Applied Multivariate Analysis

Project #2 - Classification

Omar Moustafa

900222400

Seifeldin Abdelhamid

900222395

MACT 4233

Spring 2025

Introduction

Complete Source of the Data:

The data set that will be used and analyzed is derived from the well-known Titanic dataset, which contains detailed information about the passengers aboard the RMS Titanic. It includes several personal features such as ticket class and fare, age, number of family members aboard, and survival statuses. Building on this, this data set provides valuable insights into the existing relationships between passenger demographics, ticket class, fare, and their survival outcomes, portraying which groups were more likely than others to survive the 1912 disaster. Furthermore, it is also a useful dataset for performing classification analysis and assessing model accuracy because it contains quantitative and categorical variables.

What Are the Main Objectives?

The primary objective of this study is to use multiple classification techniques to predict the passenger class ("Pclass") of Titanic passengers based on available demographic and socioeconomic data. The three classes within this categorical variable — 1st Class, 2nd Class, and 3rd Class — are represented by 1, 2, and 3, respectively. Specifically, this analysis seeks to:

1. Compare and contrast two highly prominent classification methods: Fisher's Linear Discriminant Analysis (FLDA) and Multinomial Logistic Regression.
2. Evaluate and compare the Internal and External Validation performance of these two classification models.
3. Apply FLDA2 projection analysis in order to visualize and assess class separability by reducing the dimensionality of the data, from three to two dimensions, while also maintaining the differences between the passenger classes.

Description of the Data

What Are The Observations?

The dataset comprises 887 observations, each representing a passenger aboard the RMS Titanic ship. It contains multiple different demographic and socioeconomic attributes, allowing for a thorough analysis and understanding of patterns among different passenger groups, which are first, second, and third-class passengers. By examining these observations and testing the accuracy of multiple notable classification methods, valuable insights can be derived about the distribution and classifications of passengers and the factors that could have played a role in their fates during the traumatic disaster.

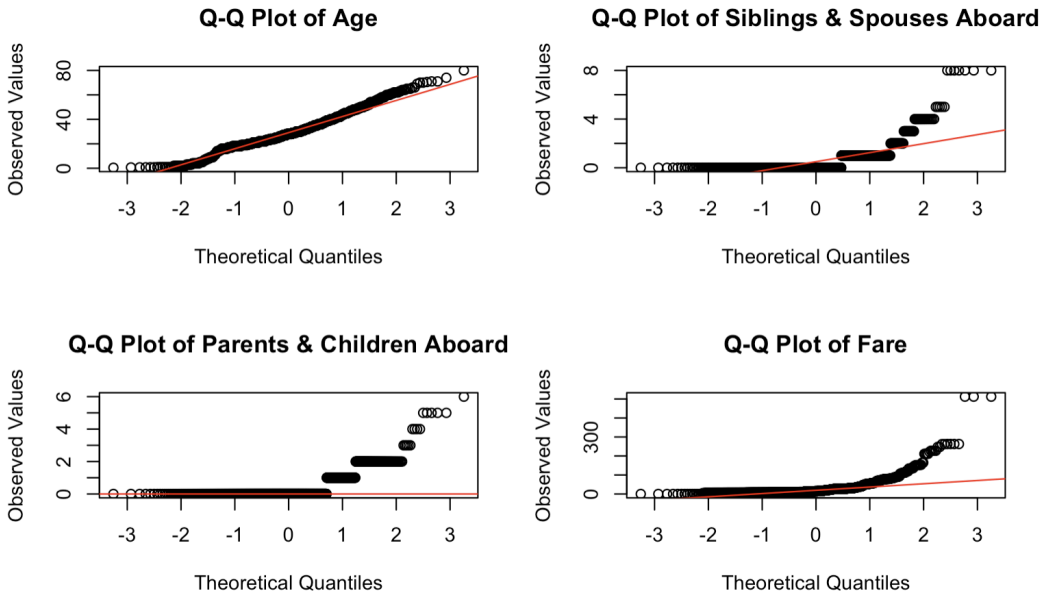
Full Definition of the Variables:

