**Assignment- II Solution**

**Class: Semester 6 (AIDS) Subject: BDA Date of Issue: 28-03-25**

**Submission Date: 04-04-2025**

1. A financial institution wants to classify loan applicants as either "Approved" or "Rejected" using a Decision Tree model. Using Entropy and Information Gain, determine which feature should be the root node of the Decision Tree.

* Compute the entropy of the target variable.
* Compute the information gain for both "Income Level" and "Credit Score."
* Select the best root node.

|  |  |  |
| --- | --- | --- |
| Income Level | Credit Score | Loan Status (Target) |
| High | Good | Approved |
| High | Poor | Approved |
| Medium | Good | Approved |
| Medium | Poor | Rejected |
| Low | Good | Rejected |
| Low | Poor | Rejected |

1. A telecom company uses logistic regression to predict customer churn based on monthly bill amount (X1​) and contract duration (X2​). The logistic regression model is:

P(Churn=1) = 1/ 1+e−(β0​+β1​X1​+β2​X2​). Given model parameters: β0=−2.5, β1=0.05, β2=−0.02. For a customer with X1=120 (monthly bill) and X2​=12 (contract duration in months):

* Compute the probability of churn.
* If the company considers a customer "at risk" if P(Churn)>0.5, determine if this customer is likely to churn.

1. An email classification system uses Naïve Bayes to detect spam. The dataset contains two features:

Contains "Discount" (X1): 1 if the email contains the word “Discount,” 0 otherwise. Contains "Free" (X2): 1 if the email contains the word “Free,” 0 otherwise.

From previous data:

P(Spam=1) = 0.4, P(Spam=0) = 0.6

Conditional probabilities:

P(X1=1∣Spam=1) = 0.7, P(X1=1∣Spam=0) = 0.3

P(X2=1∣Spam=1) = 0.8, P(X2=1∣Spam=0) = 0.2

For an email containing both "Discount" and "Free" (X1 = 1, X2 = 1), compute: The probability that the email is spam and Classify whether the email is spam or not (threshold = 0.5).

1. A bank has developed a **fraud detection model** using supervised learning. The model was tested on a dataset with **500 transactions**, where fraud cases were identified manually. The confusion matrix is:

|  |  |  |
| --- | --- | --- |
|  | Predicted Fraud | Predicted Not Fraud |
| Actual Fraud (Fraud) | 50 | 30 |
| Actual Not Fraud (Not Fraud) | 20 | 400 |

Calculate the following metrics: **Accuracy, Precision, Recall, and F1-score**

1. A company uses **Linear Regression** to predict sales based on advertising expenditure. The model's predictions and actual values for 5 test cases are:

Actual sale: [100,150,200,250,300], predicted sales:[110,140,220,230,310].

Compute the following regression evaluation metrics:

Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² (Coefficient of Determination)

**1. Decision Tree – Entropy & Information Gain**

**Step 1: Entropy of Target Variable (Loan Status)**

Target classes:

* Approved = 3
* Rejected = 3  
  Total = 6

Entropy(S)=−p+log⁡2(p+)−p−log⁡2(p−)=−(36)log⁡2(36)−(36)log⁡2(36)=1.0\text{Entropy}(S) = -p\_+\log\_2(p\_+) - p\_-\log\_2(p\_-) = -\left(\frac{3}{6}\right)\log\_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log\_2\left(\frac{3}{6}\right) = 1.0Entropy(S)=−p+​log2​(p+​)−p−​log2​(p−​)=−(63​)log2​(63​)−(63​)log2​(63​)=1.0

**Step 2: Information Gain for “Income Level”**

**Group by Income Level:**

| **Income Level** | **Approved** | **Rejected** |
| --- | --- | --- |
| High | 2 | 0 |
| Medium | 1 | 1 |
| Low | 0 | 2 |

