Model Evaluation Techniques

Introduction

- Model Evaluation is an integral part of the model development process.
- It helps to find the best model that represents our data and how well the chosen model will work in the future.
- Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and overfitted models.

Introduction Cont…

- Validation is the process of assessing how well your mining models perform against real data.
- It is important that you validate your mining models by understanding their quality and characteristics before you deploy them into a production environment.

Introduction Cont…

- Evaluating a model is a core part of building an effective machine learning model.
- There are several evaluation metrics, like confusion matrix, cross-validation, AUC-ROC curve, etc.
- Different evaluation metrics are used for different kinds of problems

Introduction Cont…

Methods for Testing and Validation of Data Mining Models

There are many approaches for assessing the quality and characteristics of a data mining model.

- Use various measures of statistical validity to determine whether there are problems in the data or in the model.
- Separate the data into training and testing sets to test the accuracy of predictions.
- Ask business experts to review the results of the data mining model to determine whether the discovered patterns have meaning in the targeted business scenario

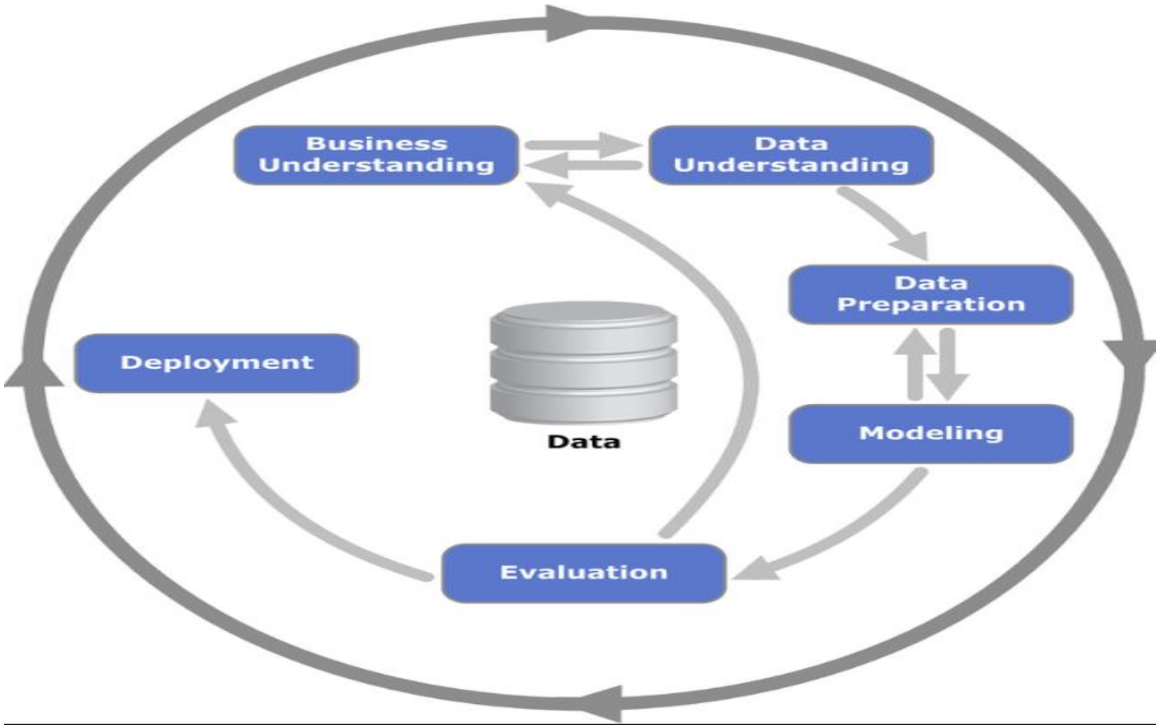
Introduction Cont…

Tools for Testing and Validation of Mining Models

- Partitioning data into testing and training sets.
- Filtering models to train and test different combinations of the same source data.
- Measuring lift and gain. A *lift chart* is a method of visualizing the improvement that you get from using machine learning algorithms, when you compare it to random guessing.
- Performing cross-validation of data sets.
- Generating classification matrices. These charts sort good and bad guesses into a table so that you can quickly and easily gauge how accurately the model predicts the target value.
- Creating scatter plots to assess the fit of a regression formula.
- Creating profit charts that associate financial gain or costs with the use of a mining model, so that you can assess the value of the recommendations.

