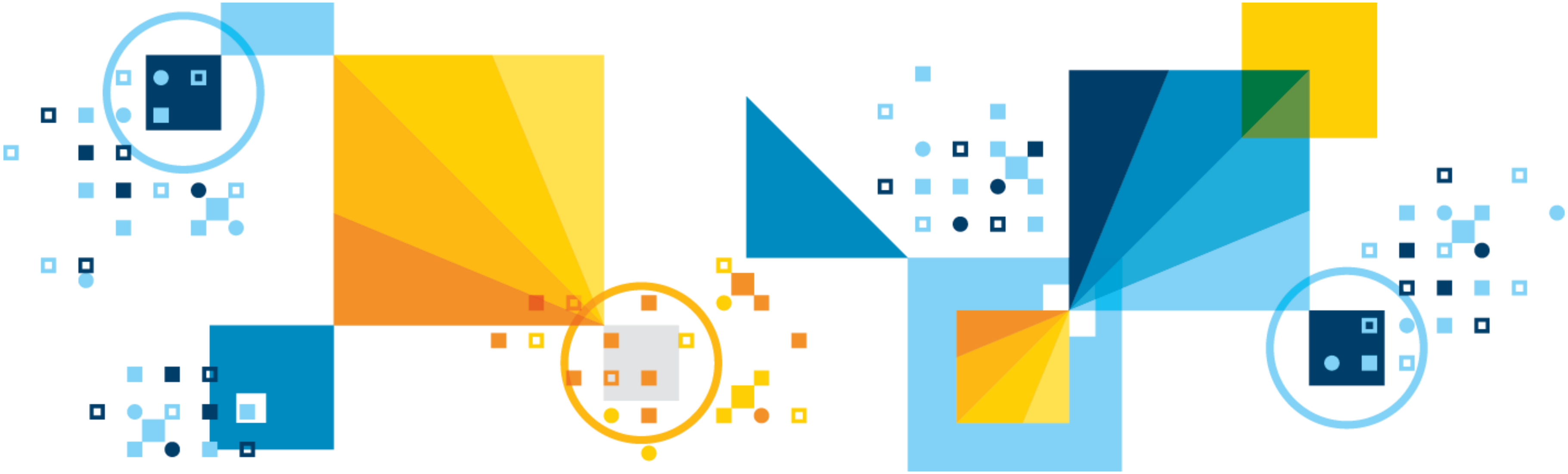
Predictive Modeling Fundamentals I
Lesson 4



Setting the Stage….

Why this is important to know…

1. Fundamental introduction to Data Mining and its application to business problems
2. Ability to utilize software tools for advanced analytics

After this session, you will be able to…

1. Understand common techniques and metrics for classification model evaluation
2. Apply predictive model on test and new data
3. Use SPSS Modeler to assess model performance and accuracy

Speaking to you today…

Armand Ruiz

Product Manager

Mikhail Lakirovich

Product Marketing Manager

Agenda

- Lesson 3 review: Core Data Mining Tools
- Lesson 3 review: Training and Testing
- Lesson 3 review: Sampling Data in Classification
- Metric for Performance Evaluation
- Accuracy as Performance Evaluation
- Overcoming Limitations of Accuracy Measure
- ROC Curves
- Lab 4

Lesson 3 Review: Core Data Mining Tools

- Supervised Learning
- Describes and distinguishes classes for future prediction (on new data) based on training data
- Classification & Prediction
- Common Methods: Decision Trees, Regression, Nearest Neighbors, Neural Networks
- Unsupervised Learning
- Analyzes data where labels are unknown to create groups/classes for objects that are similar to each other (within the group) but dissimilar to objects in other clusters
- Cluster analysis
- Common Methods: K-means, Hierarchical, Two-Step
- Association
- Analyzing data for events or instances that occur together (i.e. diapers and beer commonly purchased together)
- Association Rules
- Common Methods: Apriori, CARMA

