# Introduction to Predictive Modeling

LITYX™

EVALUATE. PREDICT. OPTIMIZE.

# Notices



The information in this document is meant to be delivered in a classroom setting, and is not complete without the associated interactions with the instructor.

**SERVICES AND SOLUTIONS**

Lityx offers world-class analytic services and solutions to help our clients become analytically-driven, successful organizations. Please see www.lityx.com for more information.

**CONTACT INFORMATION**

By phone
888-LITYX-IQ (phone)
717-388-4326 (fax)

On the web
info@lityx.com
www.lityx.com/contact-us.aspx

Postal mail
100 West Road, Suite 300
Towson, MD 21204

# Agenda

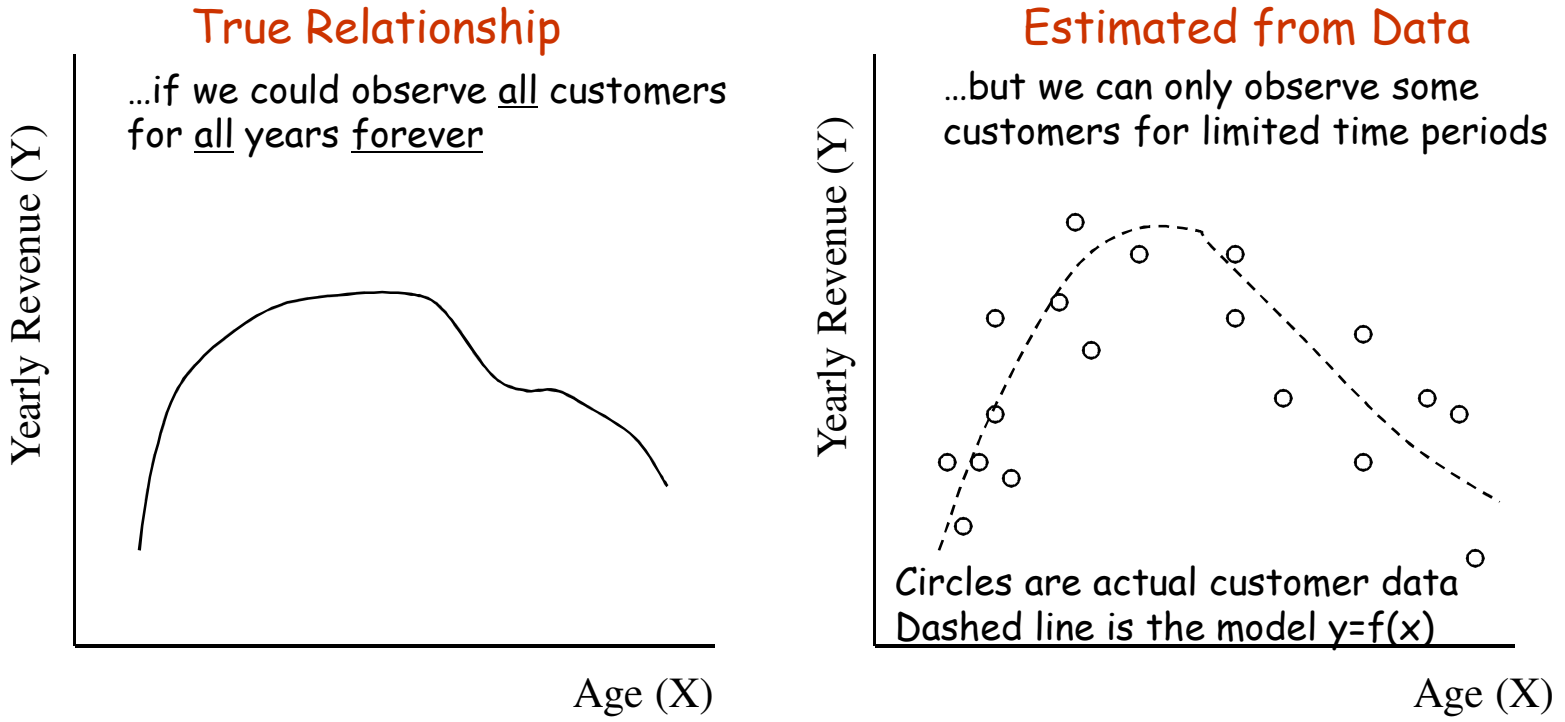
- Basics of Modeling
- The modeling process