Introduction Cont…

- **Cross-industry standard process for data mining**, known as **CRISP-DM**, is an open **standard process** model that describes common approaches used by data mining experts.
- It is the most widely-used analytics model **for data science projects.**



Introduction Cont…

- By the time we arrive at evaluation phase, modeling phase has already generated one or more candidate models.
- Critical importance that models are evaluated before deployment
- Deployment of models usually represents a capital expenditure and investment on the part of the company.
- If the models in question are invalid, then the company’s time and money are wasted.

Introduction Cont…

- Nestled between the modeling and deployment phases comes the **crucial evaluation phase**
- we examine model evaluation techniques for each of the six main tasks of data mining:
    - ✓ description,
    - ✓ estimation,
    - ✓ prediction,
    - ✓ classification,
    - ✓ clustering, and
    - ✓ association

MODEL EVALUATION TECHNIQUES FOR THE DESCRIPTION TASK

- Learned how to apply EDA to learn about salient characteristics of a data set
- Powerful technique for applying the descriptive task of data mining
- Evaluating the efficacy of these techniques can be elusive
- The watchword is common sense
- If insists on using a quantifiable measure to assess description, then one may apply the minimum descriptive length principle
- Best representation of a model or body of data is the one that minimizes the information required to encode model

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

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

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 = $\sqrt{MSE}$.

```
Regression Analysis: Rating versus Sugars

The regression equation is
Rating = 59.9 - 2.46 Sugars


76 cases used, 1 cases contain missing values


Predictor       Coef   SE Coef       T      P
Constant      59.853     1.998   29.96  0.000
Sugars       -2.4614    0.2417  -10.18  0.000

S = 9.16616    R-sq = 58.4%    R-sq(adj) = 57.8%


Analysis of Variance

Source          DF       SS      MS       F      P
Regression       1   8711.9  8711.9  103.69  0.000
Residual Error  74   6217.4    84.0
Total           75  14929.3
```

Regression results, with MSE and *s* indicated (Minitab regression output)

- **MSE =84.0 and s=9.16616**
- s=9.16616 indicate estimated prediction error
- Is this good enough to proceed to model deployment? Depends on business research problem.
- Prediction error is too large to consider for deployment

- 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

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

where SST(Sum of Squares Total) is give by

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

SSR (Sum of Squares Regression)

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

- 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:

$$\boxed{\text{Mean absolute error} = \text{MAE} = \frac{\sum |y_i - \hat{y}_i|}{n}}$$

- To calculate MAE analyst may perform following steps:

CALCULATING THE MEAN ABSOLUTE ERROR (MAE)

1. Calculate the estimated target values, $\hat{y}_i$.
2. Find the absolute value between each estimated value, and its associated actual target value, y_i, giving you $|y_i - \hat{y}_i|$.
3. Find the mean of the absolute values from step 2. This is *MAE*.

MODEL EVALUATION MEASURES FOR THE CLASSIFICATION TASK

- How do we assess how well our classification algorithm is functioning?
- Which evaluative methods should we use to assure ourselves that classifications made by our data mining algorithm are effective and accurate?
- 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.

MODEL EVALUATION MEASURES FOR THE CLASSIFICATION TASK Cont…

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

Contingency table for the C5.0 model

		Predicted Category		
		50 K	> 50 K	Total
Actual category	50 K	18,197	819	19,016
	> 50 K	2561	3423	5984
	Total	20,758	4242	25,000

MODEL EVALUATION MEASURES FOR THE CLASSIFICATION TASK Cont…

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.
- 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.

