtained concerning which method performed better under certain conditions. The results indicate that LDA performs better than QDA when classifying online stores' customers based solely on their purchasing habits, but also reveal an insensitivity of LDA to changes in both normality and equality of covariance matrices. With these results, businesses with online stores will be able to choose wisely which classification method to use depending on the type of distribution contained in the dataset.

ACKNOWLEDGMENTS

I humbly acknowledge the guidance of Dr Douglas Darbro, who helped structure my thinking about how the research process should be done efficiently. Through his experience and guidance, I was able to organize this thesis into different chapters and for each chapter, into their respective composition. This organization enabled me to be goal-oriented in the research process and by this, I was able to complete this thesis in time.

TABLE OF CONTENTS

Chapter	Page
ABSTRACT	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS	v
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER 1	1
CHAPTER 2	13
CHAPTER 3	18
CHAPTER 4	27
CHAPTER 5	41
REFERENCES	45
APPENDIX A	46
BIBLIOGRAPHY	56

LIST OF TABLES

Table	Page
Table 1. Dataset structure for normally distributed data	6
Table 2. Dataset structure using a Poisson distribution and a Binomial distribution.....	8
Table 3: Data organization for second-level data	10
Table 4: Normally distributed data.	24
Table 5: Poisson distributed data.	24
Table 6: Data from a multivariate Binomial distribution	24
Table 7: Dataset produced after running LDA and QDA on a first-level dataset.....	24
Table 8: Descriptive statistics for normally distributed data with equal covariance matrices.....	29
Table 9: Descriptive statistics for normally distributed data with unequal covariance matrices.	30
Table 10: Descriptive statistics for non-normally distributed data with equal covariance matrices.	31
Table 11: Descriptive statistics for non-normally distributed data with unequal covariance matrices.....	32
Table 12: Descriptive statistics for the normality property.	34
Table 13: Descriptive statistics for the equal covariance matrix property.....	34
Table 14: Tukey post-hocs for the effect of normality on LDA's performance.	38
Table 15: Tukey post-hocs for the effect of normality and equality of covariance on QDA	39

LIST OF FIGURES

Figure	Page
Figure 1. LDA and QDA performances for normally distributed data with equal covariances.....	29
Figure 2. LDA and QDA performances for normally distributed data with unequal covariances	31
Figure 3. LDA and QDA performances for non-normally distributed data with equal covariances	32
Figure 4. LDA and QDA performances for non-normally distributed data with unequal covariances	33
Figure 5. LDA performance by normality and equality of covariances.....	35
Figure 6. QDA performance by normality and equality of covariances.....	35

CHAPTER 1

This chapter presents an overview of the study, within which the discriminant analysis techniques of Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are informally introduced, including statements about their importance and applications. Emphasis is laid on the classification performances of these techniques since this study compares the performances of these two techniques. Also included in this overview is a description of the target population involved and the problem of concern, as well as a discussion about the significance and how the study findings can benefit businesses. A summary of related literature is also presented and is used to show how this study is related to current research in discriminant analysis. The problem statement and data organization are described clearly, and the chapter ends with a definition of certain terms and acronyms.

INTRODUCTION

Discriminant analysis is a very important statistical method for classifying data samples and has been studied extensively in research. It has been applied in credit scoring, mental state analysis using brain EEG data, and prostate cancer tissue classification. This study used an empirical approach to compare the prediction performance of Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis under the absence of normality and equal covariance matrices. Linear Discriminant Analysis uses a linear combination of certain features of data samples to separate data into different classes, while Quadratic Discriminant Analysis uses surfaces that can be described by a quadratic relationship of the features to separate classes. The data which was used in this study was a collection of randomly generated customer sales data which together represent varying customer databases for online stores. Many studies have compared these two techniques of classification, some using empirical methods, and

others not. Also, for most of the studies reviewed, the data under analysis was quite rigid as regards variation in normality. The expectation was that using randomly generated datasets that have different probability distributions would provide more generalizable knowledge as regards the performance of these two methods. The model performance metric that was used for each technique is the ROC – AUC. By comparing these performance values for the different models, information was obtained about which model performed better under varying conditions of normality and covariance matrices.

Online stores have become very common places where people go to buy products and these stores' owners are always looking for ways to maximize gains. Being able to categorize customer data based on different properties is beneficial for such companies since it can reduce marketing costs by tailoring marketing campaigns for target audiences. Also, by predicting customer categories through classification, online stores can provide customer-based product suggestions that match customer preferences. This study used different datasets representing customer data in different scenarios and applied LDA and QDA to them for classification. The expectation was that one technique was better than the other for data with certain distribution properties. The activities of data collection and data analysis were performed using R Studio.

BACKGROUND OF THE PROBLEM

The issue of which classification method is better for different use cases is an ongoing research area, where much research has been done in comparing the performance of different classification methods. Whereas some studies compared more than two classification methods, this study focused on comparing LDA and QDA only.

Also, the issue of analyzing prediction performance under varying sample properties like group sample size imbalance has been researched. Reviewed literature also shows that classification models can be compared because from these models, performance metrics like the ROC – AUC can be obtained.

Some studies like (Iain and Mues 2012) found that LDA performed better than QDA for most of the datasets involved, while (Subhrangsu and Tibarewala 2012) showed a significantly superior performance for QDA over LDA. Similarly, (Laurinda et al. 2017) showed that QDA outperformed LDA, while in (Haoyuan et al. 2017), LDA's performance was slightly better than QDA's.

Though some reviewed studies have analyzed the effect of covariance matrices on classification as in (Jing-Hao and Titterington 2007), which studied this effect only for LDA, most reviewed literature used datasets that appear not to vary in terms of normality constraints as in the cases of credit scoring data, forest fire data, and brain EEG data. So, it is natural to ask if such results are generalizable. The little variation in the datasets in terms of normality seems to be true also for the constraint on covariance structure since in most of the reviewed literature the covariance does not seem to vary much. So, it is natural to ask how distribution properties like normality and covariance structures affect prediction performance.

It is in this light that this study attempted to offer more generalizable knowledge concerning which classification method to use on datasets with different normality properties and covariance structures. By generating such datasets randomly and by applying each classification technique to them, performance metrics were gathered, and model performances were compared. Based on these, the authors were able to determine if one method was significantly superior to the other method in classifying data with certain constraints.

STATEMENT OF THE PROBLEM AND SIGNIFICANCE OF THE STUDY

This study addressed the following questions:

- When controlling for equal covariance matrices, does QDA perform better than LDA when classifying normally distributed data?
- When controlling for equal covariance matrices, does QDA perform better than LDA when classifying non-normally distributed data?
- Can the interaction between equal covariance matrices and normality explain the variation in classification performance between LDA and QDA?

By comparing the performance of one model to the other under varying conditions of normality and covariance structures, it was expected that more insight would be gained concerning the type of data each model would be most suitable for. As a result of this, companies that run online stores will be able to make more informed decisions concerning how to efficiently classify customers based on provided data.

PURPOSE OF THE STUDY

This study's objective was to show that one classification method is better than the other for normally distributed data and for non-normally distributed data in both cases where we have equal covariance structures and non-equal covariance structures. Also, a secondary objective was to analyze the interaction effect between normality and covariance structure on prediction performance. The results of the comparison should provide better confidence in the choice of method to use based on the

known distribution properties of the data. The results of the analysis of the interaction will provide even more support when choosing one method over another.

RESEARCH QUESTIONS

The following research questions were addressed:

- How does classification performance compare when using QDA vs LDA to classify data which is normally distributed and has equal covariance matrices?
- How does classification performance compare when using QDA vs LDA to classify data which is normally distributed and has unequal covariance matrices?
- How does classification performance compare when using QDA vs LDA to classify data which is not normally distributed and has equal covariance matrices?
- How does classification performance compare when using QDA vs LDA to classify data which is not normally distributed and has unequal covariance matrices?
- To what extent do normality and equality of covariance matrices affect the classification performance of QDA?
- To what extent do normality and equality of covariance matrices affect the classification performance of LDA?

RESEARCH DESIGN

Two levels of datasets were involved in the study. The first level of datasets contained randomly generated customer data on which classification was applied. These datasets were grouped into the following four categories:

- Normally distributed data and equal covariance matrices
- Normally distributed data and non-equal covariance matrices
- Non-normally distributed data and equal covariance matrices
- Non-normally distributed data and non-equal covariance matrices

The first two groups had a similar structure and differed only in covariance structure. For the last two groups, there was a mixture of both the Poisson distribution and the Binomial distribution in the datasets irrespective of whether they had equal covariance structure or not.

Tables 1 and 2 below describe the information which was generated for the first-level datasets.

Table 1. Dataset structure for normally distributed data

Variable	Type	Purpose
CustId	Number	The customer's identifier.
Age	Number	The customer's age.
Gender	Text	The customer's gender. The possible values are "Male" and "Female".
AmntPurchElect2022	Number	The amount spent purchasing electronics in 2022. This is a

		normally distributed variable.
AmntPurchElect2023	Number	The amount spent purchasing electronics in 2023. This is a normally distributed variable.
AmntPurchClothing2022	Number	The amount spent purchasing clothing in 2022. This is a normally distributed variable.
AmntPurchClothing2023	Number	Amount spent purchasing clothing in 2023. This is a normally distributed variable.
LikesFashion.	Dichotomous	A dichotomous variable that represents whether the customer likes fashion or not. The possible values are “Yes” and “No”.
LikesElectronics	Dichotomous	A dichotomous variable that represents whether the customer likes electronic items or not. The possible values are “Yes” and “No”.

Table 2. Dataset structure using a Poisson distribution and a Binomial distribution

Variable	Type	Purpose
CustId	Number	The customer's identifier.
Age	Number	The customer's age.
Gender	Text	The customer's gender. The possible values are "Male" and "Female".
NumPurchFashion2022	Number	The number of clothing items purchased in 2022. This variable follows a Poisson distribution.
NumPurchFashion2023	Number	The number of clothing items purchased in 2023. This variable follows a Poisson distribution.
NumPurchElect2022	Number	The number of electronic items purchased in 2022. This variable follows a Poisson distribution.
NumPurchElect2022	Number	The number of electronic items purchased in 2022. This variable follows a Poisson distribution.
NumWeeksPurchCloth2022	Number	The number of weeks in 2022 where purchases were made for fashion items. This variable follows a Binomial distribution.

