CZECH UNIVERSITY OF LIFE SCIENCE PRAGUE

Faculty of Economics and Management Department of Informatics

Data Mining

Binary Logistics Regression Semestral Project

Instructor: Prof. Ing. Tomáš Hlavsa, Ph.D

Submitted by :

Seid Zemzem Mustofa

Ekeh Izuchukwu Mark-Anthony

Mighty Pasurai

Jan ,2025

Contents **Page**

List of tables **Page**

List of Figures **Page**

1. Introduction to Data Mining

Data mining is the process of extracting meaningful patterns, trends, and insights from large datasets. By combining techniques from statistics, machine learning, and database systems, it transforms raw data into actionable knowledge. It is widely used across industries like healthcare, finance, and marketing for tasks such as classification, clustering, and prediction.

Using frameworks like CRISP-DM, data mining ensures a structured approach to analyzing data, enabling organizations to make informed decisions, optimize processes, and gain a competitive edge in today's data-driven world.

2. Enhancing Data

It is refining measurement scales, adjusting variable types, and ensuring proper labeling to make the data more interpretable.

Upload data sets

BSDM_Titanic_train.csv data is uploaded to the IBM SPSS modeler through Var.File node from Sources option as shown in the figure below.

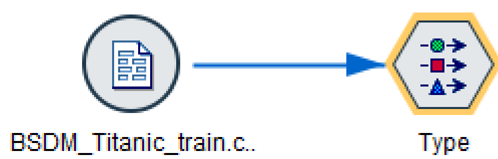

Figure 1. Uploading *BSDM_Titanic_train.csv* data

2.1.Adjust Measurements and labels of the dataset.

Data are classified into two. Quantitative data refers to information that can be measured, counted, or expressed in numerical terms. It has two types continuous: which consists of distinct for example Hight and discrete: which consists of a whole number of siblings. Another type of data is qualitative data which describes characteristics, attributes, or qualities that cannot be measured with numbers but can be observed and categorized. It can be further classified into nominal, for example, gender, and ordinal which shows rank or order such as level of education.

The given training and testing data set has quantitative and qualitative variables, In order to process in IBM SPSS modeler it needs adjustment in measurement and labels.

It processes modifying the measurement level or scaling of variables to ensure they are properly configured for analysis. Type node (figure.1) from the Field option is connected to the data *BSDM_Titanic_train.csv* . which helps to manage variable properties, including their type, and measurement level. As shown in the figure below various measurements are adjusted.

Measurement types in IBM SPSS modeler are,

Categorical: Categorical variables are variables that represent categories or groups. They can be either **nominal** or **ordinal**.

Nominal: Variables with categories that have no intrinsic order or ranking. Sex (Female and male) and Cabin (A10, A14) are adjusted as nominal variables.

Ordinal: Variables with categories that have a specific order, but the differences between the categories are not defined. Passenger class (1,2,3) is adjusted as an ordinal variable.

Continuous: Continuous variables represent numerical data that can take an infinite number of values within a given range, Age, Siblings/Spouses, and Parch are adjusted as continuous variables.

Flag: it is a binary variable used to indicate whether a certain condition or criterion. **Survived** 1, or not survived 0 is adjusted as a flag variable.

Typeless: a variable does not have a defined measurement type or category. From the data set Name and Ticket are adjusted as typeless variables.

Labels are adjusted in filed Colum (Table 1). based on the given information in the question. survived yes and no is labeled as 1 and 0 respectively. Passenger classes lower, middle, and upper are labeled as 1,2 and 3 respectively.

Table 1. Adjusting measurements and labels

Types Format Annotations

Read Values | Clear Values | Clear All Values

Field	Measurement	Values	Missing	Check	Role
PassengerId	Continuous	[1,891]		None	Record ID
Survived	Flag	1/0		None	Target
Pclass	Ordinal	1,2,3		None	Input
Name	Typeless			None	None
Sex	Nominal	female,male		None	Input
Age	Continuous	[0,80]		None	Input
SibSp	Continuous	[0,8]		None	Input
Parch	Continuous	[0,6]		None	Input

View current fields   View unused field settings

2.2.Data understanding

It is a process of exploring, cleaning, and preparing the dataset for meaningful analysis. It handles missing values, extremes and outliers. Which affects the model quality , such as overfilling.

In the training data set has 177 missing values in Age group, extremes and outlier in Sibsp, Parch and fare also age group are observed by using Data audit node . Such problems can be solved by grouping the variables. In this project SibSp Parch and Fare are grouped by using Binning node (Quartile = 4). However, Missing values in age group are filled by median 28 using Filler node. Newly created groups are used to develop the model.

In the testing dataset , there was extremes and outliers. Because of test data set should have same presentation with training dataset, SibSp , Parch and Fare are grouped. A newly created data set will be used for testing the model.

Table 2. Data validation (handling outliers, extremes, and missing values)

Audit Quality Annotations

Complete fields (%): 100%   Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Com
PassengerId	Continuous	0	0	None	Never	Fixed	
Survived	Flag	--	--	--	Never	Fixed	
Pclass	Ordinal	--	--	--	Never	Fixed	
Sex	Nominal	--	--	--	Never	Fixed	
Age	Continuous	7	0	None	Never	Fixed	
SibSp	Continuous	23	7	None	Never	Fixed	
Parch	Continuous	9	6	None	Never	Fixed	
Fare	Continuous	17	3	None	Never	Fixed	
Parch_TILE4	Nominal	--	--	--	Never	Fixed	
Fare_TILE4	Nominal	--	--	--	Never	Fixed	
SibSp_TILE4	Nominal	--	--	--	Never	Fixed	

3. Exploratory Data Analysis (EDA)

It is an approach to analyzing datasets to summarize their main characteristics, often using visual methods. It involves exploring and understanding the data before applying formal statistical modeling or machine learning algorithms.

3.1.Descriptive Statistics and Visualization

Descriptive statistics summarize and describe the essential features of a dataset, often providing insight into the central tendency, spread, and shape of the data distribution, key measures are

Measure of Central tendency: It describes the center, or typical value, of a dataset. such as mean, median, and mode.

