Assignment4_Group2

LOGISTIC REGRESSION

Contents

1 Loading files and libraries

```r
# Load necessary libraries
library(ggplot2)

# Step 1: Load + Clean Data (change as per your dataset location)
setwd("C:/Users/atuly/Documents/assesment/group work")
Train = read.csv("train.csv")
Test = read.csv('test.csv')
```

```r
# Display structure and summary of the training data
str(Train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```r
summary(Train)
```

```
##   PassengerId       Survived          Pclass          Name          
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891        
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character  
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309                     
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                     
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000                     
##                                                                     
##      Sex                 Age            SibSp           Parch       
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000  
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000  
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000  
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816  
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000  
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000  
##                     NA's   :177                                     
##     Ticket               Fare           Cabin             Embarked        
##  Length:891         Min.   :  0.00   Length:891         Length:891        
##  Class :character   1st Qu.:  7.91   Class :character   Class :character  
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character  
##                     Mean   : 32.20                                        
##                     3rd Qu.: 31.00                                        
##                     Max.   :512.33                                        
## 
```

2 Finding missing values for column “Age”

```r
# Fill in missing values for Age with mean
Train$Age[is.na(Train$Age)] = mean(Train$Age, na.rm = TRUE)
Test$Age[is.na(Test$Age)] = mean(Test$Age, na.rm = TRUE)

# Create DataFrame of independent/dependent variables, removing unnecessary columns
nonvars = c("PassengerId","Name","Ticket","Embarked","Cabin")
Train = Train[, !(names(Train) %in% nonvars)]
```

```r
str(Train)
```

```
## 'data.frame':    891 obs. of  7 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : chr  "male" "female" "female" "female" ...
##  $ Age     : num  22 38 26 35 35 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

Data processing & cleaning: The Titanic dataset was loaded and processed. Missing values in the ‘Age’ column were imputed with the mean of age. Additionally, unnecessary columns such as ‘PassengerId’, ‘Name’, ‘Ticket’, ‘Embarked’, and ‘Cabin’ were removed as they were not expected to contribute significantly to the model’s predictive power. # Logistic Regression Model On Training Data

```r
# Step 2: Build a Logistic Regression Model
# Logistic Regression function to be called is glm()
# Fitting process is not so different from the one used in linear regression.
TitanicLog = glm(Survived ~., data = Train, family = binomial)