ACCURACY AND OVERALL ERROR RATE

- *accuracy is given by,*

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

- *overall error rate is given by,*

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

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

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{N} = \frac{18,197 + 3423}{25,000} = 0.8648$$

$$\text{Overall error rate} = 1 - \text{Accuracy} = \frac{\text{FN} + \text{FP}}{N} = \frac{2561 + 819}{25,000} = 0.1352$$

SENSITIVITY AND SPECIFICITY

Next, we turn to sensitivity and specificity, defined as follows:

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

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

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

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Total actually positive}} = \frac{\text{TP}}{\text{TAP}} = \frac{3423}{5984} = 0.5720$$

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Total actually negative}} = \frac{\text{TN}}{\text{TAN}} = \frac{18,197}{19,016} = 0.9569$$

SENSITIVITY AND SPECIFICITY

- In some fields, such as information retrieval, sensitivity is referred to as *recall*.
- Of course, a perfect classification model would have sensitivity=1.0=100%.
- A null model which simply classified all customers as positive would also have sensitivity=1.0.
- Clearly, it is not sufficient to identify the positive responses alone.
- A good classification model should have acceptable levels of both sensitivity and specificity.
- What constitutes acceptable varies greatly from domain to domain.

FALSE-POSITIVE RATE AND FALSE-NEGATIVE RATE

- Our next evaluation measures are *false-positive rate* and *false-negative rate*.
- These are additive inverses of sensitivity and specificity

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

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

For our example, we have

$$\text{False positive rate} = 1 - \text{specificity} = \frac{\text{FP}}{\text{TAN}} = \frac{819}{19,016} = 0.0431$$

$$\text{False negative rate} = 1 - \text{sensitivity} = \frac{\text{FN}}{\text{TAP}} = \frac{2561}{5984} = 0.4280$$

- Our low false-positive rate of 4.31% indicates that we incorrectly identify actual low-income customers as high income only 4.31% of the time.
- The much higher false-negative rate indicates that we incorrectly classify actual high income customers as low income 42.80% of time.

PROPORTIONS OF TP, TN, FP, AND FN

- Our next evaluation measures are proportion of true positives and proportion of true negatives, and are defined as follows:

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

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

For our income example, we have

$$\text{Proportion of true positives} = \text{PTP} = \frac{\text{TP}}{\text{TPP}} = \frac{3423}{4242} = 0.8069$$

$$\text{Proportion of true negatives} = \text{PTN} = \frac{\text{TN}}{\text{TPN}} = \frac{18,197}{20,758} = 0.8766$$

- That is, probability is 80.69% that a customer actually has high income, has classified it **as high income.**
- While probability is 87.66% that a customer actually has low income, classified it **as low income**.

PROPORTIONS OF TP, TN, FP, AND FN Cont…

- The proportion of false positives and proportion of false negatives, which are additive inverses of proportion of true positives and proportion of true negatives, respectively.

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

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

$$\text{Proportion of false positives} = 1 - \text{PTP} = \frac{\text{FP}}{\text{TPP}} = \frac{819}{4242} = 0.1931$$

$$\text{Proportion of false negatives} = 1 - \text{PTN} = \frac{\text{FN}}{\text{TPN}} = \frac{2561}{20,758} = 0.1234$$

- 19.31% likelihood that low income customer, classified it as high income.
- There is 12.34% likelihood high income customer, classified it as low income.

(Note: TPP-True Positive Proposition || TPN = Total Predicted Negative)

- 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.

PROPORTIONS OF TP, TN, FP, AND FN Cont…

- As an aside, in the parlance of hypothesis testing, as the default decision is to find that applicant has low income, we would have the following hypotheses:

  H_0: income ≤ 50,000

  H_a: income > 50,000

  where H_0 represents **default, or null**, hypothesis, and

  H_a represents **alternative hypothesis**, which requires evidence to support it.

- 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.

MISCLASSIFICATION COST ADJUSTMENT TO REFLECT REAL-WORLD CONCERNS

- Consider this situation from the standpoint of the lending institution.
- Which error, a false negative or a false positive, would be **more damaging** from lender's point of view?
- If lender commits a **false negative**, an applicant who had high income gets turned down for a loan: **an unfortunate but not expensive mistake.**
- However, if the lender commits a **false positive**,
  - an applicant who had low income would be awarded loan
  - increases chances that applicant will default on loan (**expensive**).
- Lender would **consider false positive** to be **more damaging**.
- Prefer to **minimize proportion of false positives**.