Measure of Spread (Dispersion): it describes variability or diversity within the dataset by analyzing the extent to which data points in a dataset deviate from the central value such as range, standard deviation and variance.

3.2. Graphical representation (Visualization)

Histogram and Bar graph

A bar graph is the graphical representation of categorical data for examples, survived (0 and 1) , sex (male and female) passenger class (Upper , Lower and Middle) using rectangular bars where the length of each bar is proportional to the value they represent. A histogram is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar for example , number of Siblings/Spouses, fare, parch and age.

Survived passengers

The **Survived** variable represents binary data, where 0 indicates individuals who did not survive, and 1 represents those who survived. Out of 891 records, approximately 38.4% of individuals survived, as indicated by the means of 0.384. Most individuals did not survive, reflected in the median and mode, both of which are 0. The data shows a relatively low spread, with a standard deviation of 0.487, and is evenly distributed within its binary range of 0 to 1.

Table 3. Statistics of survived passengers

Survived

Statistics

Count	891
Mean	0.384
Min	0
Max	1
Range	1
Variance	0.237
Standard Deviation	0.487
Standard Error of Mean	0.016
Median	0
Mode	0

Table Graph Annotations

Value	Proportion	%	Count
0		61.62	549
1		38.38	342

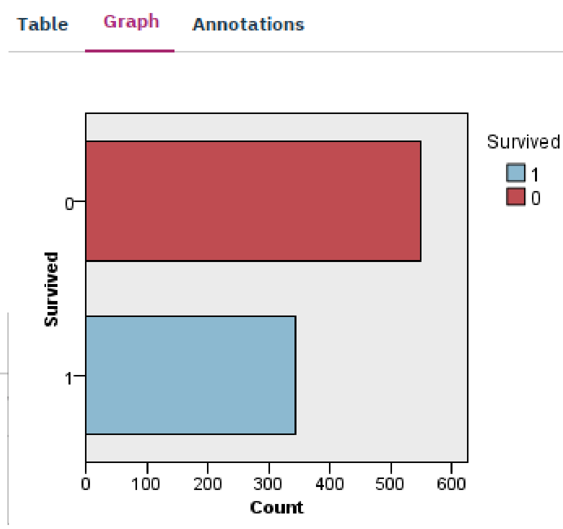

Figure 2. Distribution table and Bar graph of Survive

Sex

The dataset comprises 891 individuals, with males representing 64.76% (577 individuals) and females accounting for 35.24% (314 individuals). This indicates that most of the population is male, comprising nearly two-thirds of the total count.

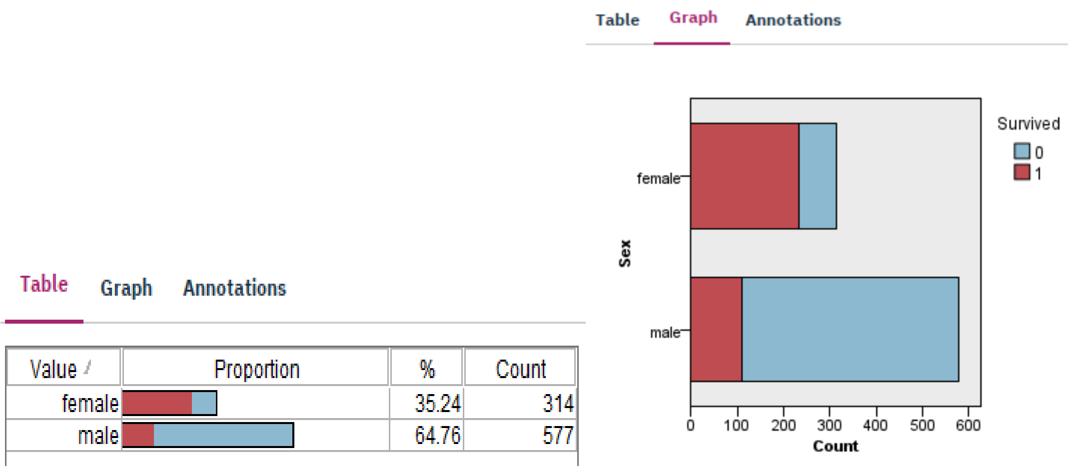

Figure 3. Distribution table and Bar graph of sex

Passenger class

The data set has 891 Passenger class observations, with a mean of 2.309 and values ranging from 1 to 3. The standard deviation is 0.836, indicating a moderate spread, while the variance is 0.699. The median and mode are both 3, showing that most observations belong to the third class. The standard error of the mean is 0.028, highlighting the precision of the mean estimate.

Table 4. Statistics of Passenger class

Pclass

Statistics

Count	891
Mean	2.309
Min	1
Max	3
Range	2
Variance	0.699
Standard Deviation	0.836
Standard Error of Mean	0.028
Median	3
Mode	3

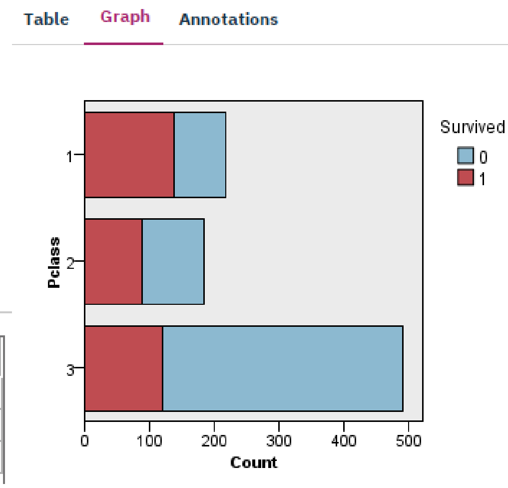

Table Graph Annotations

Value	Proportion	%	Count
1		24.24	216
2		20.65	184
3		55.11	491

Figure 4. Distribution table and Bar graph of Passenger class

Age

The **Age** variable consists of 891 records, with ages ranging from 0 to 80 (range: 80). The average age is 29.346 years, and the median age is 28, suggesting a fairly symmetric distribution. The standard deviation is 13.028, indicating notable variability in ages, while the standard error of the mean is 0.436, showing that the average age is estimated precisely. The mode is 0, indicating that 0 appears most frequently in the data, possibly representing missing or special cases.