# Display the summary of the logistic regression model
summary(TitanicLog)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = Train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.960445   0.532937   9.308  < 2e-16 ***
## Pclass      -1.084297   0.139119  -7.794 6.49e-15 ***
## Sexmale     -2.762930   0.199011 -13.883  < 2e-16 ***
## Age         -0.039702   0.007797  -5.092 3.55e-07 ***
## SibSp       -0.350725   0.109552  -3.201  0.00137 **
## Parch       -0.111963   0.117400  -0.954  0.34024
## Fare         0.002852   0.002361   1.208  0.22718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  788.73  on 884  degrees of freedom
## AIC: 802.73
##
## Number of Fisher Scoring iterations: 5
```

3 Regression Model

A logistic regression model is created using the ‘glm’ function . The model is trained to predict the probability of (binary outcome) survival (‘Survived’) variable based on all predictor variables (passenger class (‘Pclass’), sex (‘Sex’), age (‘Age’),Number of Siblings/Spouses Aboard(‘SibSP’),Number of Parents/Children Aboard(‘Parch’) and fare(‘Fare’)) in the “Train”data set.

4 Result Interpretation

Logistic regression method is used to predict a dependent variable,given a set of independent variables,Such that the dependent variable is categorical.The response binary variable holding values like 0 or 1, yes or no,A,B or c.In logistic regression, we use the logistic function $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$to fit the model.

In a logistic regression model, increasing X by one unit changes the log odds by 1, or equivalently it multiplies the odds by e^ 1,because the relationship between p(X) and X in is not a straight line 1 does not correspond to the change in p(X) associated with a one-unit increase in X. 1.First line has the original call to the generalized linear model with glm() function. 2.Coefficients Estimate:the intercept(b0)and the beta coefficient estimates associated to each predictor variable.The coefficient estimate of the variable PClass is b= -1.084297 which is negative,this means that an increase in pclass will be associated with a decreased likelihood/probability of being survived.

Standard error: We use standard error of the coefficients to measure the precision of the estimate of the coefficient.The smaller standard error,the more precise the estimate.

Z value:the z-statistic which is coefficient estimate divided by the standard error of the estimate, so a large value of the z-statistic indicates evidence against the null hypothesis.

P-value :The p-value corresponding to the z-statistic,the smaller the p-value the more significant the estimate is.In our above model variables pclass(6.49e-15),Sexmale(< 2e-16),Age(3.55e-07) have <0.05,based on smaller p-values indicates more significant model.

3.Null and Residual deviance – Null deviance shows how well the response variable is predicted by model that includes only the intercept(grand mean).Residual deviance shows how well the response variable is predicted with inclusion of independent variables. 4.AIC – Akaike information criterion which in this context, is just the residual deviance adjusted for number of parameters in the models. AIC can be used to compare one model to another model 5.Fisher scoring iterations -which just tells us how quickly the glm() function converged on the maximum likelihood estimates for the coefficients.

5 Using the model to predict survivability for Test Data

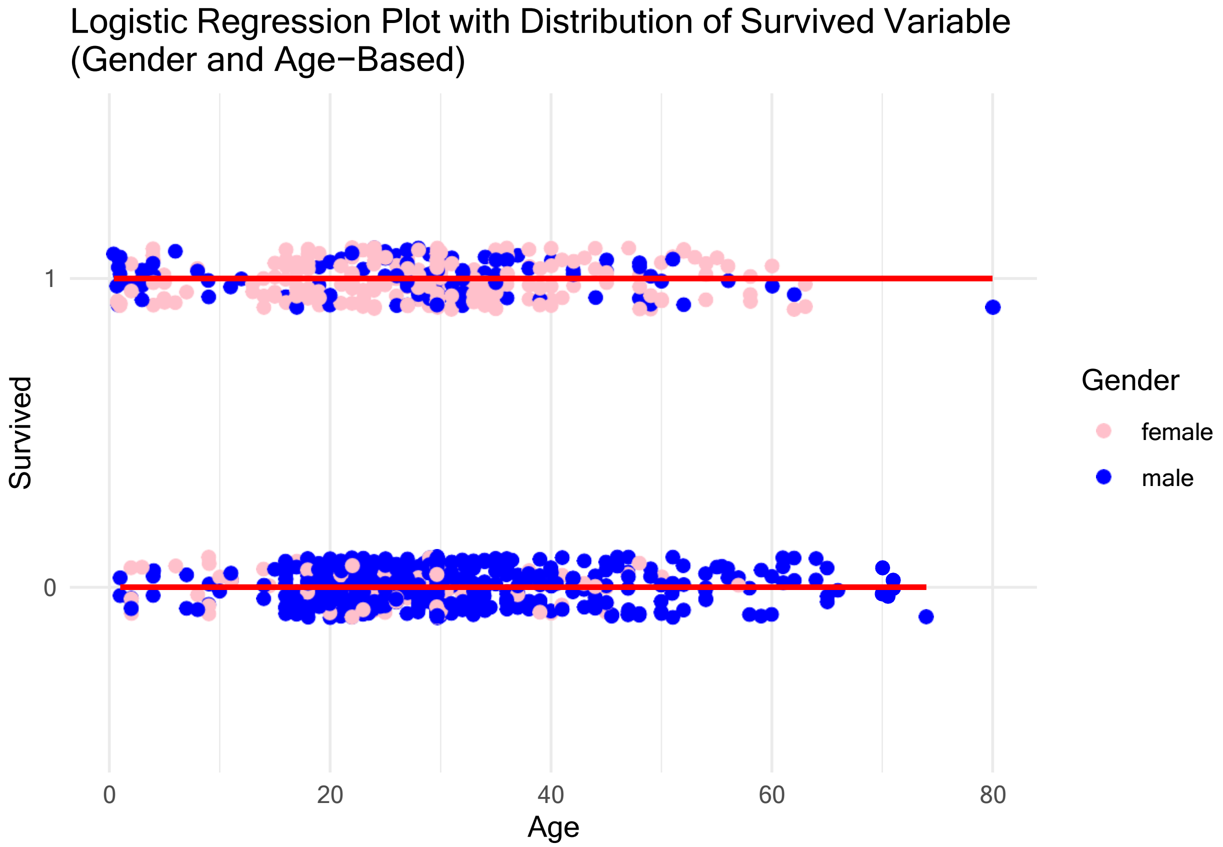
```r
predictTest = predict(TitanicLog, type = "response", newdata = Test)