- The analyst would therefore adjust **C5.0 algorithm's misclassification cost matrix** to reflect lender's concerns.
- Suppose, analyst increased false positive cost from 1 to 2, while the false negative cost remains at 1.
- False positive would be considered twice as damaging as a false negative.
- How would you expect **misclassification cost adjustment to affect the performance of the algorithm?**
  - ✓ Proportion of false positives should decrease, since the cost of making such an error has been doubled.
  - ✓ Proportion of false negatives should increase, because fewer false positives usually means more false negatives.
  - ✓ Sensitivity should decrease.
  - ✓ Specificity should increase.

MISCLASSIFICATION COST ADJUSTMENT TO REFLECT REAL-WORLD CONCERNS Cont…

Contingency table after misclassification cost adjustment

		Predicted Category		
		≤ 50 K	> 50 K	Total
Actual category	≤ 50 K	18,711	305	19,016
	> 50 K	3307	2677	5984
	Total	22,018	2982	25,000

Comparison of evaluation measures for CART models with and without misclassification costs (better performance in bold)

Evaluation Measure	CART Model	
	Model 1: Without Misclassification Costs	Model 2: With Misclassification Costs
Accuracy	**0.8648**	0.8552
Overall error rate	**0.1352**	0.1448
Sensitivity	**0.5720**	0.4474
False-positive rate	**0.4280**	0.5526
Specificity	0.9569	**0.9840**
False-negative rate	0.0431	**0.0160**
Proportion of true positives	0.8069	**0.8977**
Proportion of false positives	0.1931	**0.1023**
Proportion of true negatives	**0.8766**	0.8498
Proportion of false negatives	**0.1234**	0.1502

DECISION COST/BENEFIT ANALYSIS

- 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.

DECISION COST/BENEFIT ANALYSIS Cont…

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

DECISION COST/BENEFIT ANALYSIS Cont…

Contingency table for the C5.0 model

		Predicted Category		
		50 K	> 50 K	Total
Actual category	50 K	18,197	819	19,016
	> 50 K	2561	3423	5984
	Total	20,758	4242	25,000

Contingency table after misclassification cost adjustment

		Predicted Category		
		≤ 50 K	> 50 K	Total
Actual category	≤ 50 K	18,711	305	19,016
	> 50 K	3307	2677	5984
	Total	22,018	2982	25,000

- 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)**

LIFT CHARTS AND GAINS CHARTS Cont…

- 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.

LIFT CHARTS AND GAINS CHARTS

- Define lift as proportion of true positives, divided by the proportion of positive hits in the data set overall.

$$\text{Lift} = \frac{\text{Proportion of true positives}}{\text{Proportion of positive hits}} = \frac{\text{TP}/\text{TPP}}{\text{TAP}/N}$$

Now, earlier we saw that, for model 1,

$$\text{Proportion of true positives} = \text{PTP} = \frac{\text{TP}}{\text{TPP}} = \frac{3423}{4242} = 0.8069$$

$$\text{Proportion of positive hits} = \frac{\text{TAP}}{N} = \frac{5984}{25,000} = 0.23936$$