Table 5. Statistics of Age

Age
Statistics

Count	891
Mean	29.346
Min	0
Max	80
Range	80
Variance	169.734
Standard Deviation	13.028
Standard Error of Mean	0.436
Median	28
Mode	28

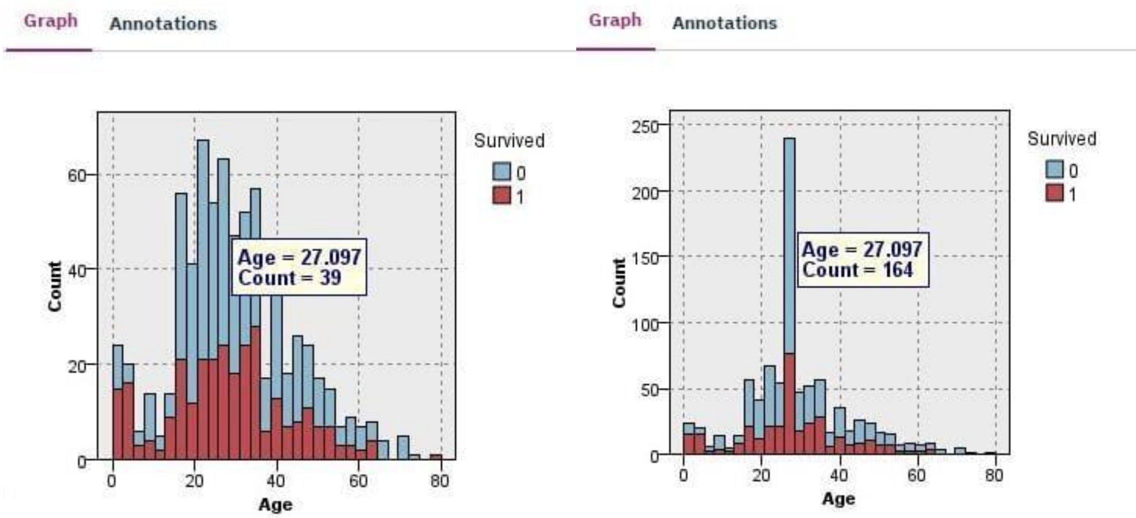

Figure 5. Histogram of Age before and after filling in missing values by median (28)

Siblings/Spouses

The descriptive statistics of the "SibSp" variable indicate that it has 891 observations. The mean value is 0.523, reflecting a low average number of Siblings/Spouses aboard. The minimum and mode values are 0, while the maximum value is 8, yielding a range of 8. The data has a variance of 1.216 and a standard deviation of 1.103, indicating moderate variability. The standard error of the mean is 0.037, suggesting a precise estimate of the mean. Both the median and mode are 0, showing that most individuals had no siblings or spouses aboard.

Table 6. statistics of Siblings/Spouses

SibSp

Statistics

Count	891
Mean	0.523
Min	0
Max	8
Range	8
Variance	1.216
Standard Deviation	1.103
Standard Error of Mean	0.037
Median	0
Mode	0

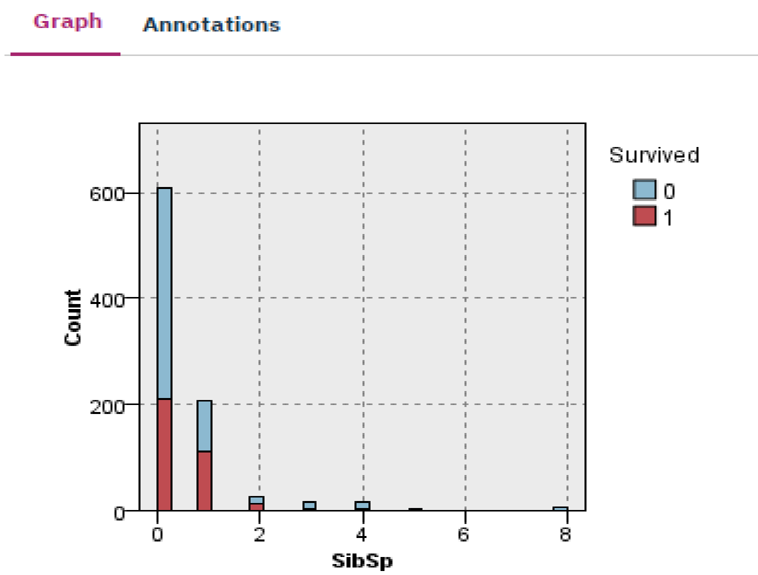

Figure 6. Histogram of Siblings/Spouses

Parch

The descriptive statistics for the "Parch" variable show that it contains 891 observations. The mean is 0.382, indicating a low average number of parents or children aboard. The minimum and mode values are both 0, while the maximum value is 6, resulting in a range of 6. The variance is 0.650, and the standard deviation is 0.806, suggesting relatively low variability. The standard error of the mean is 0.027, implying a precise estimate of the mean. Both the median and mode are 0, indicating that most individuals traveled without parents or children aboard.

Table 7. Statistics of parch

Parch

Statistics

Count	891
Mean	0.382
Min	0
Max	6
Range	6
Variance	0.650
Standard Deviation	0.806
Standard Error of Mean	0.027
Median	0
Mode	0

