**Name:** Matheus Vinicius Ferreira Figueiredo Teixeira

**Student Number:** 301236904

Assignment 4 – Logistic Regression

# Exercise 1

## Introduction

This exercise is about logistic regression, where the output is a discrete variable. It was proposed to us to import data about the Titanic tragedy and investigate the relationship between some features and the passenger’s survival. Did any of these features have any influence on if the passenger would survive or die?

## b.4

4 columns will not be used as features for the regression namely:

1. PassengerId
2. Name,
3. Ticket, and
4. Cabin.

The first 3 columns mentioned have unique values for every single data point (row); therefore, no patterns can be extracted from them. The fourth has many data (rows) missing; the table has 891 rows, and this particular column only has information about 204 points (rows); this corresponds to less than 25%.

## c.1.a

Figure 1: Proportion of Titanic survivors plotted against the passenger classes

I have plotted a stacked bar chart showing the proportion of people that have survived in each class. As one can notice, the number of survivors is clearly related to the passenger’s class: the higher the class, the higher the proportion of survivors. The first class had a little bit more than 60% of survivors; the second class had almost 50%, and the third class had only 30% of survivors (roughly).

## Chart, bar chart Description automatically generatedc.1.b

Figure 2: Proportion of Titanic survivors plotted against the gender of the passenger.

From the Figure 2, one can notice the number of survivors is close related to the gender of the passenger. Almost 80% of the women survived; however, only about 20% of the men has survived.

## c.2

1. More than half of the passengers died,
2. Roughly 50% of the people were traveling in the 3rd class, thus the majority of the fares were low price.
3. People traveling with big families tended to die; all families with either 5 or 8 siblings and spouses have died; also, all people traveling either with 4 or 6 parents and children have died.
4. A considerable number of people were traveling with no siblings or spouse or with no parents or children.
5. The lower the class, the bigger the family tends to be.
6. The first class had a broad range of fare prices. The second and third classes have almost the same range of prices.

Chart, box and whisker chart

Description automatically generated

Figure 3: Scatter Matrix with features and the output.

## d.12

Looking at figure 4, one can notice that:

1. the great majority of people who embarked through gate Q died,
2. a smaller number of people who embarked through gate C have died,
3. Most people who have embarked through gate S have survived.

Thus, based on the previous conclusions, it is fair to assume that gate Q was probably the one used to embark the majority of the 3rd class; most people in the 1st class must be embarked through gate S; it seems that people in the 2nd class used more than one gate to embark, it was probably a mix of gates C and S.

A picture containing text, crossword puzzle

Description automatically generated

Figure 4: histogram of each variable.

## e.5

Looking at the table below, I would recommend the 35-65% split because it has the higher mean score, but the split 10-90%, and 15-85% have a mean score almost as high as the 35-65% split. On top of that the 10-90% split also have the higher minimum score and the higher maximum score which makes it a very good option.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | run 1 (30%) | run 10% | run 15% | run 20% | run 25% | run 30% | run 35% | run 40% | run 45% | run 50% |
| Min score | 0.725806 | **0.7375** | 0.723684 | 0.694444 | 0.671642 | 0.725806 | 0.724138 | 0.735849 | 0.714286 | 0.688889 |
| Mean score | 0.776805 | 0.795201 | 0.795263 | 0.790982 | 0.790366 | 0.776805 | **0.796189** | 0.786303 | 0.767347 | 0.77096 |
| Max score | 0.825397 | **0.875** | 0.866667 | 0.873239 | 0.850746 | 0.825397 | 0.862069 | 0.87037 | 0.836735 | 0.822222 |

## b.8 and b.9

|  |  |  |
| --- | --- | --- |
|  | run (threshold = 0.5) | run (threshold = 0.75) |
| Accuracy | 0.83 | 0.81 |
| Precision | 0.87 | 0.78 |
| Recall | 0.88 | 1.00 |

## b.10

Comparing the 2 tables, it seems that the accuracy using the test that is very close to the accuracy of the training data.

## b.11

Looking to the table presented in the section 1.8 we can notice that with a higher thresholder for this particular model, one will get a worse accuracy and precision, but will get a better recall.