Variable Name:	**Variable Type:**	**Unit of Measurement:**
Survived	Categorical (Binary)	N/A (0 = Died, 1 = Survived)
Pclass	Categorical	N/A (1 = First Class, 2 = Second Class, 3 = Third Class)
Name	Nominal	N/A (Name of the passenger)
Sex	Categorical (Binary)	N/A (Male, Female)
Age	Quantitative	Age of the passenger
Siblings.Spouses.Aboard	Quantitative	Number of siblings and spouses aboard the ship
Parents.Children.Aboard	Quantitative	Number of parents and children aboard the ship
Fare	Quantitative	The ticket price/fare

Data Analysis

Checking the Assumption of Normality:

Before applying any classification methodology, verifying the assumption of normality for the numerical variables in the data set is essential. Fisher's Linear Discriminant Analysis (FLDA) assumes explicitly that each class follows a multivariate normal distribution with a shared covariance structure, and breaches of this assumption can affect classification performance. To ensure the validity of upcoming analyses, we assessed the normality of the numeric variables using Q-Q plots, which are presented directly below:



The Q-Q plots above display the normality assumptions for the four quantitative variables: "Age," the number of "Siblings & Spouses Aboard," the number of "Parents & Children Aboard," and "Fare," which represents the ticket price.

- The Q-Q plot for Age nearly tracks the diagonal reference line, which indicates that it meets the normality assumption.
- The Q-Q plots for Siblings & Spouses Aboard, Parents & Children Aboard, and Fare exhibit extreme departures from normality owing mainly to clustering at discrete points and the existence of extreme outliers.

- The variable, “Fare,” shows extreme skewness and values far from the reference line.
- The two other numeric variables, “Siblings & Spouses Aboard,” and “Parents & Children Aboard,” appear to have stepwise distributions, indicating they follow discrete or countable distributions rather than continuous normal distributions.

Therefore, since only one of the four quantitative variables satisfies the assumption of normality, the data set fails to fulfill this requirement. As an unfortunate result, any conclusions reached at the end of the investigation will be entirely unreliable and should not be expected to be held universally.

Assumption of Equal Variances Across All Classes:

Checking the assumption of normality was only one of two vital assumptions in this analysis. The second assumption is that all classes or categories within the categorical variable have equal variances. The Fisher’s Linear Discriminant Analysis (FLDA) method assumes that the within-class covariance matrices are equal across all categories and that this assumption is automatically satisfied. Generally speaking, this particular assumption is taken as given and does not require explicit tests, so no verification was performed in this analysis.

Applying Internal Validation on FLDA:

Fisher’s Linear Discriminant, FLDA, is a supervised classification method that works to find a linear combination of variables that separates two or more classes of objects, with this particular separation maximized as much as possible.

In this investigation, Internal Validation was first applied on FLDA using the flda() function to train the model on the data set and output a confusion matrix to assess classification accuracy. This confusion matrix shows the number of correct and incorrect classifications across the three classes of the categorical variable, “Pclass.”

In addition, the function also outputs the overall error rate, quantifying the proportion of misclassified observations. If the error rate is low, it indicates that FLDA performed well in distinguishing between the three passenger classes. In contrast, a relatively high error rate percentage indicates that FLDA is not the ideal classification method for this data set. The output for this first procedure was as follows:

```
Fisher Linear Discriminant:
class   1   2   3
    1 142   0  74
    2   9   0 175
    3   9   0 478
Error Rate = 30.10147 %
```

Analyzing the output above reveals that across the three Class categories, the FLDA method accurately categorized 142 passengers in Class 1 and 478 passengers in Class 3. Still, it did not accurately classify any Class 2 passengers. Of the 184 Class 2 observations, 9 were incorrectly classified as Class 1 passengers, and 175 were misclassified as Class 3 passengers. The model shows it had a lot of success in classifying third-class passengers, as out of the 487 total observations, 475 were correctly classified, and only 9 were misclassified as first-class passengers. As for Class 1, the model was a sufficient but not ideal capability. Of the 216 first-class passengers, 142 were correctly classified as such, but 74 were misclassified as third-class passengers.