NumWeeksPurchCloth2023	Number	The number of weeks of 2023 where purchases were made for fashion items. This variable follows a Binomial distribution.
NumWeeksPurchElect2022	Number	The number of weeks of 2022 where purchases were made for electronic items. This variable follows a Binomial distribution.
NumWeeksPurchElect2023	Number	The number of weeks of 2023 where purchases were made for electronic items. This variable follows a Binomial distribution.
LikesFashion.	Dichotomous	A dichotomous variable that represents whether the customer likes fashion or not. The possible values are “Yes” and “No”.
LikesElectronics	Dichotomous	A dichotomous variable that represents whether the customer likes electronic items or not. The possible values are “Yes” and “No”.

--	--	--

The second level of data was created by running QDA and LDA on the first-level datasets. The following table describes the structure of this data.

Table 3: Data organization for second-level data

Variable	Type	Purpose
ID	Number	This is a number representing the dataset's identifier.
IsNormal	Dichotomous	This is a dichotomous variable that determines if the dataset contains data that is normally distributed. The possible values are "Yes" and "No".
EqualCovariance	Dichotomous	This is a dichotomous variable that determines whether the data subsets defined by target classes have equal covariance matrices. The possible values are "Yes" and "No".
RocAUCLDA	Number	This is a number that represents the ROC - AUC after classification by LDA.

RocAUCQDA	Number	This is a number that represents the ROC - AUC after classification by QDA.
-----------	--------	---

ASSUMPTIONS

The generated data was random, and it was assumed that the total spent on purchases by a particular customer followed a normal distribution while data like the number of purchases made within a year and the number of weeks in a year where purchases were made followed Poisson and Binomial distributions respectively. The study also assumed a maximum of 100 customers per data set.

One limitation of the study was the number of different probability distributions used to model different customer profiles. These distributions were: the Normal distribution, the Poisson distribution, and the Binomial distribution. An avenue for future studies will be to include other probability distributions like the Geometric Distribution, the Uniform Distribution, and the Gamma Distribution.

SUMMARY

This chapter presented an overview of the study, including the definition of the problem and the target population involved. A summary of current literature as regards the topic's area was also included, and showed how the study fits itself in the research area.

Research questions were presented, as well as a description of how data was collected and organized including an explanation of the use of each variable involved. Finally, the limitations and assumptions inherent in this study, and indications for future research were also introduced.

DEFINITION OF TERMS

LDA. *Linear Discriminant Analysis.*

QDA. *Quadratic Discriminant Analysis.*

Normal distribution. *This is a probability distribution for a continuous random variable given by the formula:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{ where } \mu \text{ is the mean and } \sigma \text{ the standard deviation.}$$

Poisson distribution. *This is a probability distribution for a discrete random variable given by the formula:*

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \text{ where } \lambda \text{ is the mean.}$$

Binomial distribution. *This is a probability distribution for a discrete random variable given by the formula:*

$$f(x) = \binom{n}{x} p^x q^{n-x}, \text{ where } n \text{ is the number of trials, } p \text{ the probability of success and } q \text{ the}$$

probability of failure.

ROC – AUC. *Receiver Operating Curve Area Under the Curve.*

EEG. *Electro-encephalogram*

CHAPTER 2

This chapter presents a review of related literature in the field of comparing different methods of classification. A survey of the different types of classification techniques that have been compared is introduced. It can be noticed that different studies compared different sets of methods. This shows the importance of being able to choose the appropriate method for a particular classification task. Also, related studies have differed in the types of performance metrics used for comparison, so in this chapter, these different choices are analyzed. Also, the extent to which the data sets varied in terms of normality and equality of covariance matrices is analyzed to determine if the effects of these two properties were considered in the studies.

It is natural to investigate how the performances of the different methods were compared in related studies. An analysis of the different methodologies used is presented and arguments will be made concerning the strengths and weaknesses of the different approaches. The result of this analysis is to provide a good reason why this study uses an empirical approach to evaluate statistically whether there is a significant difference in performance across the methods.

The chapter ends with a summary of the results provided by the different studies, including how they compare to each other.

CLASSIFICATION TECHNIQUES INVOLVED

Related studies have compared different classification techniques. Some studies, as in (Iain and Mues 2012) compared many different techniques like Logistic regression, Neural networks, LDA, QDA, and Support Vector Machines, while others like (Laurinda et al. 2017) compared only LDA and QDA. In most of the reviewed studies, at least two different techniques have been compared. The only exception is in (Jing-Hao and Titterington 2007) where the performances of different LDA-based models were

compared. This shows how much research has been put into analyzing the performances of different techniques, indicating the importance of using the most suitable technique in practice.

Though some studies compared more than 3 different techniques, this study compares only LDA and QDA. Further studies could include more techniques for comparison, but the objective of this study is to investigate how LDA and QDA behave under certain imposed constraints.

VARIATION IN NORMALITY AND COVARIANCE

Normality and the equality of covariance matrices are important effects that are considered in this study. Related literature shows little variation in terms of these two properties. In (Iain and Mues 2012), given the variation in class sample sizes, different classes could seem to have different covariance matrices since the sample covariance is dependent on the sample size. Nevertheless, it is not clear to what extent the classes varied in covariance matrices. Also, in (Iain and Mues 2012), different financial datasets were used, which could give the impression that there was a variation of normality. Most of the datasets had a mixture of categorical and continuous variables, which would indicate non-normally distributed data, so there was little or no variation in terms of normality. Similarly, in (Subhrangsu and Tibarewala 2012), it is not clear whether the different classes varied in covariance matrices since what was studied was the effect of different feature combinations on classification performance. Although different feature combinations could yield classes with different covariance matrices there was no indication of a variation in covariance matrices from class to class.

On the other hand, different studies used single datasets for comparing classification performance, such as in (Laurinda et al. 2017), (Haoyuan et al. 2017), (Mahmodi, Mostafaei, and Mirzaee-Ghaleh 2019), and (Karami, Rasekh, and Mirzaee-Ghaleh 2020). It can be argued that because of these single datasets, there can be no variation in terms of normality and class covariance.

The absence of adequate variation in terms of these properties requires an investigation as to whether the performance of one method was influenced negatively or positively due to whether the data used for classification was normally distributed or not. The same applies to the issue of equality of covariance matrices across the different target classes. For this reason, this study investigates the effects of these two properties on classification performance. ANOVA techniques are used to investigate whether the interaction between normality and equality of covariance has a significant effect on classification performance. Also, by using a dependent sample T-test, more information can be obtained as to whether one classification method was better than the other in classifying data with given normality and covariance properties.

CHOICE OF PERFORMANCE METRIC

Different types of performance metrics have been used in related studies including the ROC-AUC, the percentage of correct classification rate, sensitivity, and specificity. So why this variety of metric types? Which one is better? It will seem that the different metric types have their strengths and weaknesses. Based on (Iain and Mues 2012), the ROC-AUC does not include information about class distributions and error costs. It will be preferable to have a metric that is not influenced by class distributions since, in this study, the effect of class distribution on performance is not considered. On the other hand, not considering classification errors can provide metrics that do not portray the true effectiveness of the model. In this light, other studies like (Laurinda et al. 2017), (Jing-Hao and Titterington 2007), (Mahmodi, Mostafaei, and Mirzaee-Ghaleh 2019), and (Karami, Rasekh, and Mirzaee-Ghaleh 2020), used a mix of different metric types. This practice of using different metrics types provides different views of a particular model in cases where a model can have a high ROC-AUC value but also a high error rate. In such cases, only looking at the ROC – AUC may be misleading. In contrast to studies that used more than one metric type, this current study used only the ROC – AUC. This is simpler

than using a set of different metric types and still provides a good indication of how well a method classifies.

METHODOLOGY

Repeated sampling usually provides better estimates of a population parameter or at least provides more confidence concerning that parameter over using just a single sample. Most of the reviewed studies did not use the technique of repeated sampling to provide better estimates of the difference in performance of the different techniques. On the other hand, in (Iain and Mues 2012) and (Jing-Hao and Titterington 2007), statistical tests were used to compare the performance of different methods using metrics collected from different samples. (Iain and Mues 2012) used Friedman's average rank test while (Jing-Hao and Titterington 2007) used the Wilcoxon signed rank test. This should provide better information concerning whether there was a significant difference in performance or not.

The current study follows a similar approach to those used in (Iain and Mues 2012) and (Jing-Hao and Titterington 2007) but instead used a dependent sample T-test to compare the performance of LDA and QDA. Furthermore, ANOVA is used in this study to determine whether the properties of normality and equality of covariance matrices are significant effects in explaining the variation in performance for each classification technique. This should provide even more information concerning which method is more influenced by these properties than the other method.

RESULTS OF PERFORMANCE COMPARISON

Results varied across the different studies reviewed but most seem to indicate that LDA and QDA differ in performance. For most of the studies, QDA showed superior classification performance than LDA. For instance, in (Subhrangsu and Tibarewala 2012), (Laurinda et al. 2017), (Mahmodi, Mostafaei, and Mirzaee-Ghaleh 2019), (Jie Liu et al. 2024), and (Karami, Rasekh, and Mirzaee-Ghaleh

2020), results indicated that QDA outperformed LDA. On the other hand, (Iain and Mues 2012) showed that LDA was significantly better than QDA, a result that was like the one obtained in (Haoyuan et al. 2017). However, in (Haoyuan et al. 2017), statistical tests were not involved and there was only a slight difference in performance between LDA and QDA.

CONCLUSION

Since many related studies have compared the performance of different classification techniques, it can be argued that it is important to compare the performance of different classification techniques. Most of the reviewed literature showed some maturity in the methodology by using an empirical approach. Although most of the studies did not use hypothesis testing to verify whether there was a significant difference in performance between the compared techniques, some used hypothesis testing, and some results showed a significant difference in performance across the samples involved.

However, none of the related studies investigated the relationship between normality and equality of covariance matrices and whether one technique is better due to normality and equality of covariance matrices.

