Overview of Business Metrics for Model Assesment

Dr. Goutam Chakraborty

Outline

- Discuss various metrics used to assess predictive model performance
- Metrics that are routinely used by managers/marketers
  - Deciles, gains, cumulative gains, lift and cumulative lift…
- Metrics used by statisticians/data scientists
  - ASE, Misclassification Rate, Hit-ratio, Sensitivity, Specificity, Recall, Precision, F1 Score..
  - ROC curve and area under this curve (AUC), K-S statistic …

Predictive Models (Recap)

- Aim is to build a mathematical model predicting a target measure of interest. Also, referred to as *supervised learning*
- Two types of problems:
  - *Classification*: discrete (binary or nominal) target
  - *Prediction*: continuous target
  - Two types of methods for classification/prediction models:
    - Statistical (Regression : LR for classification, MR for prediction)
    - Machine Learning (Decision trees, Neural Net, others)
- One of our goals is to avoid overfitting:
  - Achieved via splitting of data into training and validation
  - Honest assessment of a model is seen on validation data

Metrics and Graphs for Model Assessment

- Most of these originated from “data base marketing or direct marketing” domain.
- Analysts in those domains have built predictive models and then applied those models to score customers/prospects to whom they want to send offers to for many years before “data science” became a buzzword!
- The metrics they use to assess models are
  - Decile analysis, Gains and Cumulative Gains, Lift and Cumulative Lift. etc.

How are Models Used by Marketers?

- Business objective: to send direct marketing offer to selected members of housefile (*not all of them)*
- Build a predictive model using a training/validation sample drawn from the housefile
  - Target variable: whether customer responded to last year's direct marketing offer
  - Assume:
    - Model is reasonable with accuracy as well as variables included.
    - Last year's **overall response rate was 3.85%**

Response Gains on Training Sample

- Score (predict) the training sample with the model built (probability of yes response).
- Rank the training sample from highest scores to lowest scores.
- Group the ranked scores into 10 *approximately* equal bins (each representing about 10% of data).
  - Bin 1 (first decile) will have the top 10% scores of your training sample.
  - Bin 2 (second decile) will have next highest 10% scores of your training sample and so on.
- For each bin, calculate response, gain, lift and cumulative lift as shown next

