DS UNIT 6 NOTES

4. What might be a drawback of evaluation measures based on squared error? How might we avoid this?

ANS:

- One of the drawbacks of the above evaluation measures is influence of outliers.
- This is because the above measures are based on the squared error, which is much larger for outliers than for the bulk of the data.
- Thus, the analyst may prefer to use the Mean Absolute Error (MAE).
- The MAE is defined as follows:

  Mean Absolute Error (MAE) = (summation |Yi – Yi^|)/n
- To calculate MAE analyst may perform following steps:
  1. Calculate the estimated target values, Yi^.
  2. Find the absolute value between each estimated value, and its associated actual target value Yi, giving you |Yi – Yi^|.
  3. Find the mean of absolute values from step 2. This is MAE.

4. Explain model evaluation techniques for the estimation and prediction tasks.

ANS:

- For estimation and prediction models, we are provided with both the estimated value ˆy and the actual value y.
- Therefore, a natural measure to assess model adequacy is to examine the estimation error, or residual, (y ˆ-y).
- Usual measure used to evaluate estimation or prediction models is the Mean Square Error (MSE):

  MSE = $$\frac{\sum_i (y_i - \hat{y}_i)^2}{n - p - 1}$$

  where p represents the number of model variables
- Models that minimize MSE are preferred.

- Square root of MSE can be regarded as an estimate of typical error.
- This is known as standard error of estimate and denoted by s = sqrt of MSE.
- Trade-off between model complexity and prediction error.
- An evaluation measure that was related to MSE is Sum of Squared Errors (SSE)

$$\boxed{\text{SSE} = \sum_{\text{Records}} \sum_{\text{Output nodes}} (\text{actual output})^2}$$

- Another measure of the goodness of a regression model that is the coefficient of determination

$$\boxed{R^2 = \frac{\text{SSR}}{\text{SST}}}$$

where SST(Sum of Squares Total) is given by

$$\text{SST} = \sum_{i=1}^{n} (y - \bar{y})^2$$

SSR (Sum of Squares Regression)

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$

6. Explain model evaluation measures for the classification task.

ANS:

In context of C5.0 model for classifying income, we examine following evaluative concepts:

- Model accuracy
- Overall error rate
- Sensitivity and specificity
- False-positive rate and false-negative rate
- Proportions of true positives and true negatives
- Proportions of false positives and false negatives
- Misclassification costs and overall model cost
- Cost-benefit table
- Lift charts

- Gains charts

- Applied a C5.0 model for classifying whether a person's income was low(≤$50,000) or high (>$50,000)
- Predictor variables which included capital gain, capital loss, marital status, and so on.
- Let us evaluate the performance of that decision tree classification model using the notions of error rate, false positives, and false negatives.
- Let TN, FN, FP, and TP represent the numbers of true negatives, false negatives, false positives, and true positives, respectively.
- Also, let
  - TAN = Total actually negative = TN + FP
  - TAP = Total actually positive = FN + TP
  - TPN = Total predicted negative = TN + FN
  - TPP = Total predicted positive = FP + TP
- Further, let N = TN + FN + FP + TP represent the grand total of the counts in the four cells.

7. Explain classification evaluation measures accuracy, overall error rate, sensitivity and specificity.

ANS:

- *accuracy is given by,*

$$\boxed{\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{FP} + \text{TP}} = \frac{\text{TN} + \text{TP}}{N}}$$

- *overall error rate is given by,*

$$\boxed{\text{Overall error rate} = 1 - \text{Accuracy} = \frac{\text{FN} + \text{FP}}{\text{TN} + \text{FN} + \text{FP} + \text{TP}} = \frac{\text{FN} + \text{FP}}{N}}$$

- Accuracy represents an overall measure of proportion of correct classifications
- overall error rate measures proportion of incorrect classifications

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Total actually positive}} = \frac{TP}{TAP} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Total actually negative}} = \frac{TN}{TAN} = \frac{TN}{FP + TN}$$

- *Sensitivity measures the ability of the model to classify a record positively*
- *While specificity measures the ability to classify a record negatively.*

8. What is the difference between the total predicted negative and the total actually negative?

ANS:

General form of the contingency table of correct and incorrect classifications

		Predicted Category 0	Predicted Category 1	Total
Actual category	0	Truenegatives: Predicted 0 Actually 0	Falsepositives: Predicted 1 Actually 0	Totalactuallynegative
	1	Falsenegatives: Predicted 0 Actually 1	Truepositives: Predicted1 Actually1	Totalactuallypositive
	Total	Total Predictednegative	Total Predictedpositive	Grandtotal

Let TN, FN, FP, and TP represent the numbers of true negatives, false negatives, false positives, and true positives, respectively.