# What are predictive models?

- Attempt to describe the mathematical relationship between a dependent variable (y) and independent variables (x1,x2,….)
- Y (also called the response variable) is a customer attribute you would like to predict, like response/no response or yearly revenue.
- The X's (also called predictors) are attributes that may be useful in making predictions of Y
  - E.g., if Y is response/no response to a campaign, X's might be Age, Income, Previous Customer?, etc.
- Ultimately, the mathematical relationship is in the form of a mathematical function like y=f(x1,x2,…)
  - The function can range from simple to very complex

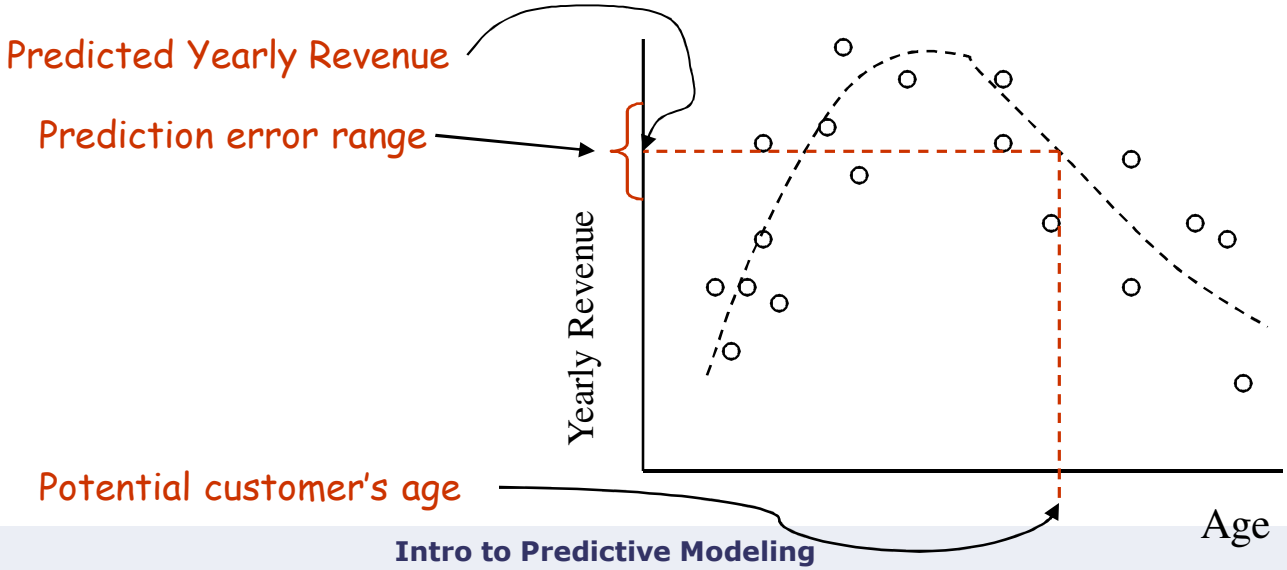
# Overview of Model Building

- Use historical customer data to understand the true relationship between Y and X's.
- The model is a simplification of the true relationship.



# Overview of Model Building

- Notice that we still do a pretty good job of estimating the true relationship between revenue and age.
- Generally, the more data we have, the better we do.
- The model now allows us to predict yearly revenue for a potential new customer whose age we know.
  - We realize that there is some "prediction error" with this prediction.



# Overview of Model Building

- A better understanding of the response variable (and therefore better predictions) typically come from more than one predictor variable
  - Maybe 10's or 100's
  - Becomes impossible to picture the relationship nicely in a graph
  - We can now only write the full relationship down using a complex mathematical formula to symbolize the model
- Finding the model and determining how good it is are what the modeling process is all about
  - Not easy, often requires many iterations
  - There are many different techniques for determining a model
    - All give different answers depending on many aspects of the modeling process

# Overview of Model Building

- Model building is a process
  - Some aspects can be automated with the help of tools
  - Overall, requires a lot of attention and iteration from experienced business people and modelers
- Good models are easy to build; great models are harder to build
  - Incremental value gained from better models can be large when scaled to the full prospect or customer base
- Requires understanding of many things
  - The business problem and the data
  - Modeling algorithms and parameters
  - Software
  - Implementation issues