# Set a threshold (0.5) for binary classification
Test$Survived = as.numeric(predictTest >= 0.5)
table(Test$Survived)
```

```
##
##   0   1
## 262 155
```

```r
# Step 4: Prepare predictions for submission
Predictions = data.frame(Test[c("PassengerId","Survived")])

# Step 5: Write predictions to a CSV file
write.csv(file = "TitanicPredictions.csv", x = Predictions)
```

6 Visualization of the logistic regression model in scatter plot

```r
ggplot(Train, aes(x = Age, y = factor(Survived), color = factor(Sex))) +
  geom_point(position = position_jitter(height = 0.1), size = 2) +
  geom_smooth(method = "glm",color = "red") +
  labs(title = "Logistic Regression Plot with Distribution of Survived Variable \n(Gender and Age-Based
       x = "Age",
       y = "Survived",
       color = "Gender") +
  theme_minimal() +
  scale_color_manual(values = c("female" = "pink", "male" = "blue"))
```



# Prediction and Visualization: Predictions on Test Data: The trained logistic regression model was applied to the test dataset to predict survival outcomes. Visualization: Visualizations such as logistic regression plots and bar plots were provided to illustrate the relationships between independent variables and survival, as well as the distribution of survival outcomes.

S-shaped Curve: The logistic regression plot displays an S-shaped curve, which is characteristic of the logistic function. This curve represents the probability of survival (Y-axis) as a function of a predictor variable (X-axis). As the predictor variable changes, the probability of survival changes smoothly from 0 to 1, fitting the binary nature of the outcome variable.

Capturing Range of Probabilities: The logistic model is adept at capturing a wide range of probabilities, unlike linear regression models that may produce predictions outside the [0,1] interval. This is essential for predicting binary outcomes accurately, as probabilities need to be bounded between 0 and 1.

Interpretation of Points: The blue dots on the plot represent individual passengers from the dataset. Each dot corresponds to a passenger, with the X-coordinate indicating whether they survived (1) or did not survive (0). By plotting these points on the logistic curve, we can visualize how well the model predicts the observed outcomes.

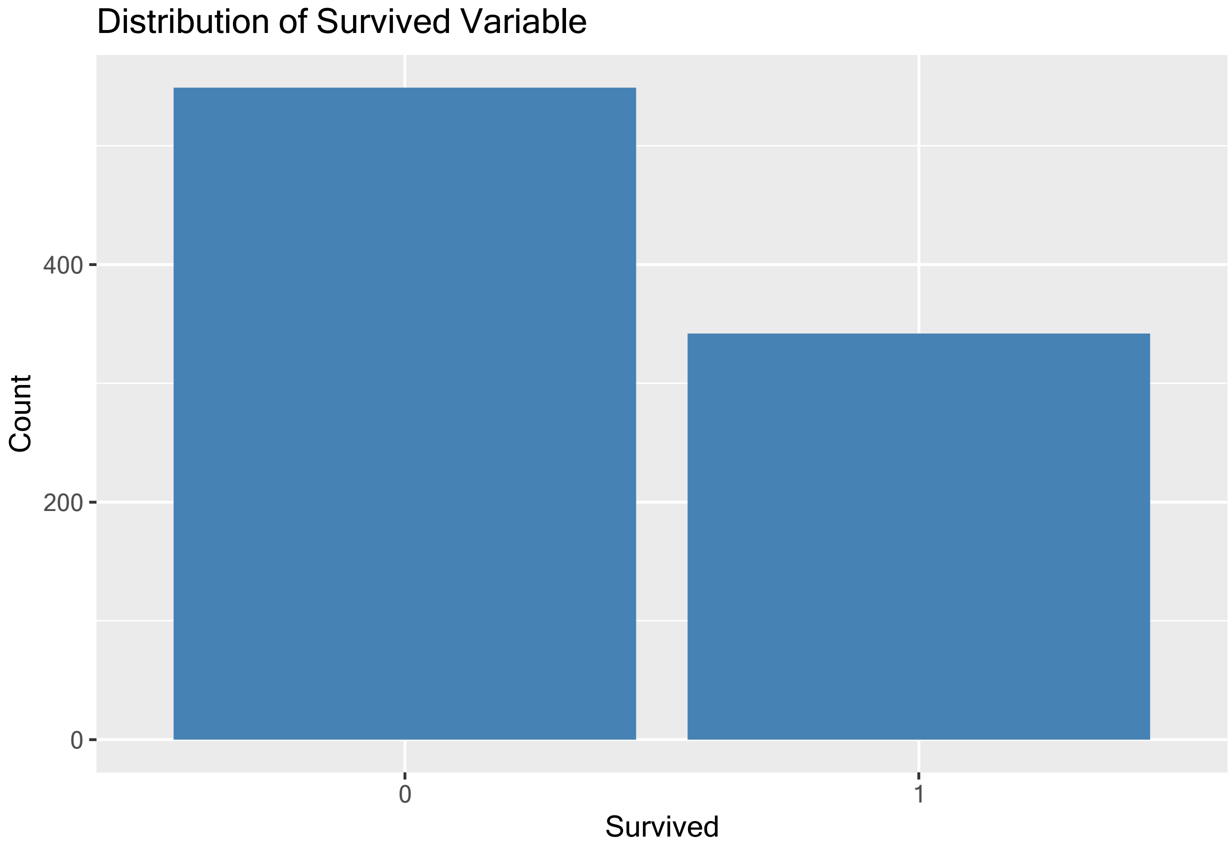
Discriminative Capacity: The logistic regression model aims to discriminate between the two classes (survived vs. not survived) based on the predictor variable(s). The plot visually demonstrates how the model distinguishes between the two classes by assigning higher probabilities of survival to passengers who actually survived (dots near 1 on the X-axis) and lower probabilities to those who did not survive (dots near 0 on the X-axis).

Threshold for Classification: Typically, a threshold probability of 0.5 is used to classify individuals into the two classes. In this plot, we can observe how the logistic curve intersects the threshold line at approximately 0.5 probability. Passengers with predicted probabilities above this threshold are classified as survivors, while those below are classified as non-survivors.

By providing this detailed interpretation, one can better understand how logistic regression models work and how they can be applied to predict binary outcomes such as survival on the Titanic. This visualization aids in explaining the model's predictive capabilities and its ability to handle binary data effectively.

7 Bar plot of the distribution of Survived variable