TPN = Total predicted negative = TN + FN

TPP = Total predicted positive = FP + TP

9. What is the relationship between accuracy and overall error rate?

ANS:

General form of the contingency table of correct and incorrect classifications

		Predicted Category		
		0	1	Total
Actual category	0	Truenegatives: Predicted 0 Actually 0	Falsepositives: Predicted 1 Actually 0	Totalactuallynegative
	1	Falsenegatives: Predicted 0 Actually 1	Truepositives: Predicted1 Actually1	Totalactuallypositive
	Total	Total Predictednegative	Total Predictedpositive	Grandtotal

Let TN, FN, FP, and TP represent the numbers of true negatives, false negatives, false positives, and true positives, respectively.

- accuracy is given by,

$$\boxed{\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{FP} + \text{TP}} = \frac{\text{TN} + \text{TP}}{N}}$$

- overall error rate is given by,

$$\boxed{\text{Overall error rate} = 1 - \text{Accuracy} = \frac{\text{FN} + \text{FP}}{\text{TN} + \text{FN} + \text{FP} + \text{TP}} = \frac{\text{FN} + \text{FP}}{N}}$$

- Accuracy represents an overall measure of proportion of correct classifications
- overall error rate measures proportion of incorrect classifications

10. Explain classification evaluation measures false-positive rate and falsenegative rate, proportions of true positives, true negatives, false positives, and false negatives.

ANS:

$$\text{False positive rate} = 1 - \text{specificity} = \frac{\text{FP}}{\text{TAN}} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{False negative rate} = 1 - \text{sensitivity} = \frac{\text{FN}}{\text{TAP}} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

$$\text{Proportion of true positives} = \text{PTP} = \frac{\text{TP}}{\text{TPP}} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

$$\text{Proportion of true negatives} = \text{PTN} = \frac{\text{TN}}{\text{TPN}} = \frac{\text{TN}}{\text{FN} + \text{TN}}$$

$$\text{Proportion of false positives} = 1 - \text{PTP} = \frac{\text{FP}}{\text{TPP}} = \frac{\text{FP}}{\text{FP} + \text{TP}}$$

$$\text{Proportion of false negatives} = 1 - \text{PTN} = \frac{\text{FN}}{\text{TPN}} = \frac{\text{FN}}{\text{FN} + \text{TN}}$$

- Using these classification model evaluation measures, analyst may compare accuracy of various models
- For example, a C5.0 decision tree model may be compared against a classification and regression tree (CART) decision tree model or a neural network model.
- Model choice decisions can then be rendered based on the relative model performance based on these evaluation measures.
- A false positive would be considered a type I error in this setting, incorrectly rejecting null hypothesis
- A false negative would be considered a type II error, incorrectly accepting null hypothesis.

12. With suitable example explain decision cost/benefit analysis.

ANS:

- Company managers may require that model comparisons be made in terms of cost/benefit analysis.
- For example, in comparing the original C5.0 model before cost adjustment (model 1) against C5.0 model using cost adjustment (model 2)
- Managers may prefer to have respective error rates, false negatives and false positives, translated into dollars and cents.
- Analysts can provide model comparison in terms of anticipated profit or loss by associating a cost or benefit with each of the four possible combinations of correct and incorrect classifications.

Cost/benefit table for each combination of correct/incorrect decision

Outcome	Classification	Actual Value	Cost	Rationale
True negative	≤50,000	≤50,000	\$0	No money gained or lost
True positive	>50,000	>50,000	−\$300	Anticipated average interest revenue from loans
False negative	≤50,000	>50,000	\$0	No money gained or lost
False positive	>50,000	≤50,000	\$500	Cost of loan default averaged over all loans to ≤50,000 group