This study continues the approach of using repeated sampling and hypothesis testing to verify whether there is a significant difference in performance between QDA and LDA and goes further to determine whether normality and equality of covariance matrices are significant effects in explaining the variation of performance for each technique. Therefore, this study should provide more knowledge concerning the impact of normality and equality of covariance matrices on classification performance.

CHAPTER 3

INTRODUCTION

This study aims to compare the performances of LDA and QDA when classifying data with different properties of normality and equality of covariances. In this chapter, the processes used to gather data and perform data analysis will be described. The target population will also be introduced including the specific data generated to represent this population. Also, the experimental setup will be presented, where the procedures used for data generation will be described precisely. The different datasets used for classification and those used to compare prediction performances will also be presented. Because this study uses statistical tests for data analysis, information will be provided concerning these statistical techniques and their reliability based on statistical power.

SETTING AND PARTICIPANTS

The data which is used for this study represents online stores' customers. More precisely, the data concerns customers' purchasing habits for two different lines of business: fashion and electronics. The goal is to determine how to better classify these customers based on their purchasing habits and reported tastes.

However, the data used is purely random and was generated using R. The details about how this data was generated will be presented and an indication of how this setup represents a real-world scenario will also be provided.

Also, some demographic data is generated for this population like age and gender. This information is not used during data analysis and provides just contextual information concerning the customers. This information can be used in further research by determining how customers' spending

habits vary by age or gender or how they can be classified into different age groups based on their spending habits.

INSTRUMENTATION

In this study, two categories of datasets were used. The first category of datasets involved randomly generated customer data drawn from some probability distribution. The probability distributions accounted for were the Normal Distribution, the Poisson Distribution, and the Binomial Distribution. R has functions for generating data from these distributions, but the difficulty faced by the authors was generating data using a given covariance structure. The approach used was to generate data in which the variables were correlated in a well-defined fashion. This was used to ensure that the data from different classes could have different covariance structures using different correlation relationships.

The relationship between covariance and correlation is given below.

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

where $\text{cor}(X, Y)$ is the correlation between X and Y, $\text{cov}(X, Y)$ is the covariance between X and Y, σ_X is the standard deviation of X and σ_Y the standard deviation of Y. From this expression, we see that if the standard deviations of X and Y are constant from one class to another, by imposing the same correlation structure, we can argue that we have approximately the same covariance across the classes.

Data that was Normally distributed or that came from a Poisson distribution was generated using the *simstudy* R package. This package enables the generation of correlated data from a multivariate Normal or Poisson distribution with a given correlation matrix. For data which was drawn from a multivariate Binomial distribution, the *copula* R package was used. It would have been desirable

to use only *simstudy* for data generation, but as of the writing of this article, the current version of *simstudy* does not support the generation of correlated data from a Binomial distribution. However, it does support the generation of Bernoulli correlated variables, but the complexity of managing many Bernoulli variables to represent a Binomial random variable was the main reason for using copulas. Based on the article:

(Wikipedia, March 23, 2024, [https://en.wikipedia.org/wiki/Copula_\(probability_theory\)](https://en.wikipedia.org/wiki/Copula_(probability_theory))), mention is made of Sklar's theorem which indicates that general joint probability distributions can be expressed using marginal probability distributions and copulas. Copulas are used to represent the dependence between the different random variables. So, in this study, copulas were used to generate randomly correlated data from a joint Binomial Distribution.

The second category of datasets was created by running QDA and LDA on the first-level datasets. These second-level datasets contained the ROC-AUC values for LDA and QDA which represented the performances of QDA and LDA when classifying the first-level datasets. Also, for these datasets, two additional variables were defined to record whether the dataset contained normally distributed data or not, and whether the classes had different covariance matrices or not.

PROCEDURE

The different algorithms used to generate the datasets will be presented in this section. The R source code used to implement these algorithms is referenced in the Appendix. For all the first-level datasets, the variables CustId, Age, and Gender are generated similarly. The CustId variable is simply a sequence of consecutive integers from 1 to 400. The Age variable is generated randomly using a Uniform distribution within the range 20 to 45 years. The Gender variable is also generated randomly using a Uniform distribution in the range 0 to 1, after which the random values are converted to the text "Male" for 0 and "Female" for 1.

The target classes used for classification are based on customers' tastes for two different lines of business: Fashion and Electronics. These tastes are represented by the variables LikesFashion and LikesElectronics. In order not to have perfect classification, a variation of tastes was applied while generating data for the 4 different types of first-level data. The years concerned with this data are the years 2022 and 2023. Below, is a description of how these different datasets were generated.

a.) *Normally distributed data with equal covariance matrices.*

A unique correlation matrix was used to generate data for the target classes represented by the variables LikesElectronics and LikesFashion. 350 different datasets of this form were created using the same correlation matrix. For each dataset, it was assumed that 70% of the customers preferred electronics over fashion, and the remaining 30% preferred fashion over electronics. This 70/30 split was used to create the target class of customers who could be predicted to prefer electronics over fashion. For this class, the average amount spent on electronics for 2022 and 2023 was about 100\$ while the average spent on fashion for the same period was about 50\$. A standard deviation of 1 was used for all the amounts of purchases. Similarly, for the class of customers who could be predicted to prefer fashion over electronics, 70% of the customers preferred fashion over electronics while 30% preferred electronics over fashion. Contrary to the class of those who preferred electronics, the average amount spent on fashion was about 100\$ while the average amount spent on electronics was about 50\$ for the same period. The idea here is that customers who prefer electronics over fashion will be expected to spend more on electronics than on fashion while those who prefer fashion to electronics will be expected to spend more on fashion than on electronics.

b.) *Normally distributed data with unequal covariance matrices.*

The procedure to generate these datasets is like the one for the previous case. The only difference was that a different correlation matrix was used to generate data for the different

target classes. The same assumptions made for the previous case also apply to this category of data.

c.) *Non-normally distributed data with equal covariance matrices.*

The non-normal distributions used were the Poisson and Binomial distributions. Similar to the generation of normally distributed data, the *simstudy* R package was used to generate correlated Poisson random data. The same 70/30 split was used for randomly generating customer tastes. However, instead of generating 350 datasets as in the previous categories, only 175 datasets were generated using the Poisson distribution. These 175 datasets were combined during classification with 175 datasets generated for the Binomial distributed data with equal covariance matrices to generate classification performance metrics for 350 datasets within this category. The variables that were modeled using a Poisson distribution were NumPurchCloth2022 which represented the number of fashion items purchased in 2022, NumPurchCloth2023 which represented the number of fashion items purchased in 2023, NumPurchElect2022 which represented the number of electronic items purchased in 2022, and NumPurchElect2023 which represented the number of electronic items purchased in 2023. For customers who preferred electronics to fashion, an average of 10 electronic items was defined for such customers compared to an average of 5 fashion items for the same category of customers for the given years. For customers who preferred fashion to electronics an average of 10 fashion items was defined compared to an average of 5 electronic items for the same category of customers. For the Binomial distribution, the variables that were defined were NumWeeksPurchElect2022 which represented the number of weeks in 2022 where a customer purchased an electronic item, NumWeeksPurchElect2023 which represented the number of weeks in 2023 where the customer purchased an electronic item, NumWeeksPurchCloth2022

which represented the number of weeks in 2022 where the customer purchased a fashion item, and NumWeeksPurchCloth2023 which represented the number of weeks in 2022 where the customer purchased a fashion item. However, for the generation of correlated data which followed a multivariate Binomial distribution, copulas were used. For customers who preferred electronics to fashion, an average of about 35 weeks for the purchase of electronic items was defined compared to 7 weeks for the purchase of fashion items. This frequency distribution was reversed for the class of customers who preferred fashion to electronics.

d.) *Non-normally distributed data with unequal covariance matrices.*

This category of datasets was generated using a similar procedure as the previous category. The only difference was that for this category, different correlation matrices were used to generate the Poisson distributed data for the different target classes. Similarly, for the Binomial distributed data, a different correlation coefficient was used when generating data for the different target classes.

Finally, the second-level datasets, which contained the performance metrics after classifying data in each first-level dataset, were generated using the *lda* and *qda* functions in R. A training set of 80% of the cases was used to create the classification models, while the remaining 20% were used for model validation. More precisely, for each dataset, since there are two taste variables: LikesElectronics and LikesFashion, different models were used to classify customers for the two different tastes. The classification was based on the numerical variables described above in sections a.) to d.) as features used to predict the classes for each taste. The average of the ROC-AUC obtained when classifying customers for each taste was used as the performance metric when classifying data in the dataset using either QDA or LDA. The following tables provide examples of the first and second-level datasets.

Table 4: Normally distributed data. The amount expressed is in US Dollars (\$).

Amount of clothing purchased in 2022	Amount of clothing purchased in 2023	Amount of electronics purchased in 2022	Amount of electronics purchased in 2023	Prefers electronics	Prefers fashion
50.881	50.553	100.040	100.378	Yes	Yes
49.030	49.236	98.688	98.856	Yes	No

Table 5: Poisson distributed data.

Number of fashion items purchased in 2022	Number of fashion items purchased in 2023	Number of electronic items purchased in 2022	Number of electronic items purchased in 2023	Prefers electronics	Prefers fashion
4	4	7	7	Yes	Yes
2	1	8	9	Yes	No

Table 6: Data from a multivariate Binomial distribution. The number of weeks per year is assumed to be 48.

Number of weeks in which fashion items were purchased in 2022	Number of weeks in which fashion items were purchased in 2023	Number of weeks in which electronic items were purchased in 2022	Number of weeks in which electronic items were purchased in 2023	Prefers electronics	Prefers fashion
4	7	35	36	Yes	Yes
9	15	41	45	No	No

Table 7: Dataset produced after running LDA and QDA on a first-level dataset. The performance values are expressed as percentages.

DatasetId	RocAUCLDA	RocAUCQDA	Is from normally distributed data	Classes have equal covariance matrices
Dataset_1	67.547	64.931	Yes	Yes
Dataset_2	66.455	68.768	Yes	Yes

DATA PROCESSING AND ANALYSIS

For each of the four datasets obtained from classification, the mean and standard deviation of the performances for LDA and QDA will be reported including a chart to view the distribution of the performances. For the primary research question, dependent sample T-tests were applied on each of the four datasets obtained from classification. Priori power analysis was performed using G*Power, which revealed that an actual power of at least 95% was achieved. The sample size was also not an issue, since each dataset had 350 cases.