**Entropy of each group:**

* **High (2 samples):** Entropy = 0 (pure)
* **Medium (2 samples):**

−(12log⁡2(12)+12log⁡2(12))=1.0-\left(\frac{1}{2}\log\_2\left(\frac{1}{2}\right) + \frac{1}{2}\log\_2\left(\frac{1}{2}\right)\right) = 1.0−(21​log2​(21​)+21​log2​(21​))=1.0

* **Low (2 samples):** Entropy = 0 (pure)

**Weighted Entropy:**

26⋅0+26⋅1+26⋅0=26=0.333\frac{2}{6} \cdot 0 + \frac{2}{6} \cdot 1 + \frac{2}{6} \cdot 0 = \frac{2}{6} = 0.33362​⋅0+62​⋅1+62​⋅0=62​=0.333

**Information Gain:**

IG=1.0−0.333=0.667IG = 1.0 - 0.333 = 0.667IG=1.0−0.333=0.667

**Step 3: Information Gain for “Credit Score”**

| **Credit Score** | **Approved** | **Rejected** |
| --- | --- | --- |
| Good | 2 | 1 |
| Poor | 1 | 2 |

**Entropy for both groups:**

* For each:

−(23log⁡2(23)+13log⁡2(13))≈0.918-\left(\frac{2}{3}\log\_2\left(\frac{2}{3}\right) + \frac{1}{3}\log\_2\left(\frac{1}{3}\right)\right) \approx 0.918−(32​log2​(32​)+31​log2​(31​))≈0.918

**Weighted Entropy:**

36⋅0.918+36⋅0.918=0.918\frac{3}{6} \cdot 0.918 + \frac{3}{6} \cdot 0.918 = 0.91863​⋅0.918+63​⋅0.918=0.918

**Information Gain:**

1.0−0.918=0.0821.0 - 0.918 = 0.0821.0−0.918=0.082

**✅ Root Node: Income Level (Higher IG = 0.667)**

**2. Logistic Regression – Probability of Churn**

**Given:**

* β0=−2.5,β1=0.05,β2=−0.02\beta\_0 = -2.5, \beta\_1 = 0.05, \beta\_2 = -0.02β0​=−2.5,β1​=0.05,β2​=−0.02
* X1=120,X2=12X\_1 = 120, X\_2 = 12X1​=120,X2​=12

z=−2.5+(0.05⋅120)+(−0.02⋅12)=−2.5+6−0.24=3.26z = -2.5 + (0.05 \cdot 120) + (-0.02 \cdot 12) = -2.5 + 6 - 0.24 = 3.26z=−2.5+(0.05⋅120)+(−0.02⋅12)=−2.5+6−0.24=3.26 P(Churn=1)=11+e−3.26≈11+0.0384≈0.963P(\text{Churn}=1) = \frac{1}{1 + e^{-3.26}} \approx \frac{1}{1 + 0.0384} \approx 0.963P(Churn=1)=1+e−3.261​≈1+0.03841​≈0.963

**✅ At Risk: YES (0.963 > 0.5)**

**3. Naïve Bayes – Email Spam Detection**

Given:

* P(Spam=1)=0.4,P(Spam=0)=0.6P(Spam=1) = 0.4, P(Spam=0) = 0.6P(Spam=1)=0.4,P(Spam=0)=0.6
* P(X1=1∣Spam=1)=0.7,P(X2=1∣Spam=1)=0.8P(X\_1=1|Spam=1) = 0.7, P(X\_2=1|Spam=1) = 0.8P(X1​=1∣Spam=1)=0.7,P(X2​=1∣Spam=1)=0.8
* P(X1=1∣Spam=0)=0.3,P(X2=1∣Spam=0)=0.2P(X\_1=1|Spam=0) = 0.3, P(X\_2=1|Spam=0) = 0.2P(X1​=1∣Spam=0)=0.3,P(X2​=1∣Spam=0)=0.2