```r
ggplot(Train, aes(x = factor(Survived))) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution of Survived Variable",
       x = "Survived",
       y = "Count")
```

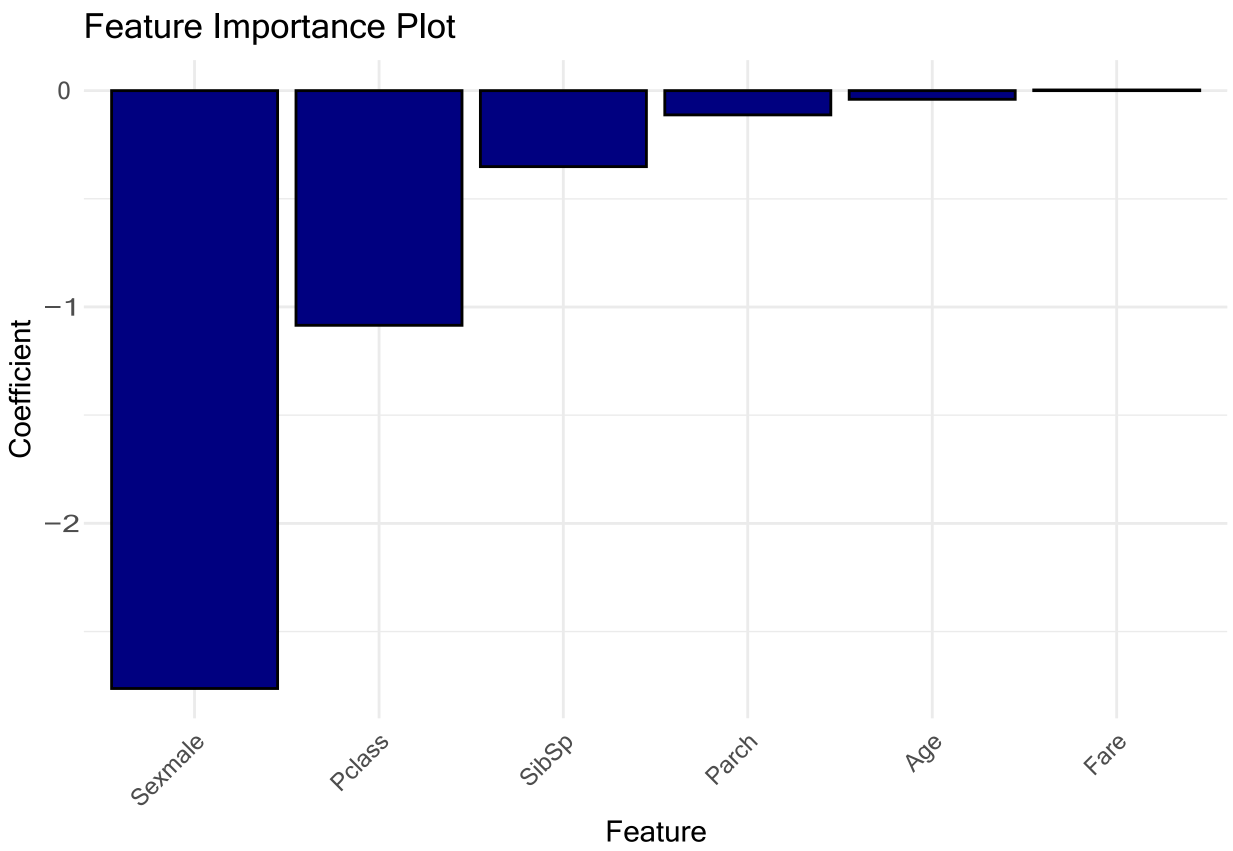


The above bar plot is showing the distribution of the "Survived" variable, with bars representing the count of each level (0 and 1;1 stands for survived and 0 stands for not survived).The x-axis represents the "Survived" variable, while the y-axis represents the count of occurrences.

8 Variable importance plot of the logistic regression model

A variable importance plot of the logistic regression model is created to visualize the coefficients of the model, which represents the importance of each variable(feature) in predicting the target variable “Survived”.