In contrast to (Iain and Mues 2012), where the statistical technique used was Friedman's average rank test, the current study uses a dependent sample T-test for comparing performances. Also, in most reviewed literature, the ROC-AUC metric was used as a performance metric to compare different classification techniques. This is the case in (Iain and Mues 2022), (Haoyuan, Naghibi, and Dashtpagerdi 2017), and (Jing-Hao and Titterington 2007). Also, this study follows the approach in (Iain and Mues 2012), where statistical inference through hypothesis testing was used to investigate performance differences.

For the secondary question which seeks to investigate the interaction effect of normality and equality of covariances on prediction performance for LDA and QDA, two separate ANOVA-based inferences were used, one for each classification method. The unique dataset used for this is a union of the four different datasets used to answer the primary research question. For this new dataset, the means and the standard deviations of prediction performances of the techniques will be reported for each normality/covariance relationship.

SUMMARY

In this chapter, a description of the methodology was presented, concerning the approach carried out to answer the research questions. The target population was described, including how data for that population was generated and analyzed. Each dataset used in the study is thoroughly explained, including how it is generated and the variables defined within it.

Also, the statistical techniques that will be used in data analysis were presented including the different statistics that will be used to describe the datasets under investigation. The next chapter will present the results of applying these methods during data analysis and interpretations that will provide information as to whether normality and equality of covariances truly affect the performances of these methods.

CHAPTER 4

This chapter presents the results of the study. The different research questions will be answered, and samples will be described statistically. Also, charts will be used to show how the performances compare. Detailed data analysis results will be shown, including p-values and where necessary, post-hocs.

INTRODUCTION

In this study, the performances of LDA and QDA have been compared when classifying data with various properties of normality and equality of covariance matrices. The data considered was a simulation of online stores' customer data. From the descriptive statistics provided for each sample, it can be seen which technique was better, but relying on such information does not reveal any relationship between the performances. For this reason, statistical tests were employed to examine if there was a significant relationship between the performance metrics for both methods.

Before going into the details of the data analysis, each sample will be described statistically using appropriate statistical measures. Charts will be used to provide additional information not provided by the descriptive statistics, enabling the reader to understand trends in the samples. After this, the results of the data analysis will be provided where significant relationships if any, between the variables under investigation will be revealed. The alpha threshold for statistical significance is $\alpha = .05$, and if necessary, the results of post-hocs will be provided to reveal more information about the relationships under investigation. This will be the case when reporting results from analyzing the interaction effect between normality and equality of covariances on prediction performance. The chapter concludes with a synthesis of the trends that are revealed by the results.

SAMPLES

Five samples will be described individually. The first four involve data concerning the primary research question which compares the performance of LDA and QDA, while the last sample is concerned with the secondary research question. All performance metrics are expressed in percentages. In all the samples under investigation, the variables that stored the performance metrics of LDA and QDA were *RocAUCLDA* and *RocAUCQDA* respectively. These two continuous variables were compared using the paired T-test technique to answer the main research question. Also, for the secondary research question, these same variables were the dependent variables for separate ANOVA tests used to determine if normality and equality of covariance matrices had a significant effect on prediction performance. The independent variables used with ANOVA were two categorical variables: *IsNormal*, which was used to indicate whether the dataset contained normally distributed data and *EqualCovariance* which was used to indicate whether the dataset had target classes with equal covariance matrices.

a.) Normally distributed data with equal covariance matrices

The table below describes this sample of performance metrics which was used in answering the research question as to whether LDA and QDA differed significantly in performance when classifying data that was normally distributed and had equal covariance metrics.

Table 8: Means and standard deviations for performance metrics when classifying normally distributed data with equal covariance matrices.

	Mean(%)	SD(%)
RocAUCLDA	69.817	3.57
RocAUCQDA	68.763	4.331

From this table, it is evident that LDA performed slightly better than QDA with a mean performance of 69.817% in classifying data that is normally distributed and within which target classes have equal covariance matrices. However, data analysis was performed to determine if this difference was significant. Also, it can be noticed that the performance of LDA varied less than the performance of QDA. This study does not aim to compare this variation, so the effect of this difference in variation is not considered. Nevertheless, there is a significant difference in the distribution of performance metrics for the two types of techniques as can be revealed in the following chart where the metrics for QDA have more outliers than those for LDA.

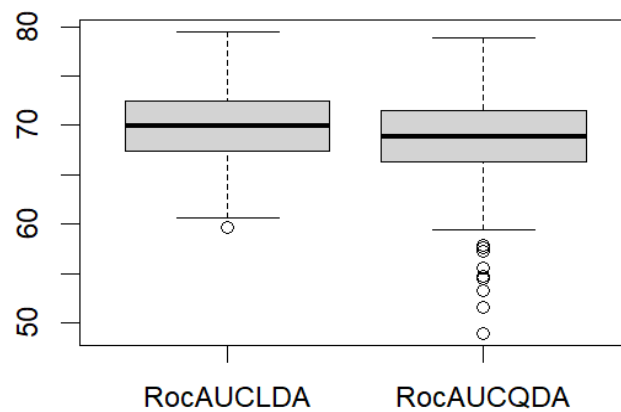


Figure 1. Box plot showing the distribution of the performance metrics for LDA and QDA when classifying normally distributed data with equal covariance matrices. The chart shows many outliers for QDA metrics.

Also, the causes for this difference in distribution are not analyzed in this study.

b.) Normally distributed data with unequal covariance metrics

For the second research question, which concerned normally distributed data with unequal covariance matrices, the gap between the performance metrics for LDA and QDA was smaller than for the previous case. The following table provides information gathered from this sample.

Table 9: Means and standard deviations for performance metrics when classifying normally distributed data with unequal covariance matrices.

	Mean(%)	SD(%)
RocAUCLDA	69.934	3.537
RocAUCQDA	69.212	3.797

The data seems to indicate that LDA and QDA performed quite similarly in performance with mean performances of 69.934% and 69.212% respectively, although LDA performed negligibly better. Also, the standard deviations of these metrics were quite close for both methods. When comparing the behaviour of QDA for this sample with its behaviour for the previous sample, a slight performance increase can be noticed. Could it mean that the equality of covariance matrices had a significant effect in explaining the variation of QDA performances? This particular question was answered using ANOVA on the sample made up of all performance metrics obtained during the study. Also, Figure 2 below shows how close the distributions of QDA and LDA performance metrics are, and a smaller number of outliers can be noticed for QDA metrics as opposed to the previous case.

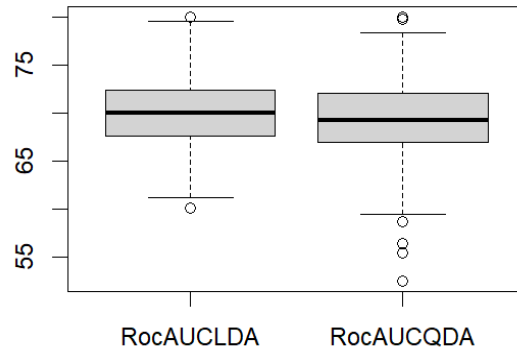


Figure 2. Box plot showing the distributions of the performance metrics for LDA and QDA when classifying normally distributed data with unequal covariance matrices. The chart shows fewer outliers for QDA metrics as opposed to the previous case.

c.) Non-normally distributed data with equal covariance matrices.

For this sample of performance metrics, a drop in classification performance can be noticed for LDA. The following table describes this sample.

Table 10: Means and standard deviations for performance metrics when classifying non-normally distributed data with equal covariance matrices.

	Mean(%)	SD(%)
RocAUCLDA	68.470	4.036
RocAUCQDA	67.345	4.810

When comparing these means with those obtained for the sample involving normally distributed data with equal covariance matrices, a slight drop in performance can be noticed for both techniques: From 69.817% to 68.470% on average for LDA and 68.763% to 67.345% on average for QDA. Similarly, there was an increase in the standard deviation for each type of performance metric: From 3.57 to 4.036 for LDA and 4.331 to 4.810 for QDA. So, it is natural to ask if normality had an influence on prediction performance. This question will also be

addressed using ANOVA. However as can be noticed from this sample, a superior performance was observed for LDA compared to QDA, and data analysis was used to investigate if this difference was significant. Also, the distributions of metrics for the two techniques differ especially due to a higher number of outliers for QDA performance metrics as shown in the following figure.

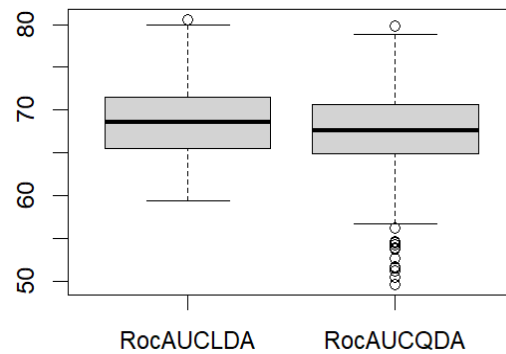


Figure 3. Box plot showing the distribution of the performance metrics for LDA and QDA when classifying non-normally distributed data with equal covariance matrices. The distributions are quite close, but QDA has many outliers.

d.) Non-normally distributed data with unequal covariance matrices.

For this sample, QDA showed the lowest mean performance of 66.668% across all samples. The following table summarizes this sample.

Table 11: Means and standard deviations for performance metrics when classifying non-normally distributed data with unequal covariance matrices.

	Mean(%)	SD(%)
RocAUCLDA	68.206	3.999
RocAUCQDA	66.668	5.095

As can also be seen in this table, QDA had the largest standard deviation of 5.095 across all 4 samples. What can explain this large variation? Although this study does not attempt to answer

this question, the secondary research question can provide indications as to what property influences this variation in performance. Comparing the distributions of performance metrics, the following figure shows a noticeable difference regarding the number of outliers in this sample for QDA.