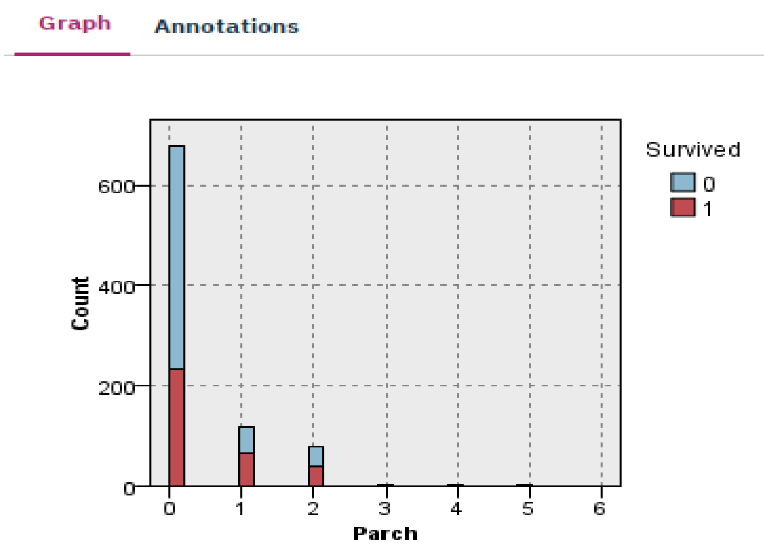

Figure 7. Histogram of parch

Passenger fare

The descriptive statistics for the "Fare" variable show that it includes 891 observations with a mean fare of 32.204. The fares range from a minimum of 0.000 to a maximum of 512.329, resulting in a wide range of 512.329. The variance is 2469.437, and the standard deviation is 49.693, indicating high variability in ticket prices. The standard error of the mean is 1.665, . indicating that lower fares were the most frequent.

Table 8. Statistics of Passenger Fare

Fare

Statistics

Count	891
Mean	32.204
Min	0.000
Max	512.329
Range	512.329
Variance	2469.437
Standard Deviation	49.693
Standard Error of Mean	1.665
Median	14.454
Mode	8.050

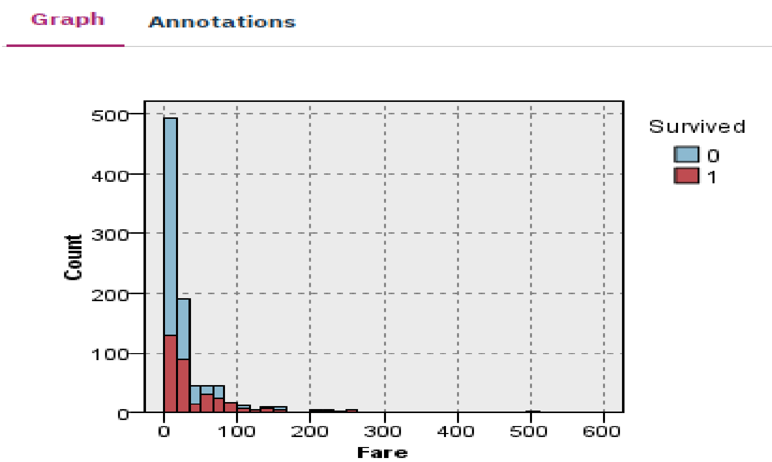

Figure 8. Histogram of fare

4. Developing a Predictive Model with Binary Logistic Regression

Binary logistic regression is a statistical method used to model the relationship between one or more independent variables and a binary dependent variable. The dependent variable is categorical and has only two possible outcomes either survived (1) or not survived (0). However, predictors can be continuous (Fare, number of Siblings/Spouses and Age) or categorical (Parch).

Training data set which has 891 rows and testing datasets which has 330 are given to develop logistic regression model.

- Odds or Chance P/(1-P)…interval (0;∞)
- Logit ln(P/(1-P)…interval (-∞; ∞)
  Where ; P is occurrence and 1-P is the component (no occurrence).

$$\boldsymbol{ln\frac{P(Y=1)}{1-P(Y=1)} = \beta o + \beta 1X1 + \beta 2X2 + \cdots + \beta kXn}$$

Where **β0**: is the intercept,

β1,β2,…,βk : are the coefficients of the predictors, it represents the change in the function for one unit change in predictors.

X1, X2,…, Xk : are the predictors.

The odds are defined as the ratio of the probability of **Y=1 to Y=0**, it is an exponential function of logit function,

$$\boldsymbol{\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta o + \beta 1X1 + \beta 2X2 + \cdots + \beta kXn}}$$

4.1. Evaluating multicollinearity among variables

It occurs when two or more independent variables in a regression model are highly correlated, meaning they provide redundant information about the dependent variable. Pearson correlation is used to analyze correlations among continuous variables. From the table below, there Is no value recorded $r > 0.75$, which shows there is no multicollinearity. Age and SibSp have a negative relationship with survival. However, parch and Fare have a positive relationship with survival.

Table 9. Pearson correlation among continuous variables

Pearson Correlations

	Survived	Age	SibSp	Parch	Fare
Survived	1.000/Perfect	-0.077/Strong	-0.035/Weak	0.082/Strong	0.257/Strong
Age	-0.077/Strong	1.000/Perfect	-0.308/Strong	-0.189/Strong	0.096/Strong
SibSp	-0.035/Weak	-0.308/Strong	1.000/Perfect	0.415/Strong	0.160/Strong
Parch	0.082/Strong	-0.189/Strong	0.415/Strong	1.000/Perfect	0.216/Strong
Fare	0.257/Strong	0.096/Strong	0.160/Strong	0.216/Strong	1.000/Perfect

4.2. Estimating parameters.

Significant relationships between predictors and target survival are summarized in the table below.

Table 10. parameter estimation of Binary logistic regression result " Event of interest 1"

Parameter Estimates

Survived[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	95% Confidence Interval for Exp (B) Upper Bound
yes	Intercept	-2.662	.429	38.574	1	<.001			
	Age	-.040	.008	25.406	1	<.001	.960	.945	.976
	[Pclass=1]	2.137	.342	39.008	1	<.001	8.473	4.333	16.568
	[Pclass=2]	1.105	.244	20.482	1	<.001	3.020	1.871	4.874
	[Pclass=3]	0[b]	.	.	0	.	.	.	.
	[Sex=female]	2.714	.199	185.707	1	<.001	15.095	10.216	22.303
	[Sex=male]	0[b]	.	.	0	.	.	.	.
	[SibSp_TILE4=1]	1.363	.412	10.963	1	<.001	3.907	1.744	8.752
	[SibSp_TILE4=2]	1.354	.392	11.958	1	<.001	3.874	1.798	8.345
	[SibSp_TILE4=3]	0[b]	.	.	0	.	.	.	.
	[Parch_TILE4=1]	.403	.340	1.404	1	.236	1.496	.768	2.914
	[Parch_TILE4=2]	.732	.367	3.976	1	.046	2.079	1.013	4.268
	[Parch_TILE4=3]	0[b]	.	.	0	.	.	.	.
	[Fare_TILE4=1]	-.250	.428	.341	1	.559	.779	.337	1.802
	[Fare_TILE4=2]	-.314	.404	.604	1	.437	.730	.331	1.614
	[Fare_TILE4=3]	-.079	.287	.076	1	.782	.924	.526	1.622
	[Fare_TILE4=4]	0[b]	.	.	0	.	.	.	.

a. The reference category is: no.

b. This parameter is set to zero because it is redundant.

