A picture containing logo

Description automatically generated

*Statistics for Data Analytics: CA2*

Submitted By : Vaibhav Tiwari  
Student Number : 10621051  
Lecturer: Dr Shahram Azizi

**Solution for Question 1 -**The R code file and the dataset have been provided with this report in the zip folder.

filename : Vaibhav-Q2-CA2  
Solution Explanation –

Age and EstimatedSalary were significant factors in predicting whether the customer Purchased the product or did not.

Although Age alone was a significant factor, building a predictive model on this factor alone gave trivial results. (58% model accuracy)

Adding EstimatedSalary to our simple model greatly improved its performance (83% model accuracy)

The test model accuracy was naturally lower than that of the training model. This makes sense because the model was built to fit the training data.

The data used in this notebook contains market information such as prospective buyers' Gender, Age, EstimatedSalary, as well as whether or not they actually purchased the product (Purchased). This notebook investigated whether a prospective buyers' Gender, Age, EstimatedSalary could determine whether or not they actually purchased the product (Purchased). First, we explored what the average consumer looked like and plotted market trends using histograms, barplots, and scatterplots. Next, Logistic Regression was used to build predictive models. From these, we learned that only customer Age and EstimatedSalary were significant predictors of product Purchased and the model built using this information could predict possible buyers with 83% accuracy.

**Solution for Question 2 -**

If the random variables X\_1, X\_2, ..., X\_n are identically independently distributed (iid) with Poisson (), then each random variable X\_i has the same Poisson distribution with the same parameter lambda. The probability mass function of a Poisson random variable X with parameter lambda is given by:

P(X = x) = (lambda^x \* e^(-lambda)) / x!, for x = 0, 1, 2, ...

This means that the probability that a Poisson random variable takes on the value x is given by the above formula. The parameter lambda represents the mean and variance of the Poisson distribution. For example, if lambda = 5, then the mean and variance of the Poisson distribution is 5.

The likelihood function (LF) for the random variables X\_1, X\_2, ..., X\_n that are identically independently distributed (iid) with Poisson() is given by:

LF = P(X\_1=x\_1, X\_2=x\_2, ..., X\_n=x\_n)

= P(X\_1=x\_1) \* P(X\_2=x\_2) \* ... \* P(X\_n=x\_n)

= ((lambda^x\_1 \* e^(-lambda)) / x\_1!) \* ((lambda^x\_2 \* e^(-lambda)) / x\_2!) \* ... \* ((lambda^x\_n \* e^(-lambda)) / x\_n!)

where lambda is the parameter of the Poisson distribution, and x\_1, x\_2, ..., x\_n are the observed values of the random variables X\_1, X\_2, ..., X\_n.

To compute the likelihood function, you need to know the values of lambda and the observed values of the random variables X\_1, X\_2, ..., X\_n. Once you have this information, you can plug it into the formula above to compute the likelihood function.

The solutions to part b,c and d have been provided in the R code file attached with the report – filename : Vaibhav-Q2-CA2

**Solution for Question 3 –**

This will be done using Chi-Square Test of Independence

Expected Frequency = (Row Total\*Column Total)/Grand Total

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Yes** | **No** | **Can’t Say** | **Row Total** |
| **Male** | **180** | **180** | **40** | **400** |
| **Female** | **270** | **270** | **60** | **600** |
| **Column Total** | **450** | **450** | **100** | **1000** |

Expected Frequency>=5 – Requirement Satisfied

Hypothesis Statement:

H0: Gender and opinion on women reservation are independent

H1: Gender and opinion on women reservation are dependent on each other

R Code:

x=matrix(c(200,250,150,300,50,50), ncol=3)

x

chisq.test(x,correct=FALSE)

qchisq(1-0.05,2)

Result:

X-squared (Test Statistic) = 16.204, df = 2, p-value = 0.000303

df = (columns-1)\*(rows-1)

= (3-1)\*(2-1)

= 2

Chi-squared Critical value= 5.991465

Decision Rule:

p-value<0.05,

Reject H0

Therefore, Gender and opinion on women reservation are dependent on each other.

The solutions and other computations have been provided in the R code file attached with the report – filename: Vaibhav-Q3-CA2

**REFERENCES:**

1. Statistics for Data Analytics using R by Sandeep Kumar– Udemy
2. R Programming for Statistics and Data Science 2022 – Udemy
3. Moodle notebooks - Statistics for Data Analytics.