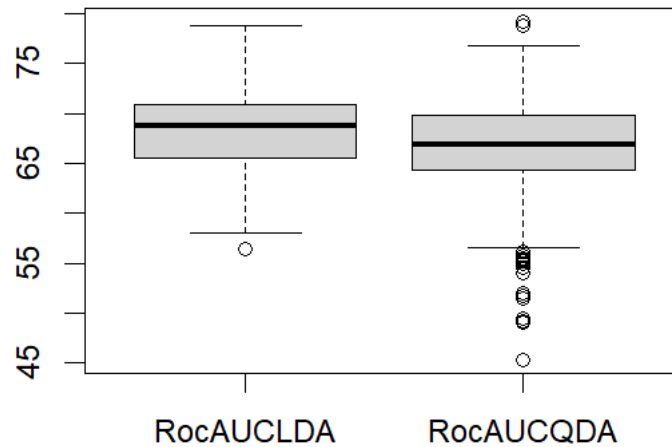


Figure 4. Box plot showing the distributions of the performance metrics for LDA and QDA when classifying non-normally distributed data with unequal covariance matrices. The distributions are quite close, but QDA has many outliers.

All four plots describing the distributions of metrics in this study, show a similar pattern concerning the metrics for QDA compared to LDA, which is the consistently high number of outliers for QDA and the relatively low number of outliers for LDA. For each sample, LDA has only 1 outlier, while QDA has many.

The description of the complete sample used to answer the secondary research question will be presented next. The following tables display the proportions and frequencies for the properties of normality and equality of covariance matrices.

Table 12: Counts and frequencies for the normality property within the complete dataset obtained by merging the previous 4 datasets.

	Proportion (%)	Frequency
Yes	50%	700
No	50%	700

Table 13: Counts and frequencies for the property of equality of covariance matrices within the complete dataset obtained by merging the previous 4 datasets.

	Proportion (%)	Frequency
Yes	50%	700
No	50%	700

As can be seen from the previous two tables, the samples are balanced in terms of the distribution of cases that are normally distributed and those that are not. A similar statement can be made for the property of the equality of covariance matrices. The following chart shows the distributions of performance metrics for LDA by the properties of normality and equality of covariance matrices.

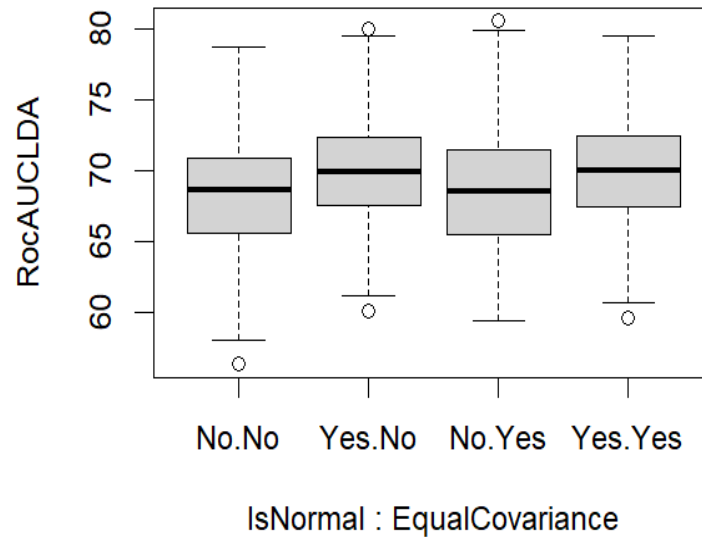


Figure 5. Box plot showing the distributions of the performance metrics for LDA by the properties of normality and equality of covariance matrices. LDA performs better when used with normally distributed data than with non-normally distributed data. The influence of equality of covariance matrices is not very noticeable since the performances do not seem to differ much when moving from a sample with equal covariance matrices to a sample with unequal covariance matrices.

For QDA, the following chart describes the distributions of performance metrics by the properties of normality and equality of covariance matrices, where the trend of having a consistently high number of outliers can be seen again.

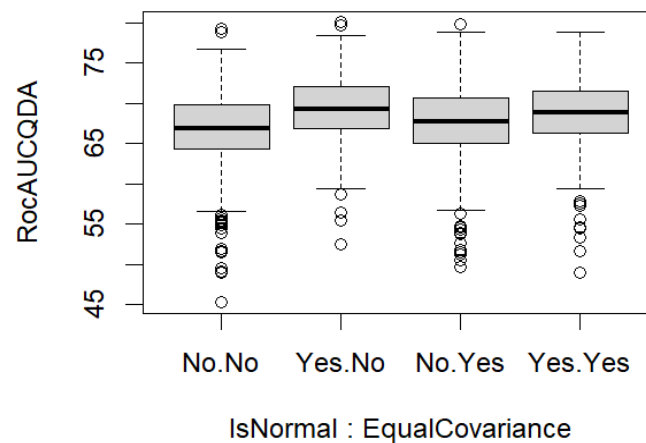


Figure 6. Box plot showing the distributions of the performance metrics for QDA by the properties of normality and equality of covariance matrices. QDA performs slightly better when used with normally distributed data with unequal covariance matrices than with normally distributed data with equal covariance matrices.

DATA ANALYSIS

In this section, the detailed results of the data analysis that was performed to answer the different research questions will be presented. Assumptions for the use of each statistical technique will be presented and information about whether statistical significance was obtained will also be provided. The α threshold used for determining statistical significance is $\alpha = .05$.

- a.) Is the performance of LDA significantly different from the performance of QDA when classifying normally distributed data with equal covariance matrices?

To answer this research question, a dependent sample T-test was used. Power analysis was performed for this sample and results indicated that power was not an issue since an actual power of 95% was achieved and required a minimum sample size of 55. A Shapiro-Wilk test for normality showed no statistical significance for the performance metrics for LDA [$W = 0.997$, $p = .866$] but showed statistical significance for the performance metrics for QDA [$W = 0.963$, $p < .0001$]. Nevertheless, the results of performing a dependent sample T-test revealed statistical significance [$t(349) = 9.275$, $p < .0001$] and a 95% CI of [0.8301738, 1.2770023]. These results indicate that there was a significant difference in performance between LDA and QDA and precisely that LDA significantly performed better than QDA.

- b.) Is the performance of LDA significantly different from the performance of QDA when classifying normally distributed data with unequal covariance matrices?

To answer this question, a dependent sample T-test was used. Power analysis was performed for this sample and results indicated that power was not an issue since an actual power of 95% was achieved and required a minimum sample size of 77. A Shapiro-Wilk test for normality showed no statistical significance for the LDA metrics [$W = .997$, $p = .934$] but showed statistical significance for the QDA metrics [$W = .985$, $p < .01$]. Nevertheless, results from running a

dependent sample T-test showed that statistical significance was obtained [$t(349) = 7.83$, $p < .0001$], 95% CI [0.5440579, 0.9087662]. So, for this research question, data seems to indicate that LDA performed significantly better than QDA.

- c.) Is the performance of LDA significantly different from the performance of QDA when classifying data that is non-normally distributed and has equal covariance matrices?

To answer this question, a dependent sample T-test was used. Power analysis was performed for this sample and results indicated that power was not an issue since an actual power of 95% was achieved and required a minimum sample size of 72. A Shapiro-Wilk test for normality showed no statistical significance for the LDA metrics [$W = .994$, $p = .142$] but showed statistical significance for the QDA metrics [$W = .964$, $p < .0001$]. Nevertheless, results from running a dependent sample T-test showed that statistical significance was obtained [$t(349) = 8.06$, $p < .0001$], 95% CI [0.8511524, 1.4005182]. So, for this research question, data seems to indicate that LDA performed significantly better than QDA.

- d.) Is the performance of LDA significantly different from the performance of QDA when classifying data that is non-normally distributed and has unequal covariance matrices?

To answer this question, a dependent sample T-test was used. Power analysis was performed for this sample and results indicated that power was not an issue since an actual power of 95% was achieved and required a minimum sample size of 61. A Shapiro-Wilk test for normality showed statistical significance for the LDA metrics [$W = .992$, $p = .047$] and for the QDA metrics [$W = .961$, $p < .0001$]. Nevertheless, results from running a dependent sample T-test showed that statistical significance was obtained [$t(349) = 8.81$, $p < .0001$], 95% CI [1.195092, 1.882473]. So, for this research question, data seems to indicate that LDA performed significantly better than QDA.

For the primary research question, we can confidently answer that data seems to indicate that LDA had superior performance when classifying data for all samples. Nevertheless, it is to be noted that since this study considers only the ROC - AUC performance metric, there may be additional information obtained from other metrics like the percentage of misclassifications. So, it will be desirable to include additional model performance metrics to have a better view of how these two techniques differ in efficiency.

The secondary research question which involves the analysis of the effect of normality and equality of covariance matrices was studied using ANOVA. Two-way ANOVA models were built to investigate the interaction effect of normality and equality of covariance matrices on LDA and QDA performances. A Shapiro-Wilk test for normality for the LDA-based model showed no statistical significance [$W = .998$, $p = .180$], which indicates that the assumption of normality was not an issue. However, for the QDA-based model, the Shapiro-Wilk test showed statistical significance [$W = .968$, $p < .0001$]. Also, Levene's test showed significance for the LDA-based model [$F(3) = 3.254$, $p = .021$], and also for the QDA-based model [$F(3) = 5.02$, $p < .01$]. Nevertheless, the LDA-based model showed that only the property of normality had a significant effect on the variation of prediction performance [$F(1) = 57.640$, $p < .0001$]. Running a Tukey post-hoc on a 1-way ANOVA model investigating the effect of normality on LDA's performance reveals the values shown in the following table:

Table 14: Tukey post-hocs for the effect of normality on LDA's performance.

Group	Lower CI Value	Upper CI Value
Yes-No	1.142	1.937

As the table above indicates, the performance of LDA significantly changes when moving from normally distributed data to non-normally distributed data.

For the QDA-based model, the interaction between normality and equality of covariance matrices had a significant effect on prediction performance [$F(1) = 5.398$, $p = .020$]. This will indicate that QDA was more sensitive to normality and equality of covariance matrices than LDA. A Tukey post-hoc on this model revealed the following results where only significant interactions are reported.