# Why build models?

- Want to predict customer behavior so we can make more efficient, profitable business and marketing decisions
- Two assumptions:
  - Past behavior can help us predict future behavior
  - Customers that have similar characteristics behave similarly
- Customer data can be used to quantify customer attributes and behavior
- Models can be built using these attributes to learn customer behavior patterns

# Examples of Models

- Marketing
  - Predict response to a campaign for customers or prospects
  - Predict likelihood of a second stay following the first for a hotel chain
  - Predict cross-sell opportunities based on product or category preferences
  - Predict number of purchases per year
  - Predict offer amount most likely to entice response but still be profitable
  - Predict revenue or profitability over a year
- Non-marketing
  - Percent gain for a stock ticker over next year
  - Likelihood of a credit card transaction being fraudulent
  - Likelihood of loan default
  - Number of bears living in a certain habitat 5 years from now

# Typical Attributes Used as Predictors

- Customer transaction history
  - Previous stays and stay lengths, previous promotion responses, previous packages purchased
- Customer demographics
  - Age, income, zip code, gender
- Appended data
  - Customer psychographics, lifestyles, attitudes
  - Often only available at block or census tract levels
- Marketing data
  - Product data, campaign data, offer data
- Etc.
  - Many possibilities – any transactional, customer, or household level data

# Classifying Models

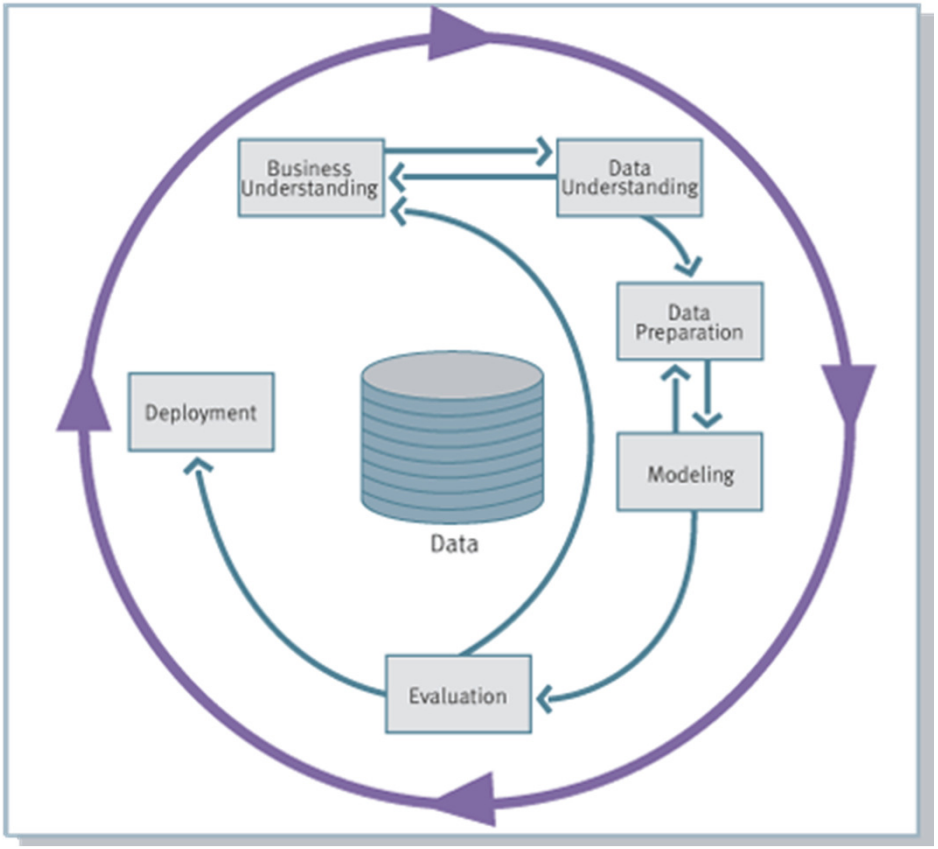
- Numeric versus Classification predictions
  - Numeric prediction model – response variable is a numeric quantity
    - E.g., revenue, profit
  - Classification model – response variable is categorical
    - E.g., response/no response, renew/doesn't renew
    - Actually we are predicting the probability of one of the categories (e.g., probability of responding to the campaign)
- Supervised versus Unsupervised models
  - Supervised – there is a response variable (like all examples so far)
  - Unsupervised – there is no response variable
    - Very different situation - not trying to predict a specific outcome
    - Goal – group similar customers together based only on predictor variables
    - Main example is customer segmentation