```r
# Getting the coefficients of the fitted regression model,excluding the intercept

coefficients <- coef(TitanicLog)[-1]

# Getting the names of the features

feature_names <- names(coefficients)

# Create a data frame of feature names and coefficients
feature_data <- data.frame(feature_names, coefficients)

# Sort the data frame by absolute coefficient values

feature_data <- feature_data[order(abs(feature_data$coefficients), decreasing = TRUE), ]

# Plotting feature importance

ggplot(feature_data, aes(x = reorder(feature_names, coefficients), y = coefficients)) +
  geom_bar(stat = "identity", fill = "navyblue", color = "black") +
  labs(title = "Feature Importance Plot",
       x = "Feature",
       y = "Coefficient") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



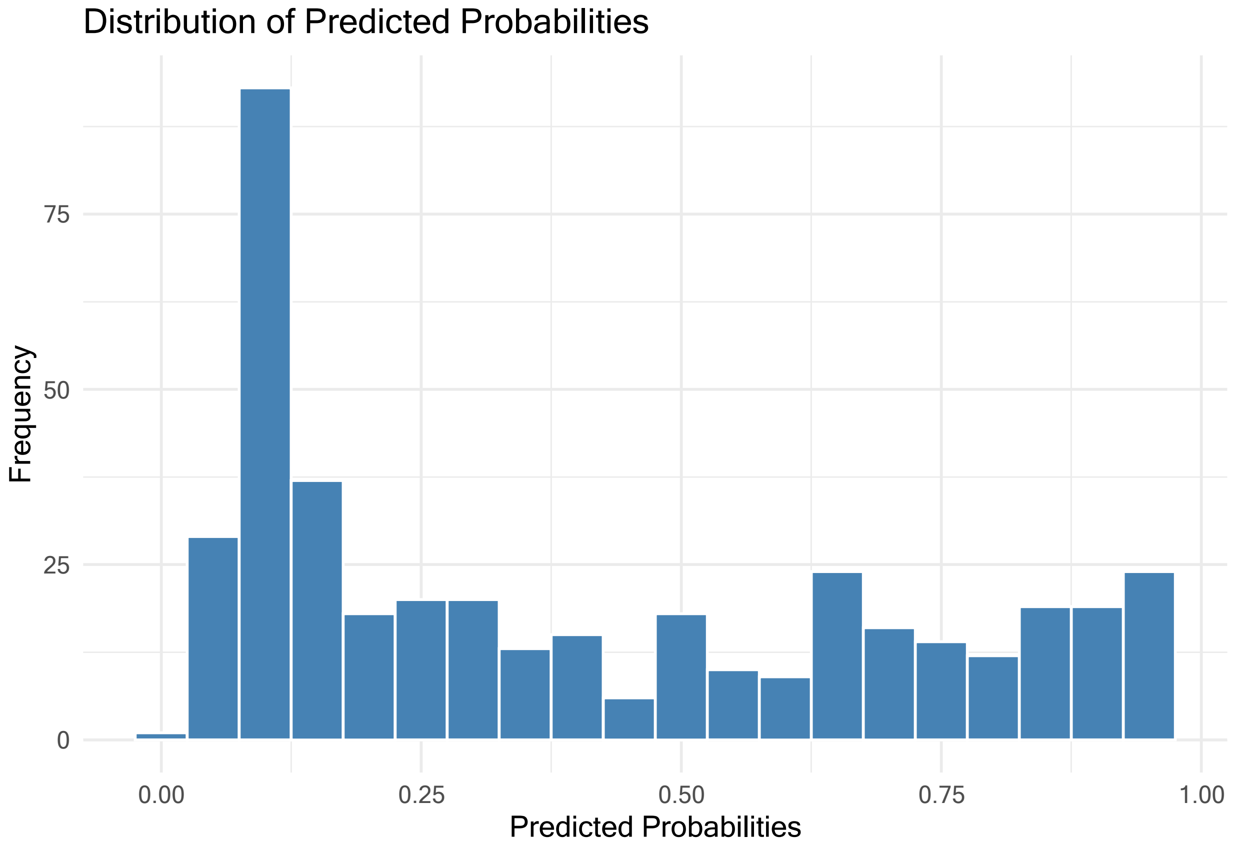
The above plot shows the importance of independent variables (features)in predicting the probability of Surviving (target variable). As we can see, the most important feature is the Sex of passengers (with coefficient -2.76). It means that the log-odds of surviving of men passengers decreases by 2.76 comparing to women passengers(coefficient named "sexmale" means that the coefficient is associated with the male category of the sex variable.The absence of a specific coefficient for "female" does not mean it's not considered in the model. The coefficient for "sexmale" captures the effect of being male compared to being female.Reference category female is implicitly included in the intercept term of the model. Therefore, when interpreting the coefficient for "sexmale," we are essentially comparing the log odds of the outcome between males and females, with females being the reference category)

The second important variable that is associated with chances of survival is Pclass (with coefficient -1.08); it means that with moving from Pclass 1 to Pclass 2, the log-odds of surviving decrease by 1.08 and moving from Pclass 2 to Pclass 3, the log-odds of surviving decrease by another 1.08.