Table 15: Tukey post-hocs for the interaction effect of normality and equality of covariance matrices on QDA's performance.

Interaction Group	Lower CI Value	Upper CI Value
Yes:No-No:No	1.663	3.426
Yes:Yes-No:No	1.214	2.977
No:Yes-Yes:No	-2.749	-0.986
Yes:Yes-No:Yes	0.536	2.30

The table above indicates that normality and equality of covariance matrices truly affect the performance of QDA. In other words, it will seem that QDA takes into account these properties. For instance, the first group suggests that moving from data that is normally distributed and has unequal covariance matrices to data that is non-normally distributed and has unequal covariance matrices shows a significant drop in performance for QDA.

CONCLUSION

The results seem to be consistent in terms of which classification method performed better when classifying the datasets involved. Results indicated that LDA outperformed QDA during classification but also showed that LDA was insensitive to changes in the property of equality of covariance matrices. Does this imply that LDA is not well suited for such datasets? Possibly, but with only the ROC-AUC as a performance metric, this question cannot be answered in this current study. It

may be the case that other performance metrics like the classification error rate could reveal that LDA performs poorly for such datasets. On the other hand, QDA seems to be sensitive to changes in the property of equality of covariance matrices and may seem to be well suited for such datasets.

CHAPTER 5

INTRODUCTION

In this chapter, a synthesis of the study will be presented including a summary of findings and recommendations. The area of data classification has numerous benefits for businesses since many companies manipulate large amounts of data. But how should one choose a classifier like LDA over QDA? By comparing the performance of each classifier over numerous data samples, insight can be obtained concerning which classifier is better. Also, it should be noted that this study does not claim to be exhaustive regarding the type of performance metrics that were compared, so further research avenues will be described that can be used to improve the results of this study.

One important aspect of management is organization for strategic decision-making. In today's economy, the online store takes a very significant share in the number of business transactions made monthly, since many people shop online. Discriminant analysis can assist companies in making better decisions if they can predict categories of customers based solely on their purchasing data. In this chapter, recommendations will be made concerning how a company can implement a classification policy based on the type of customer data.

SUMMARY OF FINDINGS AND IMPLICATIONS

This study sought to answer the following questions:

- **When controlling for equal covariance matrices, does QDA perform better than LDA when classifying normally distributed data?**
- **When controlling for equal covariance matrices, does QDA perform better than LDA when classifying non-normally distributed data?**

- **Can the interaction between equal covariance matrices and normality explain the variation in classification performance between LDA and QDA?**

The results are very consistent concerning which classification method is better in terms of the ROC-AUC when classifying customer-related data. Data seems to indicate that LDA should be preferred when classifying such data as opposed to using QDA. This result is quite contrary to what is observed in most reviewed studies where QDA seems to be the better choice.

Also, the results indicate that QDA is more sensitive to changes in the distribution of data than LDA. More specifically, the properties of normality and equality of covariance matrices had a significant effect on prediction performance for QDA-based classification, while only normality had a significant effect on the prediction performance for LDA-based classification. This supports the idea that QDA is more suitable for scenarios where there is much variation in the types of distributions inherent in the data. Compared to reviewed studies that did not investigate the effect that normality and equality of covariance matrices have on prediction performance, the current study produces insight into how these properties affect prediction performance. For example, QDA behaved better when classifying normally distributed data with unequal covariance matrices than when classifying non-normally distributed data with unequal covariance matrices. It is this sensitivity that is interesting due to the statistical significance obtained.

The results of this study can assist companies operating in the online store business in making strategic decisions based on data. Questions like “Which of our customers love tech items?”, can be answered by the correct choice of tool. If the company has a dataset comprising annual purchases for each product type, then based on the results of this study, LDA should be used for such a purpose since it promises high classification rates.

RECOMMENDATIONS

Based on the results of the current study, the authors recommend that online store companies should implement data classification policies based on the nature of the distributions inherent in datasets. For data composed only of the amount spent on purchases for each type of product, LDA should be used for such datasets provided that the target classes can be assured not to vary much in their respective covariance matrices. For data made up of information like the number of items of each product type that was purchased in each period, QDA should be the preferred technique to use due to its sensitivity to changes in normality and equality of covariance matrices. The insight that grounds such a policy is based on the results of the data analysis used to investigate the secondary research question. Results indicated that normality and equality of covariance matrices did not have an interaction effect on the performance of LDA but had an interaction effect on the performance of QDA.

Also, the choice of the performance metric used poses some issues on the generalizability of the results of this study. The ROC-AUC is not the only performance metric used to evaluate models. It is desirable that other measures like the error rate should be incorporated and that LDA and QDA should also be compared as regards these other performance metrics. So, a possibility for future research will be to compare the error rates when using LDA and QDA-based models and how they correlate with the ROC-AUC. All this should still be done with an emphasis on the influence of normality and equality of covariance matrices. Another possible study that will make the results more generalizable is the inclusion of additional probability distributions like the Geometric distribution and the Gamma distribution.

CONCLUSION

In this study, the performances of LDA and QDA were compared under varying conditions of normality and equality of covariance matrices. The study's results strongly indicate the superior performance of LDA over QDA when classifying customer sales data but also reveal a lack of sensitivity of LDA when covariance matrices vary. With these results, online store businesses can be better guided concerning how to classify their customers based only on their spending habits efficiently. Nevertheless, certain issues were highlighted as regards future research where it is desirable to include more probability distributions and account for other performance metric types which will provide a better performance profile for a given technique.

REFERENCES

Jie Liu, Pan Wang, Hua Zhang, Nan Wu. Distinguishing brain tumors by Label-free confocal micro-Raman spectroscopy. 2024.

Hamed Karami, Mansour Rasekh, Esmail Mirzaee – Ghaleh. Comparison of chemometrics and AOCS official methods for predicting the shelf life of edible oil. 2020.

Korosh Mahmodi, Mostafa Mostafaei, Esmail Mirzaee-Ghaleh. Detection and classification of diesel-biodiesel blends by LDA, QDA and SVM approaches using an electronic nose. 2019.

Laurinda F.S. Siqueira, Raimundo F. Araújo Júnior, Aurigena Antunes de Araújo, Camilo L.M. Morais, Kássio M.G. Lima. LDA vs. QDA for FT-MIR prostate cancer tissue classification. 2017.

Haoyuan Hong, Seyed Amir Naghibi, Mostafa Moradi Dashtpajardi, Hamid Reza Pourghasemi, Wei Chen. A comparative assessment between linear and quadratic discriminant analyses (LDA-QDA) with frequency ratio and weights-of-evidence models for forest fire susceptibility mapping in China. 2017.

Iain Brown, Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. 2012

Subhrangsu Aditya, D.N. Tibarewala. Comparing ANN, LDA, QDA, KNN and SVM algorithms in classifying relaxed and stressful mental state from two-channel prefrontal EEG data. 2012.

Jing-Hao Xue, D. Michael Titterton. Do unbalanced data have a negative-effect on LDA. 2007.

APPENDIX A

The following sections illustrate the R code used to generate data for this study.

NORMALLY DISTRIBUTED DATA

The following piece of code was used to generate normally distributed data.

```
install.packages("simstudy");
install.packages("purrr");
library(purrr);
library(simstudy);

# Set the seed
a <- as.numeric(Sys.time());
set.seed(a);
help(genCorData);

A <- matrix(c(1.0, 0.9, 0.3, 0.3, 0.9, 1.0, 0.3, 0.3, 0.3, 0.3, 1.0, 0.9, 0.3, 0.3, 0.9, 1.0),
nrow = 4, ncol = 4, byrow = T);

for (i in 1:350){
  output_file_path <- paste(r"(C:\datasets\dataset_)", i, ".csv");

  # Class sample for cases who like electronic

  cust_id_cls_le <- c(1:200);

  age_cls_le <- floor(runif(200, min=20, max = 46));

  gender_cls_le <- map_vec(floor(runif(200, min=0, max = 2)), function(x){ if(x == 0) 'Male' else
'Female' });

  likes_electronics_cls_le <- map_vec(rbinom(n=200, size=1, prob=0.7), function(x){ if (x == 0)
'No' else 'Yes' });

  likes_fashion_cls_le <- map_vec(rbinom(n=200, size=1, prob=0.3), function(x){ if (x == 0) 'No'
else 'Yes' });

  dframe_le <- genCorData(200, mu = c(100, 100, 50, 50), sigma = c(1, 1, 1, 1), corMatrix = A,
cnames='amnt_purch_elect_2022,amnt_purch_elect_2023,amnt_purch_cloth_2022,amnt_purch_cloth_2023')
;

  group11_cls_le <- data.frame(CustId = cust_id_cls_le, Age = age_cls_le, Gender = gender_cls_le,
AmntPurchCloth2022 = dframe_le$amnt_purch_cloth_2022, AmntPurchCloth2023 =
dframe_le$amnt_purch_cloth_2023, AmntPurchElect2022 = dframe_le$amnt_purch_elect_2022,
AmntPurchElect2023 = dframe_le$amnt_purch_elect_2023, LikesElectronics =likes_electronics_cls_le,
LikesFashion = likes_fashion_cls_le);
```

```

# Class sample for cases who like fashion
cust_id_cls_lf <- c(201:400);

age_cls_lf <- floor(runif(200, min=20, max = 46));

gender_cls_lf <- map_vec(floor(runif(200, min=0, max = 2)), function(x){ if(x == 0) 'Male' else
'Female'});

dframe_lf <- genCorData(200, mu = c(50, 50, 100, 100), sigma = c(1, 1, 1, 1), corMatrix = A,
cnames='amnt_purch_elect_2022,amnt_purch_elect_2023,amnt_purch_cloth_2022,amnt_purch_cloth_2023';

likes_electonics_cls_lf <- map_vec(rbinom(n=200, size=1, prob=0.3), function(x){ if (x == 0)
'No' else 'Yes'});

likes_fashion_cls_lf <- map_vec(rbinom(n=200, size=1, prob=0.7), function(x){ if (x == 0) 'No'
else 'Yes'});

group11_cls_lf <- data.frame(CustId = cust_id_cls_lf, Age = age_cls_lf, Gender = gender_cls_lf,
AmntPurchCloth2022 = dframe_lf$amnt_purch_cloth_2022, AmntPurchCloth2023 =
dframe_lf$amnt_purch_cloth_2023, AmntPurchElect2022 = dframe_lf$amnt_purch_elect_2022,
AmntPurchElect2023 = dframe_lf$amnt_purch_elect_2023, LikesElectronics =likes_electonics_cls_lf,
LikesFashion = likes_fashion_cls_lf);

group11 <- rbind(group11_cls_le, group11_cls_lf);

write.csv(group11, output_file_path));
}

```

The variable *output_file_path* refers to the path on the local file system which will store the resulting datasets. The file path in this case assumes that this code is to be run on a computer running Microsoft Windows. In order to run it on a Unix-based system or on a Mac-based system, use the correct syntax for naming file paths.