4.3.Hypothetical Testing Binary logistic regression model

Omnibus Chi–Square (X^2)

It is commonly used in logistic regression or logistic regression to test whether the model explains significant variance in the dependent variable. Based on (Figure 12)chi-square analysis is performed as follows.

Hypothesis

$H0$: $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = \beta7 = \beta8 = \beta9 = \beta10 = \beta11=0$ (There is no relationship between Survived and all other explanatory variables).

$H1$: $\beta1 \neq \beta2 \neq \beta3 \neq \beta4 \neq \beta5 \neq \beta6 \neq \beta7 \neq \beta8 \neq \beta9 \neq \beta10 \neq \beta11\neq0$ There is significant relationship between Survived and at least one explanatory variable).

The chi-square value of maximum likelihood = 396.439 and Pr < 0.001. Pr value is less than 0.05(at 5% level of significance), therefore it enables us to conclude that there is at least one explanatory variable that has a statistically significant relationship with survived.

Wald X^2

It is commonly employed in logistic regression to test the significance of individual coefficients (parameters) in a regression model.

Age

$H0$: $\beta1 = 0$ (There is no relationship between Age and Survived.)

$H1$: $\beta1 \neq 0$ (There is a significant relationship between Age and Survived.)

P-value = 0.01 < α. Following this, the alternative hypothesis is accepted, and the null hypothesis is rejected. This means there is statistically significant relationship between Age and survival.

Pclass = 1

$H0$: $\beta2= 0$ (There is no relationship between Pclass=1 and survived.)

$H1$: $\beta2 \neq 0$ (There is a significant relationship between Pclass=1 and survived.)

P-value = 0.01 < α. Following this, the alternative hypothesis is accepted, and the null hypothesis is rejected. This means there is a statistically significant relationship between Pclass = 1 and survival.

Pclass = 2

$H0 : \beta3 = 0$ (There is no relationship between Pclass21 and survived.)

$H1 : \beta3 \neq 0$ (There is a significant relationship Pclass=2 and survived.)

P-value = 0.01 < α. Following this, the alternative hypothesis is accepted, and the null hypothesis is rejected. This means there is a statistically significant relationship between Pclass = 2 and survival.

Sex female

$H0 : \beta4 = 0$ (There is no relationship between the sex female and survived.)

$H1 : \beta4 \neq 0$ (There is a significant relationship between the sex female and survived.)

P-value = 0.01 < α. Following this, the alternative hypothesis is accepted, and the null hypothesis is rejected. This means there is a statistically significant relationship between Sex females and survival.

SibSp_TILE4 = 1

$H0 : \beta5 = 0$ (There is no relationship between SibSp_TILE4 = 1 and survived.)

$H1 : \beta5 \neq 0$ (There is a significant relationship between SibSp_TILE4 = 1 and survived.)

P-value = 0.01 < α. Following this, the alternative hypothesis is accepted, and the null hypothesis is rejected. This means there is a statistically significant relationship between SibSp_TILE4 = 1 and survival.

SibSp_TILE4 = 2

$H0 : \beta6 = 0$ (There is no relationship between SibSp_TILE4 = 2 and survived.)

$H1 : \beta6 \neq 0$ (There is a significant relationship between SibSp_TILE4 = 2 and survived.)

P-value = 0.01 < α. Following this, the alternative hypothesis is accepted, and the null hypothesis is rejected. This means there is a statistically significant relationship between SibSp_TILE4 = 2 and survival.

Parch_TILE4 = 1

$H0 : \beta 7 = 0$ (There is no relationship between Parch_TILE4 = 1 and survived.)

$H1 : \beta 7 \neq 0$ (There is a significant relationship between Parch_TILE4 = 1 and survived.)

P-value = 0.01 > α. following this, the null hypothesis is accepted, and the alternative hypothesis is rejected. This means there is no statistically significant relationship between Parch_TILE4 = 1 and survival.

Parch_TILE4 = 2

$H0 : \beta 8 = 0$ (There is no relationship between Parch_TILE4 = 2 and survived.)

$H1 : \beta 8 \neq 0$ (There is a significant relationship between Parch_TILE4 = 2 and survived.)

P-value = 0.01 > α. following this, the null hypothesis is accepted, and the alternative hypothesis is rejected. This means there is no statistically significant relationship between Parch_TILE4 = 2 and survival.

Fare_TILE4 = 1

$H0 : \beta 9 = 0$ (There is no relationship between Fare_TILE4 = 1 and survived.)

$H1 : \beta 9 \neq 0$ (There is a significant relationship Fare_TILE4 = 1 and survived.)

P-value = 0.01 > α. following this, the null hypothesis is accepted, and the alternative hypothesis is rejected. This means there is no statistically significant relationship between Fare_TILE4 = 1 and survival.

Fare_TILE4 = 2

$H0 : \beta 10 = 0$ (There is no relationship between Fare_TILE4 = 2 and survived.)

$H1 : \beta 10 \neq 0$ (There is a significant relationship between Fare_TILE4 = 2 and survived.)

P-value = 0.01 > α. following this, the null hypothesis is accepted, and the alternative hypothesis is rejected. This means there is no statistically significant relationship between Fare_TILE4 = 2 and survival.

Fare_TILE4 = 3

$H0$: $\beta 11$= 0 (There is no relationship between Fare_TILE4 = 3 and survived.)

$H1$: $\beta 11 \neq$ 0 (There is a significant relationship Fare_TILE4 = 3 and survived.)