The plentiful list of misclassifications between the three classes of Titanic passengers described above highlights that FLDA could not categorize the passengers accurately. Specifically, it could not correctly classify any of the 2nd class passengers, revealing that the model has trouble with this area. Due to potential circumstances such as overlapping distributions or nonlinear interactions between the variables, FLDA showed its lack of capability when it comes to completely separating the passengers' classes in this dataset, as indicated by the overall error rate of 30.10%, which is a high percentage of error.

Applying Internal Validation on Multinomial:

Another suitable classification method for categorical variables consisting of three or more categories or classes is the Multinomial Logistic Regression method. This method outputs a confusion matrix that compares accurate and expected labels, and, like FLDA, it was done concerning the categorical variable “Pclass.” The percentage of accurate classifications reveals the accuracy of the model. If the model showed a low misclassification rate, it would indicate that the multinomial technique effectively represented the links between predictor variables and class labels. If not, then it would reveal that the multinomial method does not accurately classify input values. For this second procedure, the following result was reached:

```
# weights:  18 (10 variable)
initial  value 974.469100
iter  10 value 604.940687
iter  20 value 465.161649
iter  30 value 464.954015
final  value 464.936023
converged
   results
      1   2   3
  1 196  14   6
  2  11  47 126
  3   9   7 471
```

The output above shows that Multinomial Regression underwent several iterations to minimize the error, ultimately converging at a final loss value of approximately 464.94. Below this, the resulting confusion matrix indicates that of the 206 first-class passengers, 196 were correctly classified as such, with only 14 misclassified as second-class and 6 misclassified as third-class. Moving on to Class 2, there were numerous misclassifications; however, it performed significantly better than FLDA, with 47 observations accurately classified as second-class passengers compared to zero accurate classifications made by FLDA. Nevertheless, Multinomial regression still misclassified 158 second-class passengers, including 11 as first-class and 147 as third-class passengers. Additionally, this model excelled at classifying third-class passengers, as out of the 487 third-class passengers, 471 were correctly labeled, with only 9 and 7 misclassified as first and second-class passengers, respectively. In summary, while the Multinomial Regression method exhibited substantial errors in distinguishing between classes 2 and 3, it was much more effective in identifying Class 2 passengers than FLDA.

Comparing the Two Internal Validations:

When comparing the two methods with Internal Validation applied to them, Multinomial Regression provides a more balanced categorization than the FLDA method, particularly for second-class passengers, as FLDA truly failed in this classification. Highlighting this significant distinction in the outcomes of the two methods, there was some degree of misclassification in both models, and they yielded similar results when classifying third-class passengers. However, the multinomial method ended up producing significantly fewer errors when it came to categorizing first-class passengers.

One possible reason is the clear lack of linearity in class separation. This would have allowed Multinomial Regression to capture the relationships better, as seen by the reduced final loss value and the improved separation of Class 2. Furthermore, the greater error rate of 30.10% suggests that FLDA, which assumes linear separability between the classes, may not be the best or ideal fit when the primary objective is accuracy in labeling the observations, making Multinomial Regression the better option under Internal Validation and for this specific dataset.