In order to generate normally distributed data with unequal covariance matrices, all we need to do is define an additional correlation matrix to generate data for the data frame *dframe_lf*.

NON-NORMALLY DISTRIBUTED DATA WITH THE BINOMIAL DISTRIBUTION

The following piece of code is used to generate non-normally distributed data. This implementation creates data that follows a binomial distribution using copulas.

```

install.packages("purrr");

library(purrr);

library(copula);

# Set the seed

a <- as.numeric(Sys.time());

set.seed(a);

```

```

n_copula <- normalCopula(0.75, dim=4);
help(normalCopula);

# -- What follows is for the poisson data generation
help(qbinom);

for (i in 1:175){

  output_file_path <- paste(r"(C:\datasets\dataset_)", i, ".csv");

  # Class sample for cases who like electronic

  cust_id_cls_le <- c(1:200);

  age_cls_le <- floor(runif(200, min=20, max = 46));

  gender_cls_le <- map_vec(floor(runif(200, min=0, max = 2)), function(x){ if(x == 0) 'Male' else
'Female'}));

  likes_electronics_cls_le <- map_vec(rbinom(n=200, size=1, prob=0.7), function(x){ if (x == 0)
'No' else 'Yes'}));

  likes_fashion_cls_le <- map_vec(rbinom(n=200, size=1, prob=0.3), function(x){ if (x == 0) 'No'
else 'Yes'}));

  r_copula <- rCopula(200, n_copula);

  matrix_binomials_le <- cbind(qbinom(r_copula[,1], 48, 0.1), qbinom(r_copula[,2], 48, 0.2),
qbinom(r_copula[,3], 48, 0.7), qbinom(r_copula[,4], 48, 0.8));

  group21_cls_le <- data.frame(CustId = cust_id_cls_le, Age = age_cls_le, Gender = gender_cls_le,
NumWeeksPurchCloth2022 = matrix_binomials_le[,1], NumWeeksPurchCloth2023 =
matrix_binomials_le[,2], NumWeeksPurchElect2022 = matrix_binomials_le[,3], NumWeeksPurchElect2023
= matrix_binomials_le[,4], LikesElectronics = likes_electronics_cls_le, LikesFashion =
likes_fashion_cls_le);

  # Class sample for cases who like fashion

  cust_id_cls_lf <- c(201:400);

  age_cls_lf <- floor(runif(200, min=20, max = 46));

  gender_cls_lf <- map_vec(floor(runif(200, min=0, max = 2)), function(x){ if(x == 0) 'Male' else
'Female'}));

  matrix_binomials_lf <- cbind(qbinom(r_copula[,1], 48, 0.7), qbinom(r_copula[,2], 48, 0.8),
qbinom(r_copula[,3], 48, 0.1), qbinom(r_copula[,4], 48, 0.2));

  likes_electronics_cls_lf <- map_vec(rbinom(n=200, size=1, prob=0.3), function(x){ if (x == 0)
'No' else 'Yes'}));

  likes_fashion_cls_lf <- map_vec(rbinom(n=200, size=1, prob=0.7), function(x){ if (x == 0) 'No'
else 'Yes'}));

  group21_cls_lf <- data.frame(CustId = cust_id_cls_lf, Age = age_cls_lf, Gender = gender_cls_lf,
NumWeeksPurchCloth2022 = matrix_binomials_lf[,1], NumWeeksPurchCloth2023 =
matrix_binomials_lf[,2], NumWeeksPurchElect2022 = matrix_binomials_lf[,3], NumWeeksPurchElect2023
= matrix_binomials_lf[,4], LikesElectronics = likes_electronics_cls_lf, LikesFashion =
likes_fashion_cls_lf);

```



```

group21 <- rbind(group21_cls_le, group21_cls_lf);

write.csv(group21, output_file_path);

}

```

The previous code generates 175 datasets containing data that follows a Binomial distribution and has equal covariance matrices for the different target classes. To generate data that have unequal covariance matrices and follow the same distribution, we only need to use a different copula with a different correlation coefficient, when generating data in the data frame *matrix_binomials_lf*.

NON-NORMALLY DISTRIBUTED DATA WITH THE POISSON DISTRIBUTION

The following piece of code generates data that follows a Poisson distribution.

```

install.packages("simstudy");

install.packages("purrr");

library(purrr);

library(simstudy);

# Set the seed

a <- as.numeric(Sys.time());

set.seed(a);

A <- matrix(c(1.0, 0.9, 0.3, 0.3, 0.9, 1.0, 0.3, 0.3, 0.3, 0.3, 1.0, 0.9, 0.3, 0.3, 0.9, 1.0),
nrow = 4, ncol = 4, byrow = T);

# -- what follows is for the poisson data generation

for (i in 1:175){

  output_file_path <- paste(r"(C:\datasets\dataset_)", i, ".csv");

```

```

# Class sample for cases who like electronic

cust_id_cls_le <- c(1:200);

age_cls_le <- floor(runif(200, min=20, max = 46));

gender_cls_le <- map_vec(floor(runif(200, min=0, max = 2)), function(x){ if(x == 0) 'Male' else
'Female'}));

likes_electronics_cls_le <- map_vec(rbinom(n=200, size=1, prob=0.7), function(x){ if (x == 0)
'No' else 'Yes'}));

likes_fashion_cls_le <- map_vec(rbinom(n=200, size=1, prob=0.3), function(x){ if (x == 0) 'No'
else 'Yes'}));

dframe_le <- genCorGen(200, nvars = 4, params1 = c(10, 10, 5, 5), wide = TRUE, cnames =
"num_purch_elect_2022,num_purch_elect_2023,num_purch_cloth_2022,num_purch_cloth_2023", dist =
"poisson", corMatrix = A);

group21_cls_le <- data.frame(CustId = cust_id_cls_le, Age = age_cls_le, Gender = gender_cls_le,
NumPurchCloth2022 = dframe_le$num_purch_cloth_2022, NumPurchCloth2023 =
dframe_le$num_purch_cloth_2023, NumPurchElect2022 = dframe_le$num_purch_elect_2022,
NumPurchElect2023 = dframe_le$num_purch_elect_2023, LikesElectronics = likes_electronics_cls_le,
LikesFashion = likes_fashion_cls_le);

# Class sample for cases who like fashion

cust_id_cls_lf <- c(201:400);

age_cls_lf <- floor(runif(200, min=20, max = 46));

gender_cls_lf <- map_vec(floor(runif(200, min=0, max = 2)), function(x){ if(x == 0) 'Male' else
'Female'}));

dframe_lf <- genCorGen(200, nvars = 4, params1 = c(5, 5, 10, 10), wide = TRUE, cnames =
"num_purch_elect_2022,num_purch_elect_2023,num_purch_cloth_2022,num_purch_cloth_2023",
dist = "poisson", corMatrix = A);

likes_electronics_cls_lf <- map_vec(rbinom(n=200, size=1, prob=0.3), function(x){ if (x == 0)
'No' else 'Yes'}));

likes_fashion_cls_lf <- map_vec(rbinom(n=200, size=1, prob=0.7), function(x){ if (x == 0) 'No'
else 'Yes'}));

```

```

    group21_cls_lf <- data.frame(CustId = cust_id_cls_lf, Age = age_cls_lf, Gender = gender_cls_lf,
    NumPurchCloth2022 = dframe_lf$num_purch_cloth_2022, NumPurchCloth2023 =
    dframe_lf$num_purch_cloth_2023, NumPurchElect2022 = dframe_lf$num_purch_elect_2022,
    NumPurchElect2023 = dframe_lf$num_purch_elect_2023, LikesElectronics = likes_electronics_cls_lf,
    LikesFashion = likes_fashion_cls_lf);

    group21 <- rbind(group21_cls_le, group21_cls_lf);

    write.csv(group21, output_file_path);

}

```

To generate data with unequal covariance matrices, we only need to use a different correlation matrix when generating data in the data frame *dframe_lf*. Also, it is desirable that the datasets generated do not overwrite previously generated datasets. So different directories should be created in the file system to represent the four groups of data.

The previous sections of code were only involved in generating first-level datasets. The following section shows the code used to generate second-level datasets.