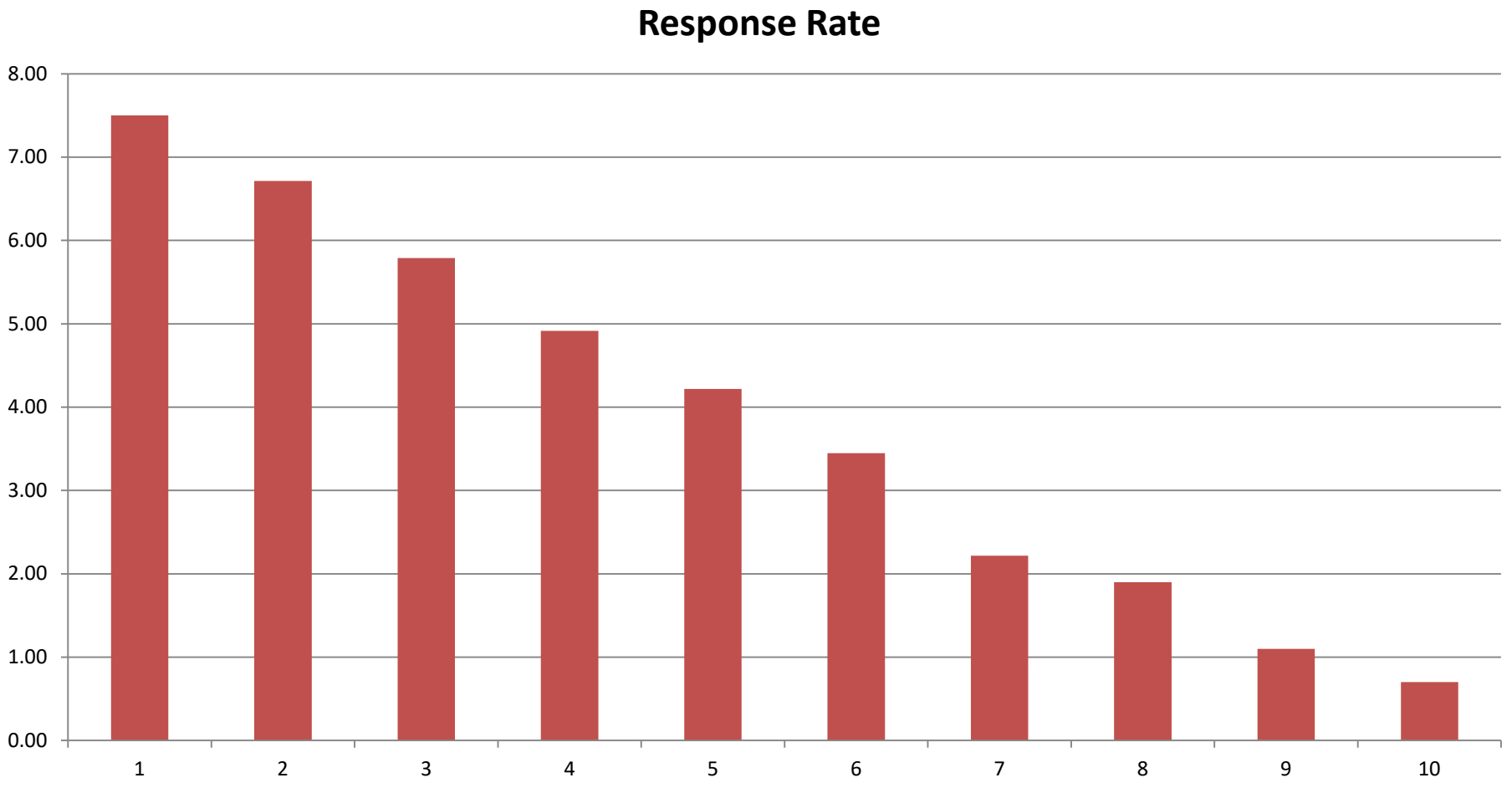
Formulas for Calculation

- Response rate in a decile = (number of responders in a decile /total number of sample in the decile)* 100
- Lift of a decile = (Response rate in a decile/ Overall response rate across all deciles)
  - Lift higher than 1 means…
  - Lift less than 1 means…
- Gain of a decile = {(Response rate in a decile – overall response rate)/overall response rate}*100
  - Positive gain means…
  - Negative gain means…

