Lab 05

1. After building the logistic regression model, with the answer for the following questions.

a) The key variables impacting the employee retention are:

→ Satisfaction Lower: Lower satisfaction increase attrition.

→ Time spent in company: Employee with 5 years tend to leave

→ Salary: Low salaries lead to higher turnover

→ No of projects & avg monthly hours: Very high or low values affect retention.

Date / /
Page

b. The accuracy of logistic regression model is 78.40%. The accuracy overall is good but the model still needs improvements.

2. Decision Tree

Consider the following dataset. Calculate entropy & information gain. w.r.t target variable 'classification'. Identify whether splitting node should be a_2 or a_3 attribute.

→

instance	a_2	a_3	classification
1	hot	High	No
2	hot	High	No
6	Cool	High	No
7	hot	High	No
8	hot	normal	Yes

Attribute a_2

Values (a_1) = Hot, Cool

S = [1+, 4-] Entropy $(S) = \frac{-1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) = 0.9709$

S_{hot} = [1+, 3-] Entropy $(S_{hot}) = -1/4 \log_2(-1/4) - 3/4 \log_2(3/4) = 0.8112$

S_{cool} = [0+, 1-] Entropy $(S_{cool}) = 0.0$

Gain (S, a_2) = Entropy $(S) - \sum_{v \in \{hot, cool\}} \frac{|S_v|}{|S|}$ Entropy (S_v)

= Entropy $(S) - \frac{4}{5}$ Entropy $(S_{hot}) - \frac{1}{5}$ Entropy (S_{cool})

$= 0.9709 - \frac{4}{5}(0.8112) - \frac{1}{5}(0.0)$

$= 0.3219$

Date __/__/__
Page _____

Attribute a_3 : Values (a_3) : High, Normal

$S_{a_3} = [1+, 4-] \Rightarrow$ Entropy $(S_{a_3}) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log\left(\frac{4}{5}\right) =$ 0.9709

$S_{High} = [0+, 4-]$: Entropy $(S_{High}) = 0.0$

$S_{Normal} = [1+, 0-]$: Entropy $(S_{Normal}) = 0.0$

$Gain(S, a_3) =$ Entropy $(S) - \frac{4}{5}$ Entropy $(S_{High}) - \frac{1}{5}$ Entropy (S_{Normal})

$= 0.9709 - \frac{4}{5}(0.0) - \frac{1}{5}(0.0)$

$= 0.9709$

$\therefore$ $Gain(S, a_2) = 0.3719$

$Gain(S, a_3) = 0.9709$ (More gain)

$\Rightarrow$ After building decision tree model with the following aims

1. For "iris.csv" dataset
   - The accuracy score of model was 1.00 (100%) meaning the model perfectly classified all samples
   - The confusion matrix shows that the model made no errors as all prediction matches the actual class : [[10 0 0] [0 9 0] [0 0 11]]
   - No misclassification is observed since all diagonal values in confusion matrix are zero.
2. For "petrol_consumption.csv" dataset
   - The regression tree structure splits data into two minimum MSE, with leaf node predicting the avg petrol consumption
   - Most important features is petrol_tax & license.
   - The regression tree predicts continuous values while a classifier predicts categories.