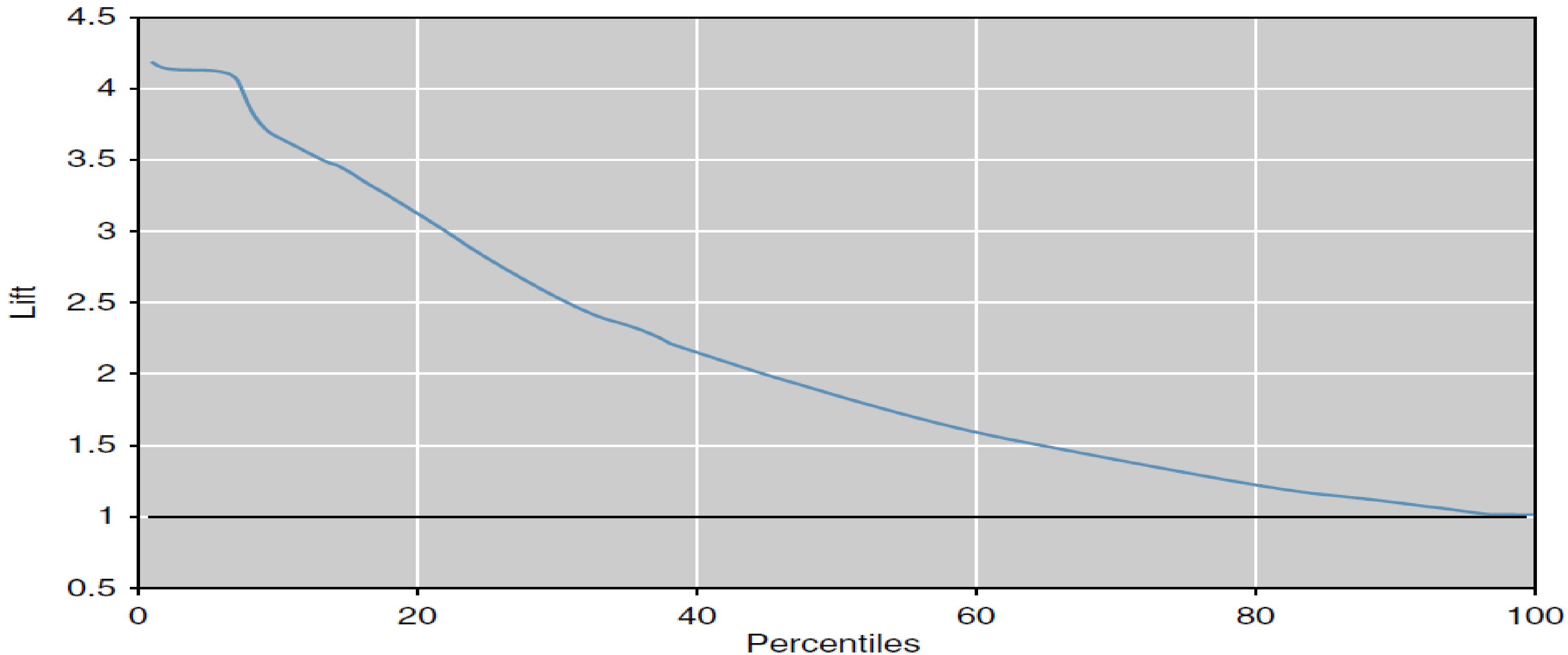
Thus, the lift, measured at the 4242 positively predicted records, is

$$\text{Lift} = \frac{0.8069}{0.23936} = 3.37$$

Lift is a function of sample size, the lift of 3.37 for model 1 was measured at n=4242 records.