P-value = 0.01 > α. following this, the null hypothesis is accepted, and the alternative hypothesis is rejected. This means there is no statistically significant relationship between Fare_TILE4 =3 and survival.

From all these tests it is possible to conclude that based on the Wald Chi-square test 2 out of 6 explanatory variables were found statistically insignificant, so they are excluded from the model.

4.4. Parameter reduction

Based on the previous hypothesis testing parameter, Fare and Parch have no significant relationship with survival they are from the model. New parameter is estimated to build new Binary logistic regression model with " Event of interest 1".

Table 11. New parameter estimation

Parameter Estimates

Survived[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
1	Intercept	-2.397	.370	42.073	1	<.001			
	Age	-.041	.008	26.599	1	<.001	.960	.945	.975
	[Pclass=1]	2.321	.245	89.774	1	<.001	10.181	6.300	16.454
	[Pclass=2]	1.147	.228	25.193	1	<.001	3.147	2.011	4.925
	[Pclass=3]	0[b]	.	.	0	.	.	.	.
	[Sex=female]	2.688	.194	191.760	1	<.001	14.704	10.051	21.511
	[Sex=male]	0[b]	.	.	0	.	.	.	.
	[SibSp_TILE4=1]	1.258	.359	12.315	1	<.001	3.519	1.743	7.104
	[SibSp_TILE4=2]	1.407	.381	13.655	1	<.001	4.085	1.937	8.618
	[SibSp_TILE4=3]	0[b]	.	.	0	.	.	.	.

a. The reference category is: 0.

b. This parameter is set to zero because it is redundant.

5. Building a logistic regression model.

Maximum Likelihood (ML) method is used to estimate the parameters (i.e., the coefficients) of the logistic regression model. The basic idea behind MLE is to find the values of the parameters that maximize the likelihood function, which measures how well the model explains the observed data. The Regression model consists of a dependent variable categorized customer revenue and the explanatory variable " X1, X2, X3, X4, X5 and X6 represent Age, Pclass=1, Pclass=2, Sex female group, SibSp_TILE4 = 2, SibSp_TILE4 = 2.

Logistic regression model by Event of interest " 1 ".

$$\boldsymbol{ln\frac{P(Y=1)}{1-P(Y=1)} = -2.397 - 0.041X2 + 2.321X2 + 1.147X3 + 2.688X4 + 1.258\ X5 + 1.407X6}$$

5.2. Odds Ratios (EXP-$\boldsymbol{\beta}$)

In logistics regression, the odds ratio measures the strength and direction of the association between two binary variables. It is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. The formula for calculating the odds ratio in the context of logistic regression is:

$$\text{OR xi} = e^{Bi}$$

where xi is the variable and Bi is the regression coefficient of a Xi.

For example, to calculate the OR of Age (X1) for the Event of interest " 1." The Coefficient of regression of Age = $\beta 1$ = - 0.041

$$ORXi = e^{-0.014}$$

$$= 0.96$$

which is the same result as the odd ratio point estimate value calculated in the IBM SPSS modeler. Negative β, results EXP(β) > 1, when age increased by 1 unit survived decreased by 1- 0.96 = 0.04 times or 4%.

Positive β results EXP(β)>1, which show the increase of target variable outcome when predictor variables are increased by one unit. For categorical variables, it means the given category has higher odds of the outcome compared to the reference category.

On the other hand, negative β results EXP(β)< 1, which shows the decreasing of the target variable survived when predictor variables are increased by one unit. For categorical variables, it means the given category has lower odds compared to the reference category.

Zero β results EXP(β) = 1, predictor variable has no effect on the odds of the outcome.

The odds ratio of predictors with the event of interest 1 is explained as follows,

Age

Exp(B) = 0.960, for every one-year increase in age, the odds of survival decrease by approximately 4% (1 - 0.960 = 0.04).

Pclass = 1

Exp(β) = 10.181, Passengers in first class are 10.18 times more likely to survive compared to the reference category Pclass = 3

Pclass = 2

Exp(β): 3.147, Passengers in second class are 3.15 times more likely to survive compared to the reference category Pclass = 3

Sex (Female) OR

Exp(β) = 14.704 Females are 14.7 times more likely to survive compared to males (reference category).

SibSp_TILE4=2

Exp(β) = 4.085, Passengers with this level of sibling/spouse count are 4.08 times more likely to survive compared to the reference category (SibSp_TILE4 = 3).

SibSp_TILE4=2

Exp(β): 3.519

Passengers with a SibSp_TILE4 = 1 (specific sibling/spouse count grouping) are 3.52 times more likely to survive compared to the reference category (SibSp_TILE4 = 3).

6. Evaluating the model quality and practical interpretations.

It is a goodness of fit in logistic regression analysis measured by Nagelkerke R Square, Confusion matrix, ROC, and AUC are commonly used to evaluate logistic regression models. In field operation, the analysis node and evaluation node are used to evaluate the model by using a testing data set.

−2 Log Likelihood (-2LL)

-2 Log Likelihood is a key statistic used to assess the fit of logistic regression models. −2LL is simply the negative of twice the log-likelihood value. It's often used because it converts the likelihood ratio into a chi-square distribution, making statistical testing easier −2 Log Likelihood

= 916.778: This is the value for the null model (intercept-only model)

=520.339: This indicates how well the model with predictors fits the data.

$\boldsymbol{\chi 2=916.778-520.339=396.439}$,

This value represents the improvement in model fit due to the inclusion of the predictors. The large reduction in −2 Log Likelihood (from 916.778 to 520.339) shows the model explains a substantial amount of variation in the outcome. Lower values of −2LL indicate the model predictions are closer to the observed outcomes which is the improvement.

Table 12. Model fitting information

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	916.778			
Final	520.339	396.439	6	<.001

Nagelkerke's R^2

It is a measure of how well a logistic regression model fits the data. It provides a value between 0 and 1, where: **0** indicates the model explains none of the variance in the outcome and **1** indicates the model explains all of the variance in the outcome.