The variable age has coefficient -0.039701625, which means that for every unit increase in age, the log odds of survival decreases by 0.039701625 units, in short as age increases the chances of survival decreases.In the same way the importance of other features can be interpreted. Fare is the least important feature with coefficient 0.002851825,which is closer to zero.Although the coefficient for fare is relatively small, it still indicates that an increase in fare is associated with a slight increase in the log-odds or odds of survival.

9 Plotting the distribution of predicted probabilities.

```
# Plot the distribution of predicted probabilities for the test set

ggplot(data = data.frame(predicted_probs = predictTest), aes(x = predicted_probs)) +
  geom_histogram(binwidth = 0.05, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Predicted Probabilities",
       x = "Predicted Probabilities",
       y = "Frequency") +
  theme_minimal()
```



Concentration of Predicted Probabilities: The plot reveals that a significant portion of predicted probabilities is concentrated in the lower range (0 to 0.25). This indicates that the model tends to predict a low likelihood of survival for many passengers in the test set.

Confidence in Predictions: Lower prediction probabilities around 0.5 suggest higher confidence in the model's predictions. Predicted probabilities closer to 0.5 indicate uncertainty, where the model is less confident in assigning passengers to either the survived or not survived category.

Model Performance Evaluation: While the distribution of predicted probabilities offers valuable insights, evaluating the model's performance requires comparing its predictions to the actual outcomes. Since actual values for the target variable are not available in the test set, an external validation dataset or techniques like cross-validation can be used to assess the model's performance.

One can can better comprehend the strengths and limitations of the logistic regression model and make informed decisions regarding its application in predicting survival on the Titanic. Additionally, it underscores the importance of continuous refinement and validation of predictive models in real-world scenarios.