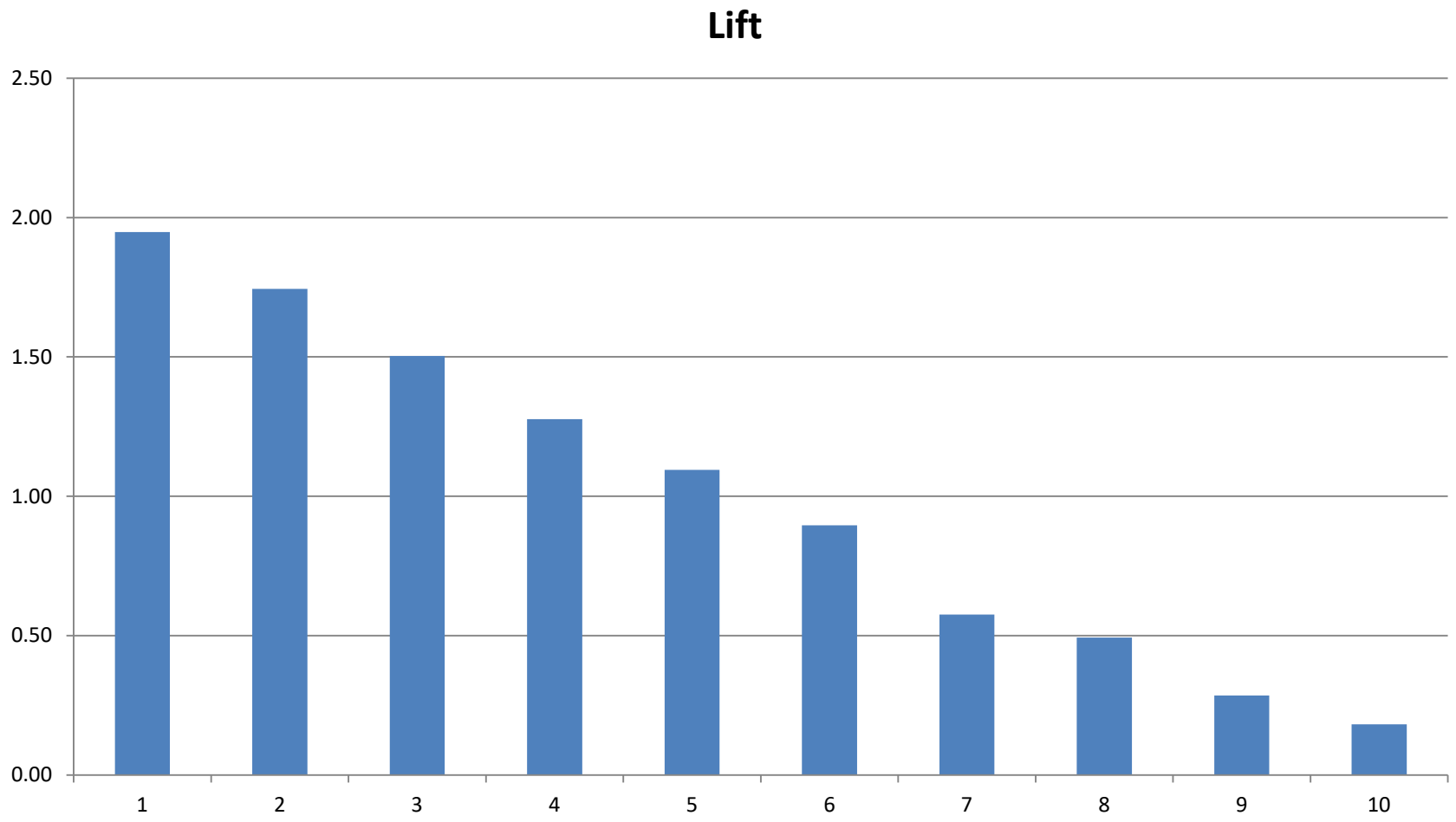
Calculations for Lift, Gains for Training Data

Decile	Score	Sample Count	No. of Responders	Response Rate (%)	Lift	Cumulative Lift	Gains
1	> 0.79	1,000	75	7.50	1.95	1.95	95
2	0.65-0.789	998	67	6.71	1.74	3.69	74
3	0.58-0.649	1,002	58	5.79	1.50	5.20	50
4	0.45-0.579	997	49	4.91	1.28	6.47	28
5	0.35-0.449	996	42	4.22	1.10	7.57	10
6	0.31-0.349	1,015	35	3.45	0.90	8.46	-10
7	0.25-0.309	992	22	2.22	0.58	9.04	-42
8	0.21-0.249	1,000	19	1.90	0.49	9.53	-51
9	0.15-0.209	1,000	11	1.10	0.29	9.82	-71
10	< 0.149	1,000	7	0.70	0.18	10.00	-82
Total		10,000	385	3.85			0