Table 13. Pseudo R-square

Pseudo R-Square

Cox and Snell	.359
Nagelkerke	.488
McFadden	.334

Based on Table 13. Nagelkerke $R^2 = 0.488$ in the context of predicting Titanic passenger survivability with logistic regression indicates that the model explains approximately 48.8% of the variability in survival outcomes. This is a relatively good fit for a logistic regression model, particularly in a scenario involving human behavior and survival, which are influenced by complex factors. However, it's important to also evaluate other performance metrics (e.g., classification accuracy or ROC AUC) to determine whether the model is good enough for practical use.

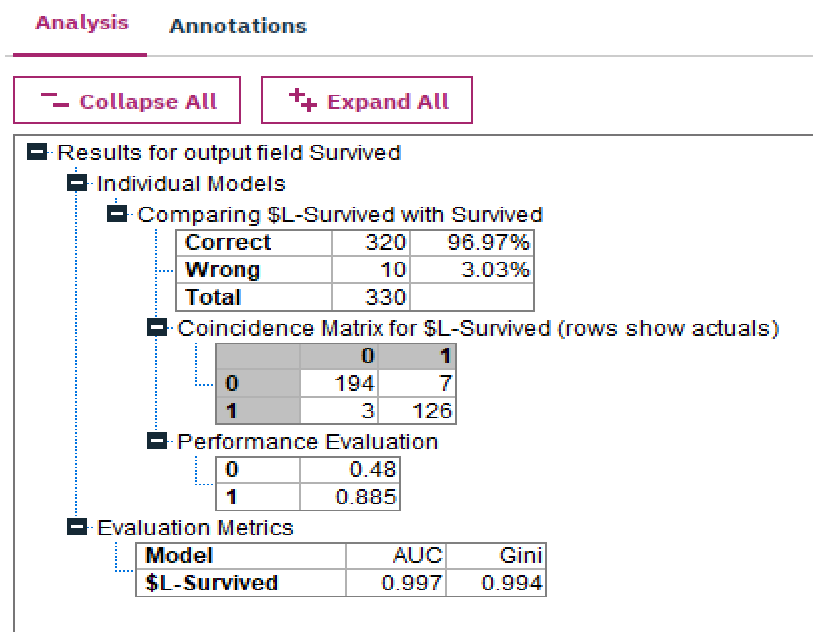

Figure 9. Result for output field survival

6.1.Confusion matrix

It is a tabular representation used to evaluate the performance of a classification model. It compares the predicted labels with the actual labels to provide insight into the accuracy and types of errors made by the model.

Table 14.Confusion matrix

Actual / Predicted	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

Where;

- **True Positives (TP)**: Correctly predicted positive cases, when the model also identifies actual surviving passengers as survived. The model classified 126 surviving passengers correctly.
- **True Negatives (TN)**: Correctly predicted negative cases, when the model also identifies the actual died passengers as died. The model classified 194 surviving passengers correctly.
- **False Positives (FP)**: Negative cases incorrectly predicted as positive, when actual died passengers are identified by the model survived. The model classified 7 died passengers as survived.
- **False Negatives (FN)**: Positive cases incorrectly predicted as negative, when actual survived passengers are identified by the model as died. The model classified 3 surviving passengers as dead.

Sensitivity True positive rate (TPR).

It measures the ability of the model to correctly identify positive cases out of all the actual positive cases, which is the true positive rate (TPR).

$$\boldsymbol{TPR = \frac{TP}{TP + FN}}$$

Based on Figure (8), the model performed a sensitivity of 0.976, in other words the ability of the model to correctly classify 1 (true positives) is 97.6% out of all 128 actual positive cases. A high sensitivity means the model classified most Survived passengers are correct.

Specificity True negative rate (TNR).

It measures the ability of the model to correctly identify negative cases out of all actual negative cases. A high TNR means the model classified most died passengers correctly .

$$TNR = \frac{TN}{TN + FP}$$

The model performed a specificity of 0.965, showing the ability of the model to correctly classify 0 (true Negatives) is 96.5% out of all 201 actual positive cases. A high specificity means the model classified most dead passengers correctly.

Accuracy

It measures the proportion of all correct predictions (both positive and negative) out of the total predictions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The model recorded 96.97% accuracy, which means out of all recorded 330 passengers 320 passengers were correctly identified, however, the model wrongly classified 10 passengers (3.03%).

It is not possible to increase both sensitivity and specificity. However, by adjusting the decision threshold, the confusion matrix helps balance sensitivity and specificity, depending on the problem's requirements. In this project, logistic regression has a 0.5 threshold.

Generally, based on the confusion matrix, it can be concluded that the model demonstrates a high level of effectiveness. This strong performance suggests that the model is reliable and capable of making accurate predictions.

6.2. Receiver Operator Characteristic (ROC) curve

It is a graphical representation used to evaluate the performance of a binary classification system as its classification threshold is varied. The ROC curve plots TPR and FPR. The ideal point on the ROC curve is the top-left corner (TPR = 1 and FPR = 0), indicating perfect classification.

Where, $TPR = \frac{TP}{TP+FN}$

and $FPR = \frac{FP}{FP+TN}$ =, $1 - specificity$

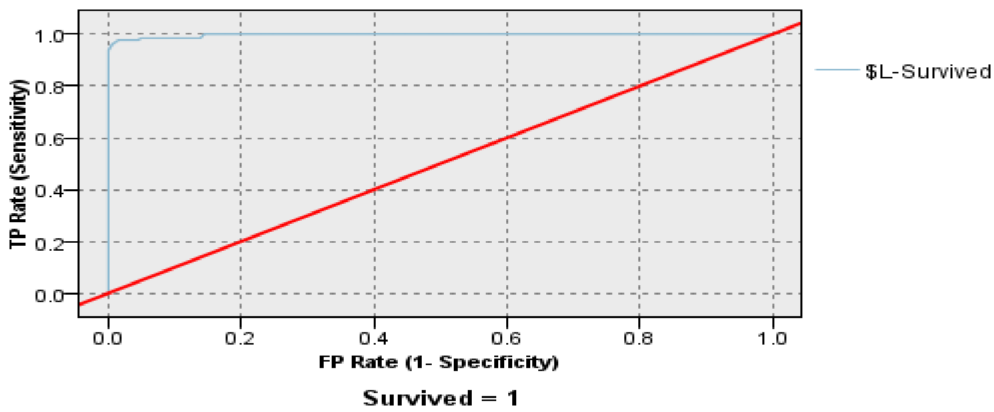

Figure 10. ROC curve of survived

6.3. Area Under the Curve (AUC)

The AUC measures the area under the ROC curve (Figure.16), providing a single scalar value that quantifies the overall ability of the model to classify outcomes. AUC ranges $0 \leq \text{AUC} \leq 1$.A value above 0.7 is generally considered acceptable, while above 0.8 is strong.