**Compute Posterior Probabilities (Unnormalized):**

* **Spam = 1:**

P=0.4⋅0.7⋅0.8=0.224P = 0.4 \cdot 0.7 \cdot 0.8 = 0.224P=0.4⋅0.7⋅0.8=0.224

* **Spam = 0:**

P=0.6⋅0.3⋅0.2=0.036P = 0.6 \cdot 0.3 \cdot 0.2 = 0.036P=0.6⋅0.3⋅0.2=0.036

**Normalize:**

P(Spam=1∣X1=1,X2=1)=0.2240.224+0.036≈0.8627P(\text{Spam}=1 | X\_1=1, X\_2=1) = \frac{0.224}{0.224 + 0.036} \approx 0.8627P(Spam=1∣X1​=1,X2​=1)=0.224+0.0360.224​≈0.8627

**✅ Classified as Spam (0.86 > 0.5)**

**4. Confusion Matrix Metrics**

|  | **Predicted Fraud** | **Predicted Not Fraud** |
| --- | --- | --- |
| **Actual Fraud** | 50 | 30 |
| **Actual Not Fraud** | 20 | 400 |

**Metrics:**

* **Accuracy:**

TP+TNTotal=50+400500=0.9\frac{TP + TN}{Total} = \frac{50 + 400}{500} = 0.9TotalTP+TN​=50050+400​=0.9

* **Precision:**

TPTP+FP=5050+20=5070≈0.714\frac{TP}{TP + FP} = \frac{50}{50 + 20} = \frac{50}{70} \approx 0.714TP+FPTP​=50+2050​=7050​≈0.714

* **Recall:**

TPTP+FN=5050+30=5080=0.625\frac{TP}{TP + FN} = \frac{50}{50 + 30} = \frac{50}{80} = 0.625TP+FNTP​=50+3050​=8050​=0.625

* **F1 Score:**

F1=2⋅Precision⋅RecallPrecision+Recall=2⋅0.714⋅0.6250.714+0.625≈0.666F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{0.714 \cdot 0.625}{0.714 + 0.625} \approx 0.666F1=2⋅Precision+RecallPrecision⋅Recall​=2⋅0.714+0.6250.714⋅0.625​≈0.666

**5. Linear Regression Evaluation Metrics**

**Actual:** [100,150,200,250,300]  
**Predicted:** [110,140,220,230,310]

**Step-by-step errors:**

| **Actual** | **Predicted** | **Error** | **Absolute** | **Squared** |
| --- | --- | --- | --- | --- |
| 100 | 110 | -10 | 10 | 100 |
| 150 | 140 | 10 | 10 | 100 |
| 200 | 220 | -20 | 20 | 400 |
| 250 | 230 | 20 | 20 | 400 |
| 300 | 310 | -10 | 10 | 100 |

* **MAE = (10 + 10 + 20 + 20 + 10)/5 = 14**
* **MSE = (100 + 100 + 400 + 400 + 100)/5 = 220**
* **RMSE = √220 ≈ 14.83**
* **R²:**

Total Sum of Squares (TSS)=∑(yi−yˉ)2=25000\text{Total Sum of Squares (TSS)} = \sum (y\_i - \bar{y})^2 = 25000Total Sum of Squares (TSS)=∑(yi​−yˉ​)2=25000 Residual Sum of Squares (RSS)=1100\text{Residual Sum of Squares (RSS)} = 1100Residual Sum of Squares (RSS)=1100 R2=1−RSSTSS=1−110025000=0.956R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{1100}{25000} = 0.956R2=1−TSSRSS​=1−250001100​=0.956

**✅ Summary of Answers**

| **Q#** | **Metric/Result** |
| --- | --- |
| 1 | Root = Income Level |
| 2 | Churn = 96.3%, Likely to churn |
| 3 | Spam Probability = 86.27%, Classified as Spam |
| 4 | Accuracy = 0.9, Precision = 0.714, Recall = 0.625, F1 = 0.666 |
| 5 | MAE = 14, MSE = 220, RMSE ≈ 14.83, R² ≈ 0.956 |