# Classifying Algorithms

- For classification or numeric prediction (or both)
- For supervised or unsupervised modeling
- Underlying assumptions of the algorithm
  - Some require data to have certain distributions, others do not
- Complexity of resulting models
- Number of algorithm parameters to fine-tune
  - Relates to ease of building models
- Length of time to build models
- Ease of interpreting results

# Typical Model Development Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment




http://www.crisp-dm.org

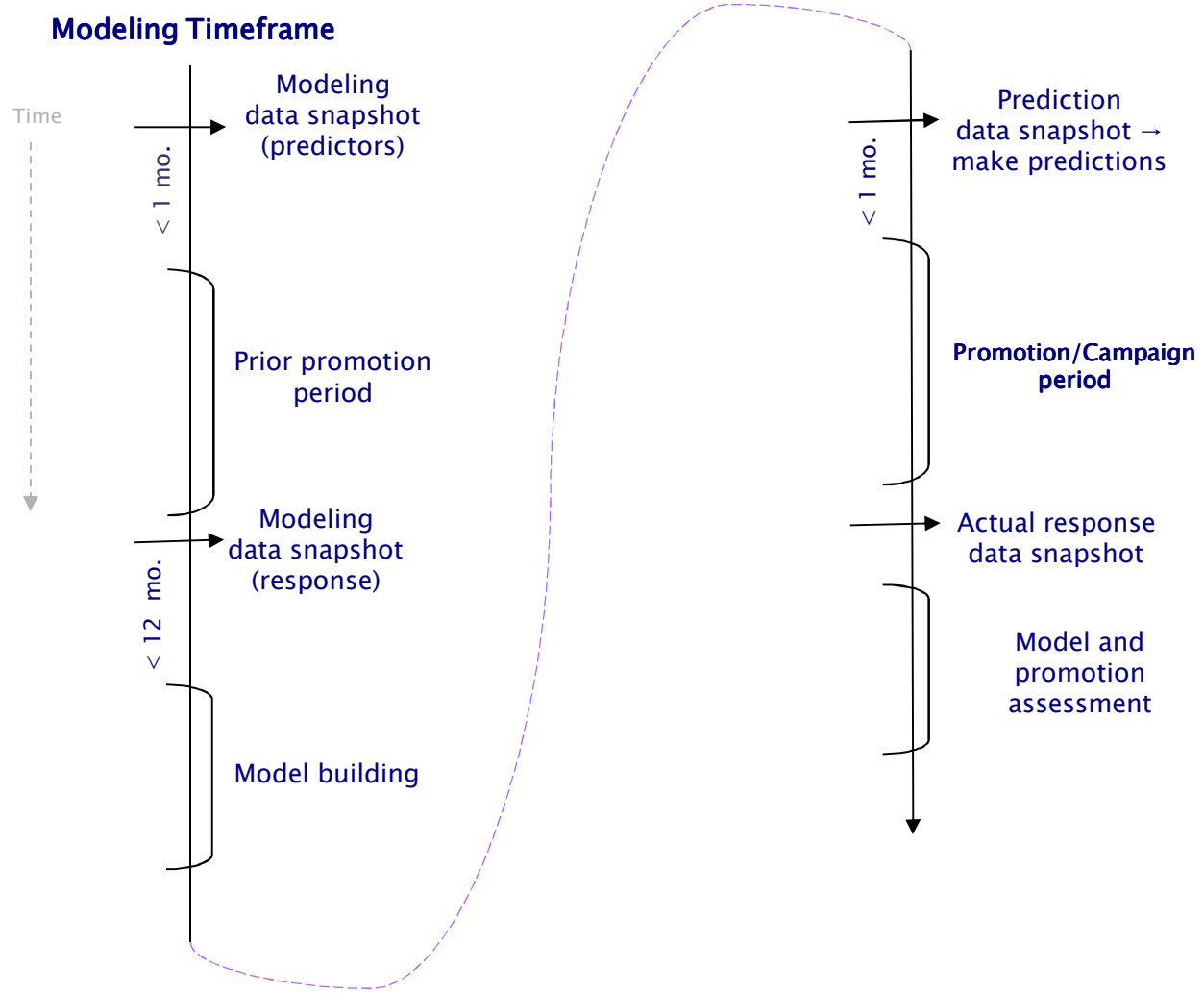
# Step 1: Business Understanding

- What are the goals from a business perspective?
  - Retain more customers
  - Grow the customer base
  - Make profitable customers more loyal
  - Increase response rates to campaigns/promotions
  - Reduce churn, increase loyalty
  - Etc.
- How will achievement be measured?
  - Profit/revenue increase
  - Increased market share
  - Number of customers
  - Retention rate
  - Etc.