In this project, the logistic regression model exhibits $\text{AUC} = 0.997$ in model evaluation, which shows the model is almost correct.

6.4. Gini Index (coefficient)

The Gini Coefficient is a measure of how well the logistic regression model discriminates between the two classes (e.g., positive and negative). In logistic regression, the Gini Coefficient is closely tied to the ROC curve and the Area Under the Curve (AUC). The Gini coefficient ranges from -1 to Gini +1.

Where, $\boldsymbol{Gini\ cofficent = 2xAUC - 1.}$

$\boldsymbol{2x0.997 - 1 = \ 0.994}$, That shows the model is highly effective.

6.5. Lift chart

In the model, the top ~20% of the population, the lift is significantly above 2.5, indicating the model is excellent at identifying survivors with high probabilities. The model demonstrates strong predictive ability for the top predicted probabilities, and its performance gradually diminishes as less confident predictions are evaluated.

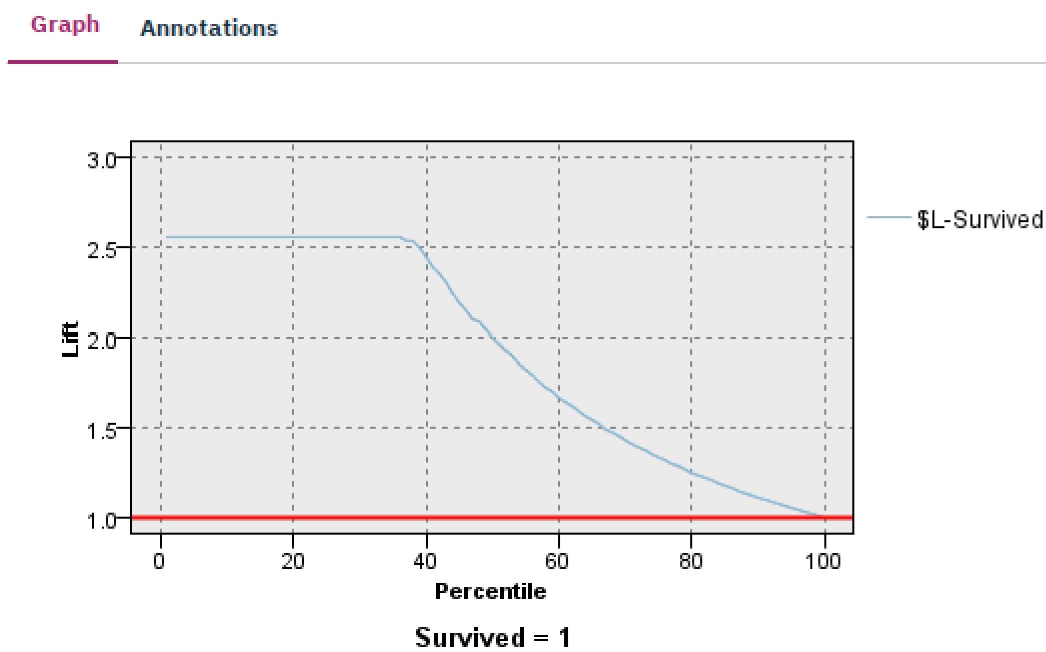

Figure 11. Lift chart of survived

6.6. Gain Charts.

Gain charts are commonly used in classification and predictive analytics to evaluate the performance of a predictive model.

The model captures about **100% of the "Survived = 1" cases within the first 40%** of the population. The blue curve flattens after 40%, showing that adding more of the population contributes little additional value to identifying "Survived = 1".

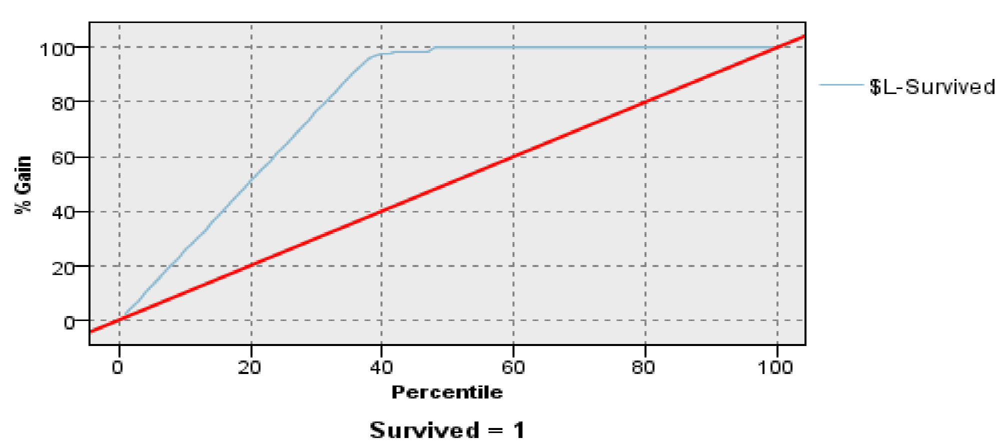

Figure 12. Gain chart of survived

7. Conclusion

The model demonstrates excellent performance with an accuracy of 96.97%, a high AUC of 0.997, and a Gini coefficient of 0.994, which indicates the model is highly effective at correctly classifying the two possible outcomes. The confusion matrix shows fewer misclassifications (10 instances), with a sensitivity of 97.6% and specificity of 96.5%, highlighting its effectiveness in predicting survivors. The Pseudo R^2 values, particularly Nagelkerke R^2 at 0.488, suggest the model explains 48.8% of the variability in the outcome, indicating a moderate but acceptable fit. Overall, the model is robust, with strong predictive power and reliable classification performance.