Applying External Validation on FLDA:

The two external validation procedures described above follow the two internal validations. First, external validation was applied to the FLDA model to assess its consistency, as this particular stage assists in evaluating if the model's classification performance holds steady and remains sufficient when new input data are added. The output of this first of the two external validations is stated directly below:

```
flda_predictions

      1   2   3

  1  44   0  25

  2   4   0  48

  3   2   0 144

External Validation Error Rate (FLDA) = 29.58801 %
```

The external validation of the FLDA method produced and outputted a confusion matrix which shows the struggle that the model went through and its lack of capabilities at accurately classifying passengers. It can be seen that while it had almost maximum or perfect success with classifying the third-class passengers, it did not have the same success at all with the two other classes and kept on classifying passengers from Class 1 and Class 2 as a part of Class 3. The error rate for this validation being almost 29.59% indicates that a significant number of misclassified instances took place. Specifically, it misclassified 25 first-classers and labeled them as members of Class 3 and did not have a single accurate classification for Class 2. This failure to differentiate between Class 1 and Class 2, to reiterate, resulted in zero classifications for Class

2, therefore, suggesting that there are potential existing issues regarding feature discrimination or class imbalance, ultimately affecting the linear decision boundaries between the classes.

Applying External Validation on Multinomial:

External validations were also applied to the Multinomial Logistic Regression model as a technique to establish the model's resilience. This step is necessary as it ensures that the results are independent of any particular splitting of the data set or the steps taken as part of the methodology. Below are the findings on this second external validation approach:

```
# weights:  18 (10 variable)
initial  value 681.139619
iter  10 value 412.623502
iter  20 value 354.958434
iter  30 value 354.858820
final  value 354.858099
converged
   mn_predictions
      1   2   3
  1  61   7   1
  2   3  13  36
  3   1   1 144
External Validation Error Rate (Multinomial Logistic Regression) = 18.35206 %
```

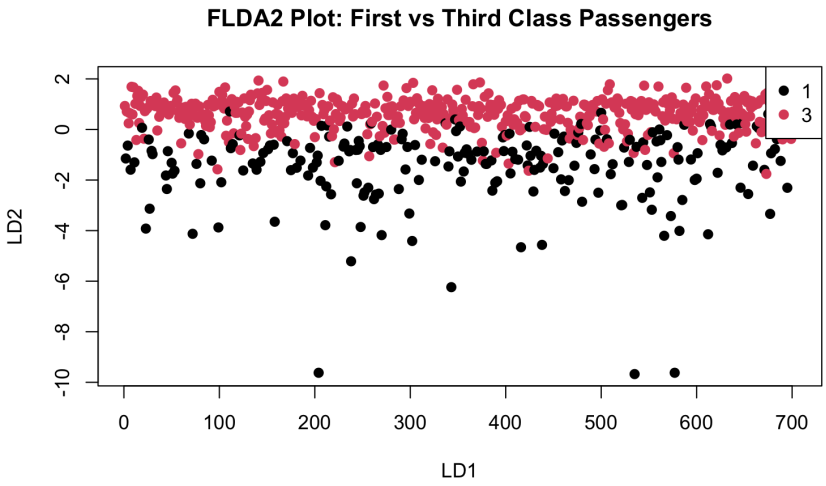
Conversely, external validation of the Multinomial Logistic Regression model displayed a better and improved performance with an error rate of about 18.35%, which is significantly lower than that of FLDA. As for the confusion matrix, it shows that this model performs better for first-class passengers as it made 61 correct predictions and 8 inaccurate ones, but it did somewhat struggle with the second and third-class passengers, misclassifying 13 Class 2 passengers as a part of Class 3. Nonetheless, it provided a much more balanced distribution of predictions across the three different classes, unlike FLDA. Additionally, the decreasing trend of training log-likelihood values, which plateau at around 354.86, verifies that the model had effectively adjusted its parameters to maximize its accuracy when classifying the passengers and what class they each belong to.

Comparing the Two External Validations:

When comparing the two models under external validation, Multinomial Logistic Regression outperformed FLDA, evidenced by the significant difference between their respective error rates of approximately 18.35% and 29.59%. The multinomial model demonstrated greater flexibility in modeling complex relationships among the variables, allowing for more accurate passenger classifications across all three ticket classes. FLDA, with its linear class boundaries, has struggled with class separation, particularly between first-class and third-class passengers. In contrast, the multinomial model showed much greater adaptability to the data distribution. This suggests that the linear assumption of FLDA could limit its performance. Meanwhile, the multinomial model effectively separates non-linearly separable data and sufficiently models non-linear decision boundaries.