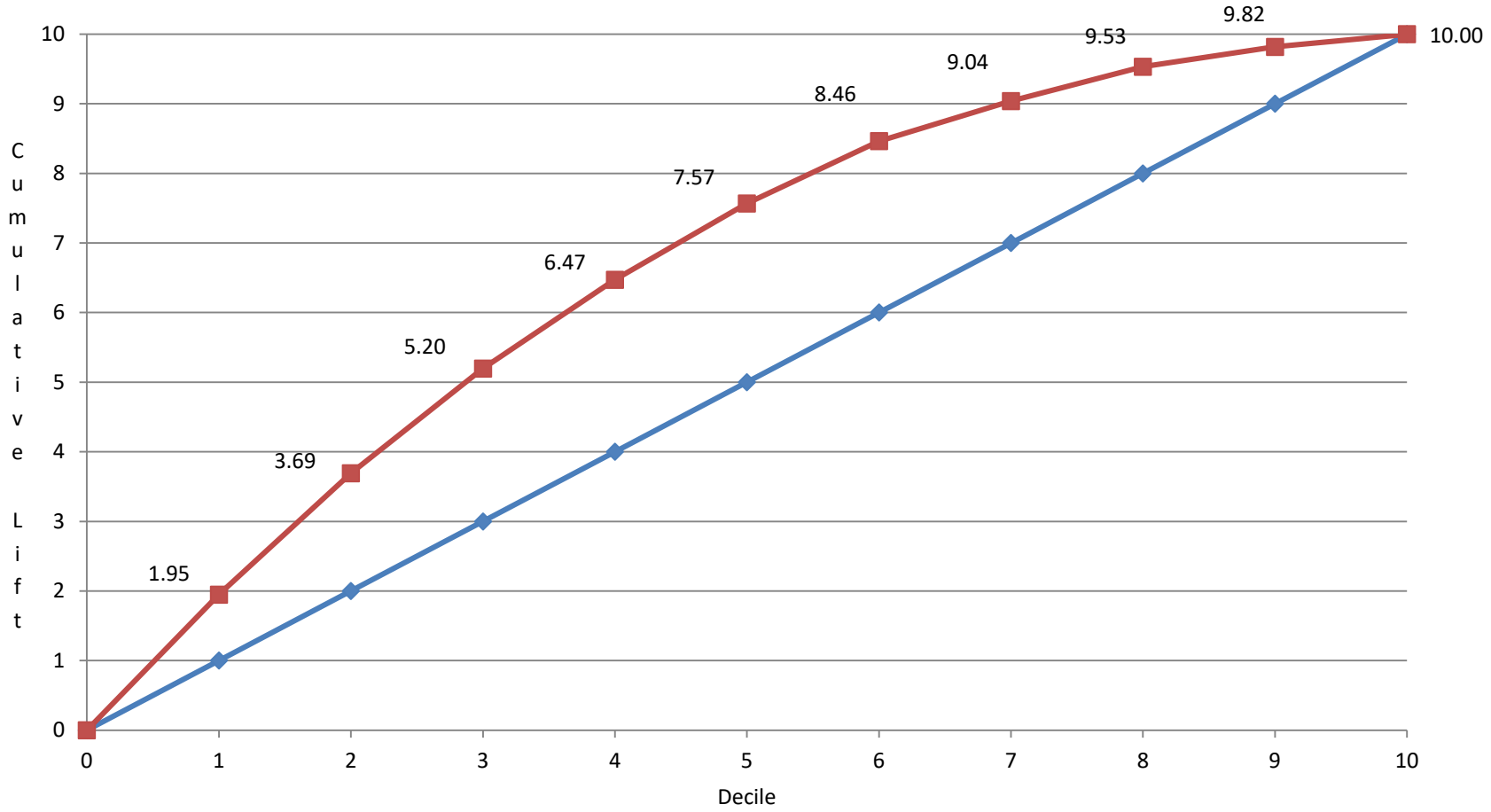
Response Rates Across Deciles



Lift Across Deciles



Cumulative Lifts Across Deciles



Cumulative lift at 3rd decile is 5.20 – what does that mean?