LIFT CHARTS AND GAINS CHARTS Cont…

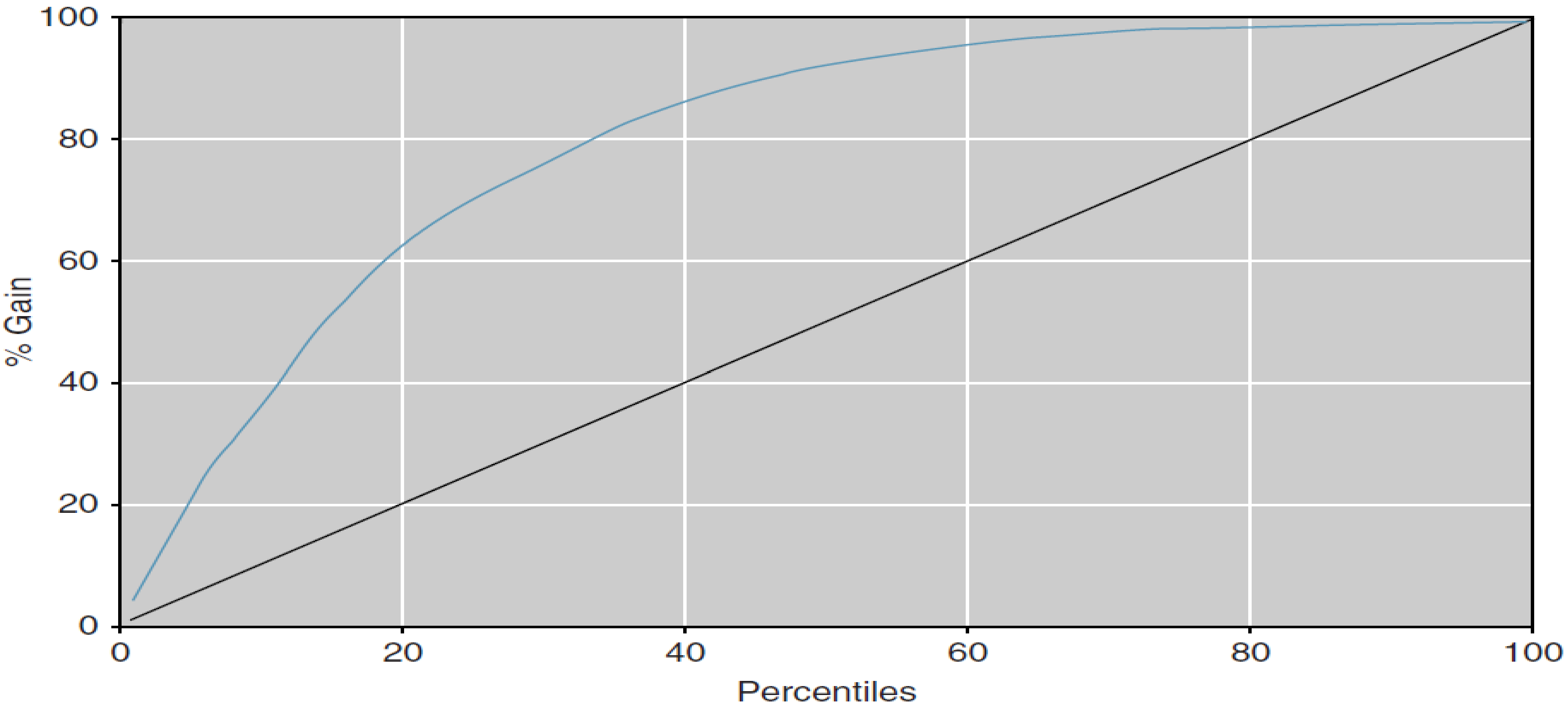
- When calculating lift, software will first sort records **by probability of being classified positive**.
- The lift is then calculated for every sample size from **n=1 to n**=the size of the data set.