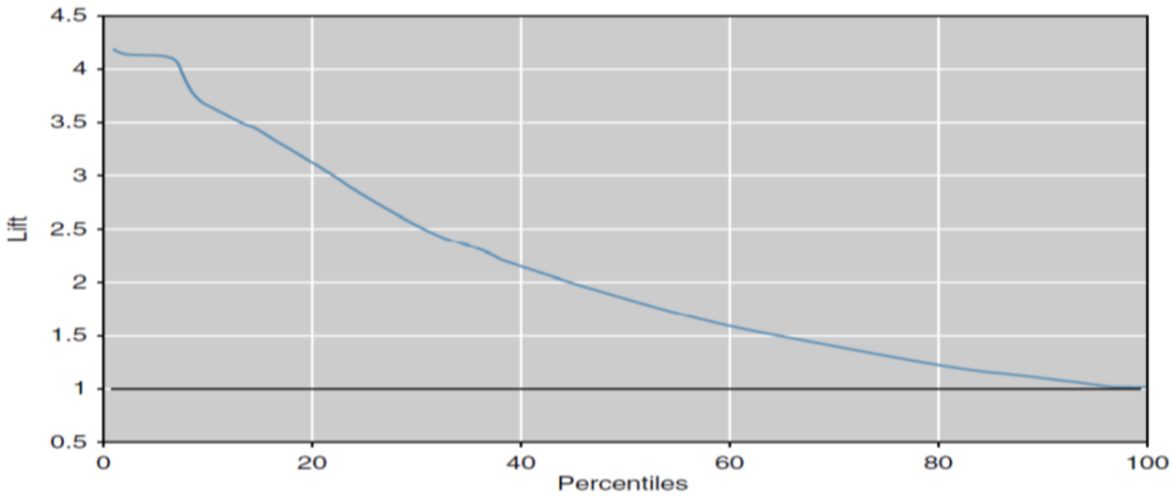
- Cost of model 1 (false positive cost not doubled):
  18, 197(\$0) + 3423(−\$300) + 2561(\$0) + 819(\$500) = −\$275, 100
- Cost of model 2 (false positive cost doubled):
  18, 711(\$0) + 2677(−\$300) + 3307(\$0) + 305(\$500) = −\$382, 900
- Negative costs represent profits.
- Thus, the estimated cost savings from deploying model 2
  −\$275, 100 − (−\$382, 900) = \$107, 800 (increases company's profit)

13. Explain use of lift charts and gains charts to compare model performance.

ANS:

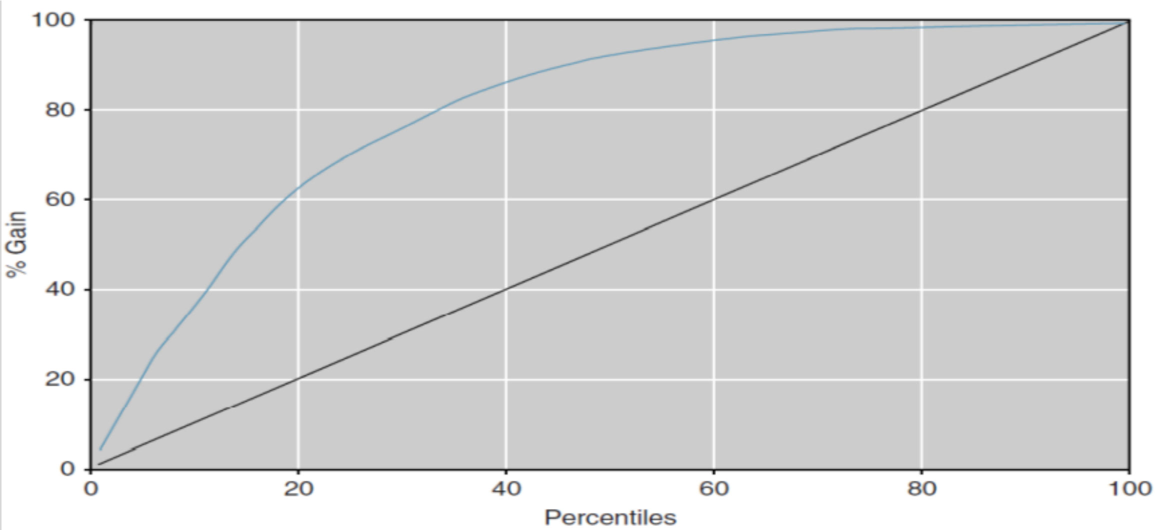
- For classification models, lift is a concept, which seeks to compare response rates with and without using classification model
- Lift charts and gains charts are graphical evaluative methods for assessing and comparing the usefulness of classification models.
- Suppose financial firm is interested in identifying high-income persons for targeted marketing campaign.
- Build a model to predict which contacts have high income, and restrict canvassing to these contacts.

- A good classification model should identify in its positive classifications, a group that has a higher proportion of positive "hits" than database as a whole.
- The concept of lift quantifies this.
- Define lift as proportion of true positives, divided by the proportion of positive hits in the data set overall.
- When calculating lift, software will first sort records by probability of being classified positive.
- The lift is then calculated for every sample size from n=1 to n=the size of the data set.



Lift chart for model 1: strong lift early, then falls away rapidly.

- Lift charts are often presented in their cumulative form, where they are denoted as cumulative lift charts, or gains charts.
- Number of cumulative actual events in each percentile of data set



Gains chart for model 1.

- Lift charts and gains charts can also be used to compare model performance.