# Step 2: Data Understanding

- Based on the business objectives, what data is available to us for modeling?
  - Attributes (both response and predictor variables)
  - Timeframe of available data
  - Quality of data available
  - Quantity of data available
- Different objectives require different views of data, different attributes, and different modeling techniques

# Data Timing Example (Response Model)



# Step 2: Data Understanding

- Retrieve the data from data sources
  - Internal databases/tables, external data, merges, etc.
  - Row/column flat file format for use in modeling tool
  - Sampling (same time and computational effort with little accuracy tradeoff)
    - Simple random sample (all records equally likely to be sampled)
    - Stratified sampling (some categories of data oversampled, e.g., higher percentage of responders than non-responders)
- Analyze data quality
  - Missing data for some attributes
  - Data errors, outliers
  - Data accuracy
  - Missing attributes

# Step 2: Data Understanding

- Data exploration
  - Interactive and iterative data analysis
  - Look for interesting patterns and correlations
  - Give initial assessment of important attributes for making predictions

# Step 3: Data Preparation

- Variable transformations
  - Reason 1: Business reasons
    - e.g., ProfitPerMonth = Total Profit / Number Months as Customer
    - e.g., Recency = Current Date – Date of last stay
  - Reason 2: Mathematical/modeling reasons (usually automated)
    - e.g., take log ratios of money variables
    - e.g., normalize all attributes to same scale
    - e.g., binary coding of categorical predictor variables
- Data cleaning
  - Remove, keep, or impute for missing values
  - Detect and fix/keep/remove outliers and bad data

# Step 3: Data Preparation

- Data reduction
  - In many situations, the number of attributes available is huge.
  - Modeling effort can grow exponentially with each additional variable.
  - Use techniques for reducing the number of attributes to be used for modeling.
    - Remove redundant or highly correlated variables.
    - Factor analysis or principal components analysis find a small number of transformations of original variables that contain most of the important information.
  - Benefits: Performance savings
  - Drawbacks: May reduce/remove important relationships
- Note: Some aspects of data prep can be integrated into the model building process (Step 4).
- Put into final format needed for modeling tool.

# Step 4: Modeling

- Now it's time to actually build the models
- Consists of a number of important sub-tasks

1) Select modeling algorithm(s) to use
2) Select parameter settings of the algorithms
    - A single algorithm is often run more than once using different parameter settings
3) Run the models ("fit the models", "build the models") using the tool
    - May build dozens of models depending on number of algorithms chosen and number of parameter settings for each

# Step 4-1: Select Modeling Algorithms

- May choose one, two, or many different algorithms to try
- Choice depends on many things:
  - Availability in your tool
  - Availability for the particular modeling problem
  - Analyst capabilities
  - Time
  - Ease of deployment
  - Need for interpretability
- Some tools automate some of this, some parts are still manual.

# Step 4-2: Select Parameter Settings

- A single algorithm actually consists of many possible models based on various parameter settings and model forms the analyst can choose.
- Algorithm parameters are one of two kinds:
  - Some parameters have to do with the mathematical form of the model itself.
  - Some parameters have to do with the details of how the model is built.
- In either case, changing a parameter setting usually results in a different model being built.
  - Note: for some algorithms, even not changing a parameter setting can result in a different model being built if you run it again!

# Step 4-2: Select Parameter Settings

- This is usually recommended only for advanced users
  - Most tools have reasonable default settings for parameters.
  - Most changes have little effect, but an advanced user can often find a combination of settings that can produce more accurate models.
  - Novice users can still play around with different settings, but might waste time if they don't really understand what they mean.

# Step 4-3: Run the Models

- This is always a computer-automated task.
  - Requires many complex mathematical calculations.
- Depending on the algorithms chosen, amount of data, number of variables, and other settings, this can take a second or two, or much, much longer.
  - I've run models that have taken days to finish, and have heard of some taking months.
- Result is a listing of completed models (i.e., their mathematical formulation) and their performance.