GENERATING METRICS FOR NORMALLY DISTRIBUTED DATA

The following piece of code shows how classification performance metrics were obtained when classifying normally distributed data.

```

library(MASS);
library(pROC);

a <- as.numeric(Sys.time());
set.seed(a);

result <- data.frame(c(), c(), c(), c(), c());

output_file_path <- r"(C:\datasets\metrics\dataset_11.csv)";

for (i in 1:350){

    rm(AmntPurchCloth2022, AmntPurchCloth2023, AmntPurchElect2022, AmntPurchElect2023,
    LikesElectronics, LikesFashion);

    input_file_path <- paste(r"(C:\datasets\dataset_)", i, ".csv");

    group11 <- read.csv(input_file_path)

    n <- floor(.80*nrow(group11));

    train_row_indices <- sample(seq_len(nrow(group11)), size=n)

    train_dataset <- group11[train_row_indices,];

    test_dataset <- group11[-train_row_indices,];
}

```

```

attach(train_dataset);

factor_train_le <- as.ordered(factor(LikesElectronics));
factor_train_lf <- as.ordered(factor(LikesFashion));

l_model_le <- lda(factor_train_le ~ AmntPurchCloth2022 + AmntPurchCloth2023 +
AmntPurchElect2022 + AmntPurchElect2023, data=train_dataset);

l_model_lf <- lda(factor_train_lf ~ AmntPurchCloth2022 + AmntPurchCloth2023 +
AmntPurchElect2022 + AmntPurchElect2023, data=train_dataset);

predict_lda_le <- predict(l_model_le, newdata=test_dataset);
predict_lda_lf <- predict(l_model_lf, newdata=test_dataset);

q_model_le <- qda(factor_train_le ~ AmntPurchCloth2022 + AmntPurchCloth2023 +
AmntPurchElect2022 + AmntPurchElect2023, data=train_dataset);

q_model_lf <- qda(factor_train_lf ~ AmntPurchCloth2022 + AmntPurchCloth2023 +
AmntPurchElect2022 + AmntPurchElect2023, data=train_dataset);

predict_qda_le <- predict(q_model_le, newdata=test_dataset);
predict_qda_lf <- predict(q_model_lf, newdata=test_dataset);

factor_test_le <- as.ordered(factor(test_dataset$LikesElectronics));
factor_test_lf <- as.ordered(factor(test_dataset$LikesFashion));

roc_l_le <- roc(factor_test_le, as.ordered(predict_lda_le$class), percent=TRUE);
roc_l_lf <- roc(factor_test_lf, as.ordered(predict_lda_lf$class), percent=TRUE);
roc_q_le <- roc(factor_test_le, as.ordered(predict_qda_le$class), percent=TRUE);
roc_q_lf <- roc(factor_test_lf, as.ordered(predict_qda_lf$class), percent=TRUE);

auc_l_le <- auc(roc_l_le);
auc_l_lf <- auc(roc_l_lf);
auc_q_le <- auc(roc_q_le);
auc_q_lf <- auc(roc_q_lf);

result <- rbind(result, c(paste('Dataset_', i), (auc_l_le + auc_l_lf)/2, (auc_q_le +
auc_q_lf)/2, 'Yes', 'Yes')));
}

names(result) <- c('DataSetId', 'RocAUCLDA', 'RocAUCQDA', 'IsNormal', 'EqualCovariance');
write.csv(result, output_file_path);

```

Code to generate the three other second-level datasets is quite similar to the previous code. For normally distributed data with unequal covariance matrices, the input path should point to the folder containing this category of datasets. The independent variables for the LDA and QDA models are the same. We will also need to change the value for the output file path and set the value of the column

EqualCovariance to *No*. For non-normally distributed data, the value of the column *IsNormal* will be *Yes*, and the value of the column *EqualCovariance* will be set depending on the category of datasets. Also, for non-normally distributed data, the independent variables for the LDA and QDA models are columns in the corresponding first-level datasets whose probability distributions have been specified as either Binomial or Poisson. However, separate loops were used for the Binomial case and the Poisson case so that metrics were generated for a particular type of distribution. The cases were then merged into a single dataset for that particular category. The following piece of code shows this.

```
library(MASS);
library(pROC);
a <- as.numeric(Sys.time());
set.seed(a);
output_file_path <- r"(C:\datasets\metrics\dataset_11.csv)";
result <- data.frame(c(), c(), c(), c(), c());
for (i in 1:175){
  input_file_path <- paste(r"(C:\datasets\poisson\dataset_)", i, ".csv");
  group21 <- read.csv(input_file_path)
  n <- floor(.80*nrow(group21));
  train_row_indices <- sample(seq_len(nrow(group21)), size=n)
  train_dataset <- group21[train_row_indices,];
  test_dataset <- group21[-train_row_indices,];
  attach(train_dataset);
  factor_train_le <- as.ordered(factor(LikesElectronics));
  factor_train_lf <- as.ordered(factor(LikesFashion));
  l_model_le <- lda(factor_train_le ~ NumPurchCloth2022 + NumPurchCloth2023 + NumPurchElect2022 +
NumPurchElect2023, data=train_dataset);
  l_model_lf <- lda(factor_train_lf ~ NumPurchCloth2022 + NumPurchCloth2023 + NumPurchElect2022 +
NumPurchElect2023, data=train_dataset);
  predict_lda_le <- predict(l_model_le, newdata=test_dataset);
  predict_lda_lf <- predict(l_model_lf, newdata=test_dataset);
  q_model_le <- qda(factor_train_le ~ NumPurchCloth2022 + NumPurchCloth2023 + NumPurchElect2022 +
NumPurchElect2023, data=train_dataset);
```

```

q_model_1f <- qda(factor_train_1f ~ NumPurchCloth2022 + NumPurchCloth2023 + NumPurchElect2022 +
NumPurchElect2023, data=train_dataset);

predict_qda_1e <- predict(q_model_1e, newdata=test_dataset);
predict_qda_1f <- predict(q_model_1f, newdata=test_dataset);
factor_test_1e <- as.ordered(factor(test_dataset$LikesElectronics));
factor_test_1f <- as.ordered(factor(test_dataset$LikesFashion));
roc_l_1e <- roc(factor_test_1e, as.ordered(predict_lda_1e$class), percent=TRUE);
roc_l_1f <- roc(factor_test_1f, as.ordered(predict_lda_1f$class), percent=TRUE);
roc_q_1e <- roc(factor_test_1e, as.ordered(predict_qda_1e$class), percent=TRUE);
roc_q_1f <- roc(factor_test_1f, as.ordered(predict_qda_1f$class), percent=TRUE);
auc_l_1e <- auc(roc_l_1e);
auc_l_1f <- auc(roc_l_1f);
auc_q_1e <- auc(roc_q_1e);
auc_q_1f <- auc(roc_q_1f);

result <- rbind(result, c(paste('Dataset_', i), (auc_l_1e + auc_l_1f)/2, (auc_q_1e +
auc_q_1f)/2, 'No', 'Yes'));
}

for (i in 1:175){
input_file_path <- paste(r"(C:\datasets\binom\dataset_)", i, ".csv");
group21 <- read.csv(input_file_path)

n <- floor(.80*nrow(group21));
train_row_indices <- sample(seq_len(nrow(group21)), size=n)
train_dataset <- group21[train_row_indices,];
test_dataset <- group21[-train_row_indices,];
attach(train_dataset);

factor_train_1e <- as.ordered(factor(LikesElectronics));
factor_train_1f <- as.ordered(factor(LikesFashion));

l_model_1e <- lda(factor_train_1e ~ NumWeeksPurchCloth2022 + NumWeeksPurchCloth2023 +
NumWeeksPurchElect2022 + NumWeeksPurchElect2023, data=train_dataset);

l_model_1f <- lda(factor_train_1f ~ NumWeeksPurchCloth2022 + NumWeeksPurchCloth2023 +
NumWeeksPurchElect2022 + NumWeeksPurchElect2023, data=train_dataset);

predict_lda_1e <- predict(l_model_1e, newdata=test_dataset);
predict_lda_1f <- predict(l_model_1f, newdata=test_dataset);

q_model_1e <- qda(factor_train_1e ~ NumWeeksPurchCloth2022 + NumWeeksPurchCloth2023 +
NumWeeksPurchElect2022 + NumWeeksPurchElect2023, data=train_dataset);

```

```

q_model_1f <- qda(factor_train_1f ~ NumWeeksPurchCloth2022 + NumWeeksPurchCloth2023 +
NumWeeksPurchElect2022 + NumWeeksPurchElect2023, data=train_dataset);

predict_qda_1e <- predict(q_model_1e, newdata=test_dataset);
predict_qda_1f <- predict(q_model_1f, newdata=test_dataset);
factor_test_1e <- as.ordered(factor(test_dataset$LikesElectronics));
factor_test_1f <- as.ordered(factor(test_dataset$LikesFashion));
roc_l_1e <- roc(factor_test_1e, as.ordered(predict_lda_1e$class), percent=TRUE);
roc_l_1f <- roc(factor_test_1f, as.ordered(predict_lda_1f$class), percent=TRUE);
roc_q_1e <- roc(factor_test_1e, as.ordered(predict_qda_1e$class), percent=TRUE);
roc_q_1f <- roc(factor_test_1f, as.ordered(predict_qda_1f$class), percent=TRUE);
auc_l_1e <- auc(roc_l_1e);
auc_l_1f <- auc(roc_l_1f);
auc_q_1e <- auc(roc_q_1e);
auc_q_1f <- auc(roc_q_1f);

result <- rbind(result, c(paste('Dataset_', 175 + i), (auc_l_1e + auc_l_1f)/2, (auc_q_1e +
auc_q_1f)/2, 'No', 'Yes'));
}

names(result) <- c('DataSetId', 'RocAUCLDA', 'RocAUCQDA', 'IsNormal', 'EqualCovariance');
write.csv(result, output_file_path);

```

The file paths are just for the purpose of illustration but they should be designed to follow an organization corresponding to the first and second level of datasets, and for the different non-normally distributions, separate directories should be used for the first-level datasets.

BIBLIOGRAPHY¹

BATE – EYA AYUK EGBE

Candidate for the Degree of

Master of Science Mathematics

Thesis: EMPIRICALLY COMPARING THE PERFORMANCE OF LDA AND QDA WHEN CLASSIFYING CUSTOMER SALES DATA WITH DIFFERENT PROPERTIES OF NORMALITY AND EQUALITY OF COVARIANCE MATRICES

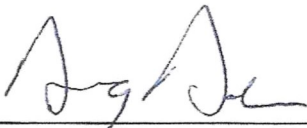
Major Field: Mathematics

Biographical: Bate-Eya Ayuk Egbe is an MSc Mathematics student at Shawnee State University. He received a BSc degree in Computer Science from the University of Yaounde I, Cameroon and an MSc degree in Computer Science from the University of Franche-Comte, France. He is an active senior software engineer who is interested in the integration of machine learning algorithms and artificial intelligence techniques for generating insight from data.

Personal Data: Email(jasonjava2003@hotmail.com), Phone: +1 4184733192

Education: BSc in Computer Science, MSc in Computer Science.

Completed the requirements for the Master of Science in Mathematics at Shawnee State University, Ohio in July 2024.



ADVISER'S APPROVAL: Dr. DOUGLAS G. DARBRO

7/22/2024