Applying FLDA2 on the Categorical Variable:

FLDA2 is a practical dimension reduction approach that maximizes class separability and is well-suited to the classification job. In this case, it will be thoroughly utilized to discriminate between two of the three passenger ticket classes. By projecting the data onto a two-dimensional space while retaining discriminative information, analyzing the between-class differences becomes more straightforward. This will allow for the determination of whether or not a sufficient separation exists between the classes and to what extent a linear model can split passengers into their ticket classes. Here, it was chosen to proceed with the first and third-class passengers, Class 1 and Class 3, and leave out the second-class passengers, Class 2. The two outputs were the following plot and error rate:



```
Confusion Matrix:

class_subset   1   3

           1 137  79

           3  14 473

FLDA2 Classification Error Rate = 13.22902 %
```

The results of the FLDA2, displayed above, strongly indicate that it achieves a moderate degree of class separation between first-class and third-class passengers, specifically between Class 1 and Class 3. The projection plot above clearly differentiates between the two classes along the linear discriminants (LD1 and LD2). Although it is not significant, some overlap between these two classes still exists. This overlap is also evident in the confusion matrix below the plot, which shows that Class 1 is more likely to be misclassified, with 79 out of 216 first-class passengers being incorrectly labeled as third-class passengers, compared to only 14 third-class passengers being misclassified as first-class passengers out of 487 third-class individuals. Thus, this indicates that Class 1 is more difficult to classify, potentially due to certain feature similarities or within-class imbalance. Moreover, the FLDA2 Classification Error Rate, currently estimated at approximately 13.23%, is not bad, suggesting that while FLDA2 serves as a competent classifier for these two specific classes, further refinement—such as feature scaling or removing outliers—could be advantageous. Overall, the results are close to the expected outcomes, indicating the success of the FLDA2 model and highlighting areas for improvement to overcome its limitations.

Conclusion

In conclusion, this study applied two prominent classification methods, Fisher's Linear Discriminant Analysis (FLDA) and Multinomial Logistic Regression, in order to accurately classify passengers of the RMS Titanic based on their ticket class. Both internal and external validation techniques were utilized in the analysis to evaluate the accuracy and reliability of each of the two models. According to the findings of internal validation, Multinomial Logistic Regression fared better than FLDA, especially when it came to accurately identifying second-class passengers, as that was where the FLDA method had failed. This pattern was also

externally validated, as the Multinomial model showed superior adaptability to complex class interactions, which was proven by achieving a much lower error rate of 18.35% compared to the 29.59% of FLDA.

Further, the distinction between the first and third-class passengers, Class 1 and Class 3, was visualized using the FLDA2 projection method. The projection plot showed considerable separation, albeit there was a handful of overlap between these two classes, especially for first-class passengers, which resulted in approximately 13.23% error rate. These findings imply that while FLDA2 works well for reducing dimensionality, as it was decreased from three to two-dimensional, it still might not be the best classification method in all situations, mainly when the decision boundary between the classes is not perfectly linear.

Overall, the Multinomial Logistic Regression method performed considerably better in its classification performance for this dataset as it demonstrated more flexibility in categorizing the three passenger classes. Although the FLDA method did provide valuable insights regarding class separability, it was still constrained by its automatic assumption of linearity. Moreover, the reliability of FLDA's conclusions is further undermined by the fact that only one of the four quantitative variables satisfied the assumption of normality. On top of the assumption of normality, FLDA also assumes that the class distributions do not differ in their variances, which was an assumption that was proceeded with as given when conducting this analysis. Any violation of this assumption could also further impact the overall classification performance. Therefore, while the two methods offer important and valuable points of view, the results emphasize the benefits of more flexible classification methods when handling a complex and non-linearly separable data set.