# Step 5: Evaluation (and Selection)

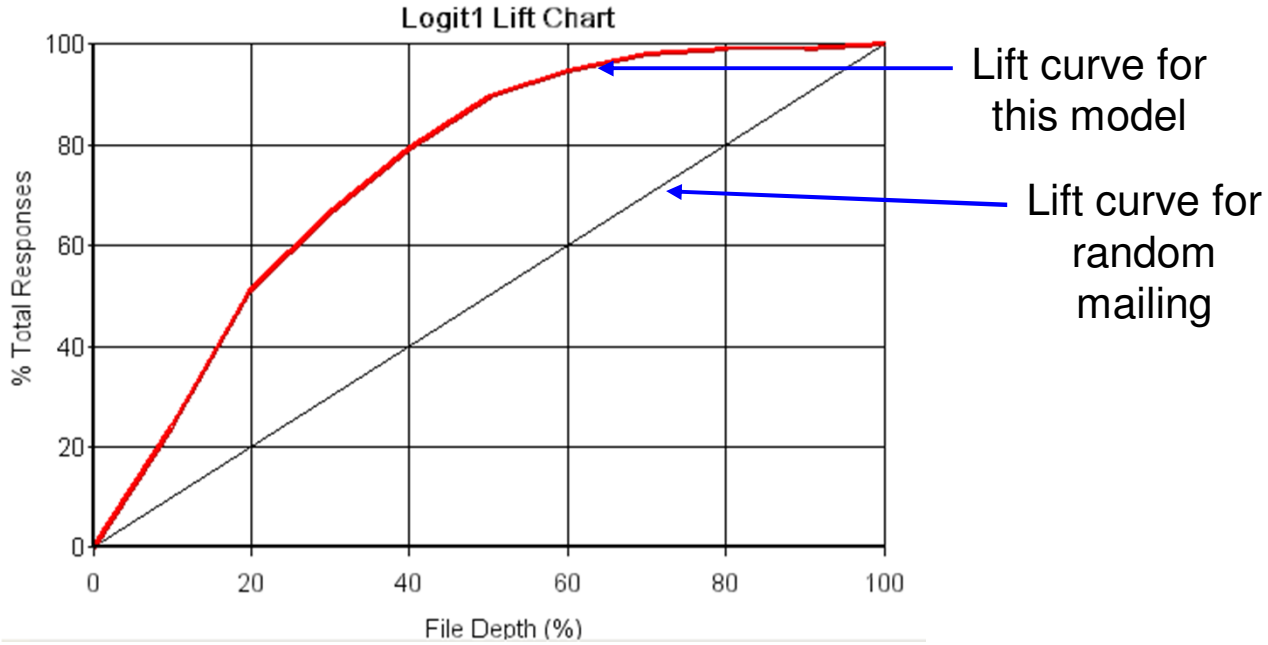
- Lots of models now built and finalized.
  - We now have the mathematical formula that represents each model
- Need to evaluate each model and select one (or more).
- Two steps:
  1) Estimate performance of the models
  2) Select one (or more) “winning” models

# Step 5-1: Estimate Model Performance

- Need to know how good each model is
  - Measure accuracy of predictions, response lift, revenue gain, etc.
- Usually performed simultaneously with the Model Building step within a software tool.
- Estimating performance has two parts:
  - Choice of performance measure:
    - Lift, percent accuracy, profit, mean squared error, etc.
  - Choice of performance estimation method
    - Holdout set, cross-validation, bootstrapping, etc.