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

LIFT CHARTS AND GAINS CHARTS Cont…

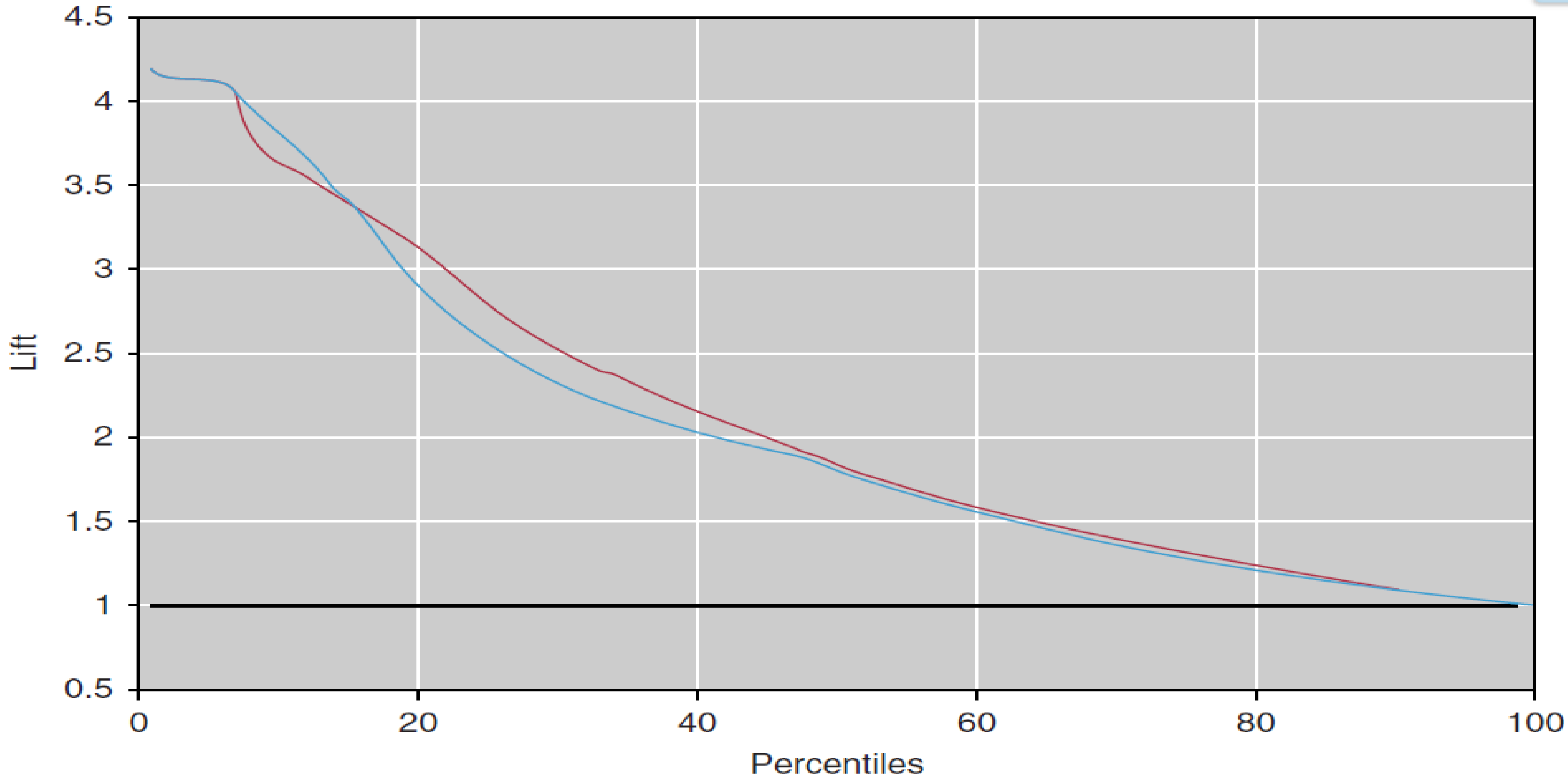
- Lift charts are often presented in their cumulative form, where they are denoted as cumulative lift charts, or gains charts.
- Number of cumulative actual events in each percentile of data set



Gains chart for model 1.

LIFT CHARTS AND GAINS CHARTS Cont…

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



Combined lift chart for models 1 and 2.

INTERWEAVING MODEL EVALUATION WITH MODEL BUILDING

- Recommend that model evaluation become a nearly **"automatic"** process, performed to a certain degree whenever a new model is generated.
- Therefore, at any point in the process, we may have an accurate measure of the quality of the current or working model.
- Therefore, it is suggested that model evaluation be interwoven seamlessly.
- Performed on models generated from each of the training set and the test set.
- For example, when we adjust provisional model to minimize the error rate on the test set, we may have at our fingertips the evaluation measures such as sensitivity and specificity, along with the lift charts and the gains charts.
- These evaluative measures and graphs can then point the analyst in the proper direction for best improving any drawbacks of working model.

CONFLUENCE OF RESULTS: APPLYING A SUITE OF MODELS

- In model selection, analyst should not depend solely on a single prediction model.
- Instead, analyst should seek a confluence of results from a suite of different data mining models.
- Can use algorithms like CART, C5.0, and neural network algorithm to identify most influential variables in dataset.

CONFLUENCE OF RESULTS: APPLYING A SUITE OF MODELS Cont…

Most important variables for classifying income, as identified by CART, C5.0, and the neural network algorithm

CART	C5.0	Neural Network
Marital_Status	*Capital-gain*	*Capital-gain*
Education-num	*Capital-loss*	*Education-num*
Capital-gain	*Marital_Status*	*Hours-per-week*
Capital-loss	*Education-num*	*Marital_Status*
Hours-per-week	*Hours-per-week*	*Age*
		Capital-loss

- All three algorithms identify ***Marital_Status, education-num, capital-gain, capital-loss*, and *hours-per-week*** as the most important variables, except for the **neural network**, where ***age*** snuck in past *capital-loss*.
- None of the algorithms identified either ***work-class* or *Gender*** as important variables, and only the neural network identified *age* as important.
- The algorithms agree on various ordering trends, such as ***education-num* is more important than *hours-per-week*.**

Thank You !!!