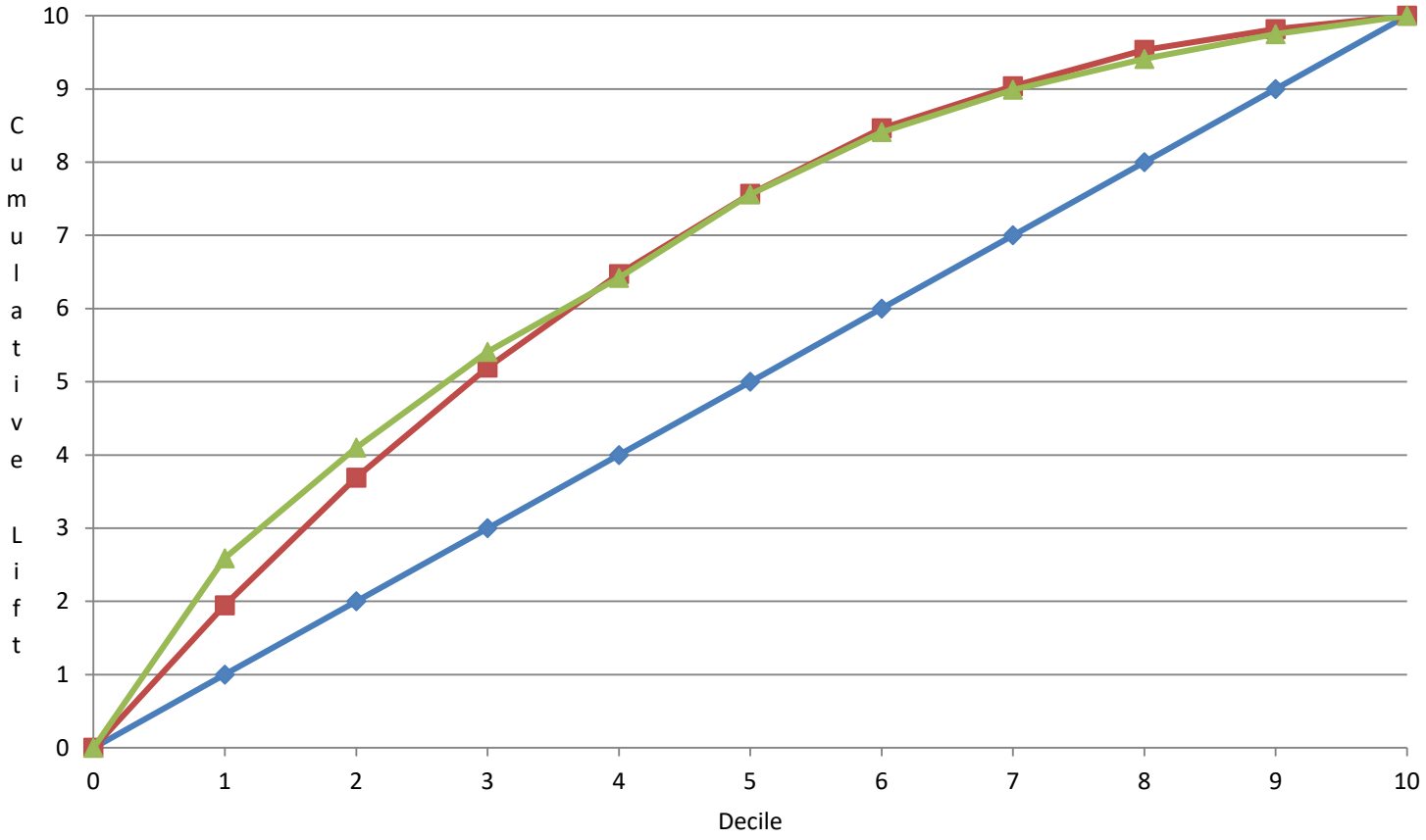
Response Gains on Validation Sample

- Score (predict) the validation data with the model built on training data.
- Rank the sample from highest scores to lowest scores (probability of yes response) .
- Define 10 bins *as determined by the training sample*. So, for example
  - Bin 1 will be scores more than 0.79.
  - Bin 2 will be scores between 0.65-0.789 and so on
- For each bucket, calculate response, gain, lift and cumulative lift *using response numbers in the validation sample*.
- Compare these numbers from validation sample with numbers from training sample.
  - You want to see stability between metrics from training and validation sample
  - But, metrics will typically …..

Comparing Training versus Validation

- Main purpose of a validation sample is to confirm the results from the analysis done on calibration samples (*to avoid overfitting of models*).
  - If you have identified multiple candidate models (each about equally good) using training sample, you can apply each of those on the validation sample and choose the one that *performs the best* in validation.
  - Instead of defining best performance *being best in overall predictions*, often direct marketers will define *best performance as in top 2, 3 or 4 deciles*!

Multiple Models Comparison



Which is a better model?

Gains, Lifts for Continuous Variables

- In my example, I have used a binary target variable.
  - But, the concept of gains, lift are applied even when the dependent variable is continuous (such as $ amount order, or profit from a campaign).
  - In that case, **instead of response rate**, the relevant metric in a decile becomes the **average $ order** in a decile or **average profit** in a decile.
    - The baseline number is the average $ order or average profit across all the deciles.
    - The formulas for gains, lifts need to be adjusted using the above metrics (instead of response or response rate)