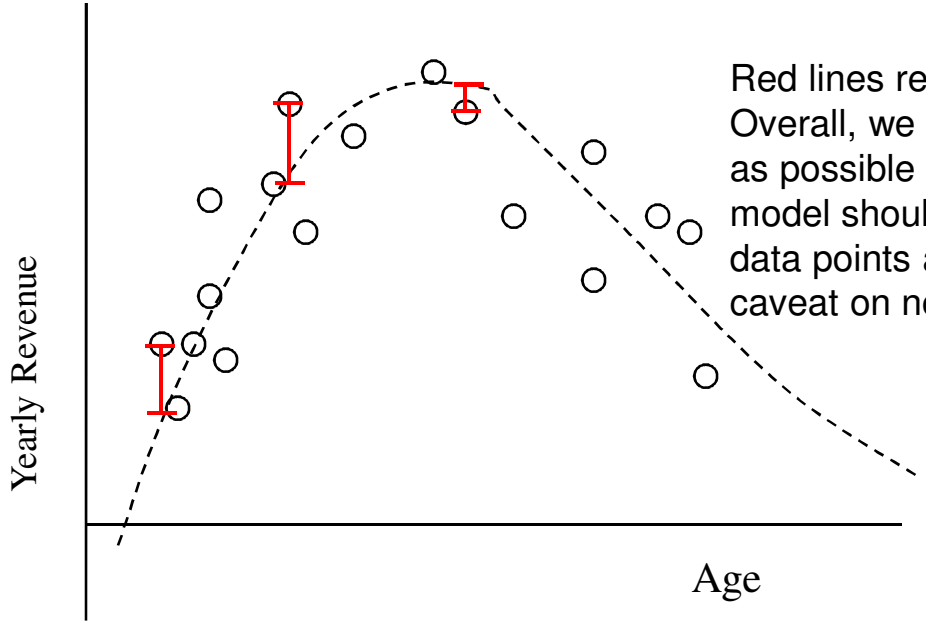
# Model Performance Measures

- Lift
  - Computes percent of responses found by model at each decile in the modeling file, and compares to percent of responses that would be returned by a random mailing.
  - Lift is measured as the total area between curves. High is good.



# Model Performance Measures

- For numeric prediction models, Root Mean Square Error (RMSE) and coefficient of determination ($R^2$) are also common accuracy measures.
  - They have to do with the amount of prediction error made by the model.




Red lines represent prediction errors. Overall, we want these to be as small as possible (in other words, our model should come as close to the data points as possible, except for caveat on next page).

# Model Performance - Caveat

- Seemingly, we would want a model with the best possible performance when measured on the modeling data.
  - Example: 100% lift in first decile for a response model, or zero error for a numeric prediction
- But, in reality, such a model is likely to be not very good, or even terrible. Usually, either one of two things can happen to put you in this situation:
  - Inappropriate modeling data leaves room for model to "cheat" and find a perfect or near perfect prediction.
  - Modeling algorithm (or mathematical form of the model) is too complex and can fit any dataset perfectly.
- This is called "overfitting" the model.

# Validating Model Performance

- To guard against overfitting, we typically use some of the modeling data to build the model, and the rest to validate its performance.
  - Gives us a way to see how well the model "generalizes" to data it has not seen yet.
  - If the modeling process came up with an overfit model, we will find out using the data set aside
  - Performance validated this way is called an "unbiased measure of performance" or "generalization performance".
- There are different techniques for doing this:
  - Holdout method (also called train/test method or out-of-sample method)
  - Cross-validation (more complex)
  - Bootstrapping

# Performance Validation Methods

- Holdout Method
  - Randomly select 70% (or other percentage) of data to build model.
  - Set aside (i.e., holdout) the other 30% and use it to measure performance of the built model.
- k-fold Cross-validation (CV)
  - Maximizes data efficacy of training/testing
  - Modeling data partitioned into k even groups (folds)
    - Leave one fold out, build model on all the rest
    - Measure performance of that model using the left-out fold
    - Do for each fold, leaving one out at a time
    - Average performance over each iteration
  - Train final model with all data

# Step 5-2: Select Model

- Based on different criteria available, pick best model
  - Performance is easiest criteria since it is numeric and easily ranks the models
  - Other things may be important
    - Does the model seem sensible?
    - Is it easy to implement?
    - Is it interpretable?
- May pick two or more good models and test each in a live setting or market test.

# Step 6: Deployment

- Now we need to implement (deploy) the chosen model
  - Export model formula to a database or other program
  - Score prospects or customers according to the model formula
  - Take appropriate marketing action based on these scores
    - Send/don't send promotion, take anti-churn action, etc.
  - Measure how well the model performed in the real-world live setting

# Iterative Nature of Entire Process

- After deploying model, we can measure how well it did
- Model performance typically degrades over time as customers, products, and markets change
- Need to continually re-assess live model performance and rebuild models periodically.

# Automation: Modeling as a Service

- Companies have come to use dozens or even hundreds/thousands of models to make predictions for a variety of business problems and situations.
- To scale to this level, models cannot be built and scored manually. Complex tools are needed to automate much of the process:
  - Data extraction and creation for modeling
  - Model building and testing automation
  - Model management and refresh
  - Model scoring
- This is the concept of analytics as a service… to give organizations the capability to do this without investing enormous dollars/time into tools, people, and process.

# The End

- If you have any questions, we'll be happy to help.

Contact Paul Maiste

maiste@lityx.com

www.lityx.com/contact