Lesson 3 Review: Training & Testing

- Splitting the data set into Training and Testing
  - Approximately 66%-75% for training and 34%-25% for testing
- Training the model
  - On the data with existing classes – supervised learning
- Testing the model
  - On the portion of the data that was not included in the training phase
- Evaluating the model
  - Comparing the accuracy of the model on the training and testing sets
  - Accuracy rate is the percentage of sample that is correctly classified by the model
  - High accuracy for both training and testing data sets
  - High accuracy on training and low on testing -> overfitting problem
- Using the model
  - or classifying future or unknown objects

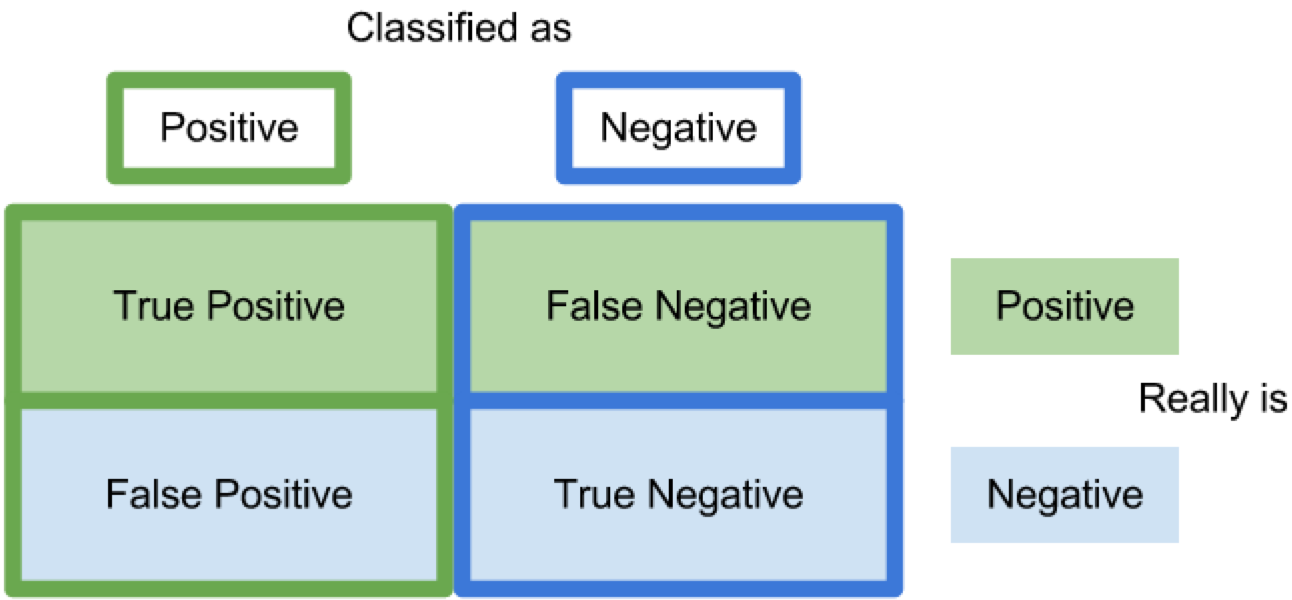
Lesson 3 Review: Sampling Data in Classification

- **Why Sample?**
  - Numerosity Reduction: dealing with a smaller subset of massive dataset that is representative of the population
- ***Simple samples***
  - I take 30% of my original sample
  - May not be appropriate for unbalanced data (1000 positive and 100 negative cases)
- ***Complex samples***
  - ***-Clustered samples***: used to sample groups or clusters rather than individual units.
- ***Stratified samples***
  - ***-Stratified samples***: Used to select samples independently within non-overlapping subgroups of the population, or strata.
  - -For example, you can ensure that men and women are sampled in equal proportions, or that every region or socioeconomic group within an urban population is represented.
  - -You can also specify a sample size for each strata

Partition

Metrics for Performance Evaluation

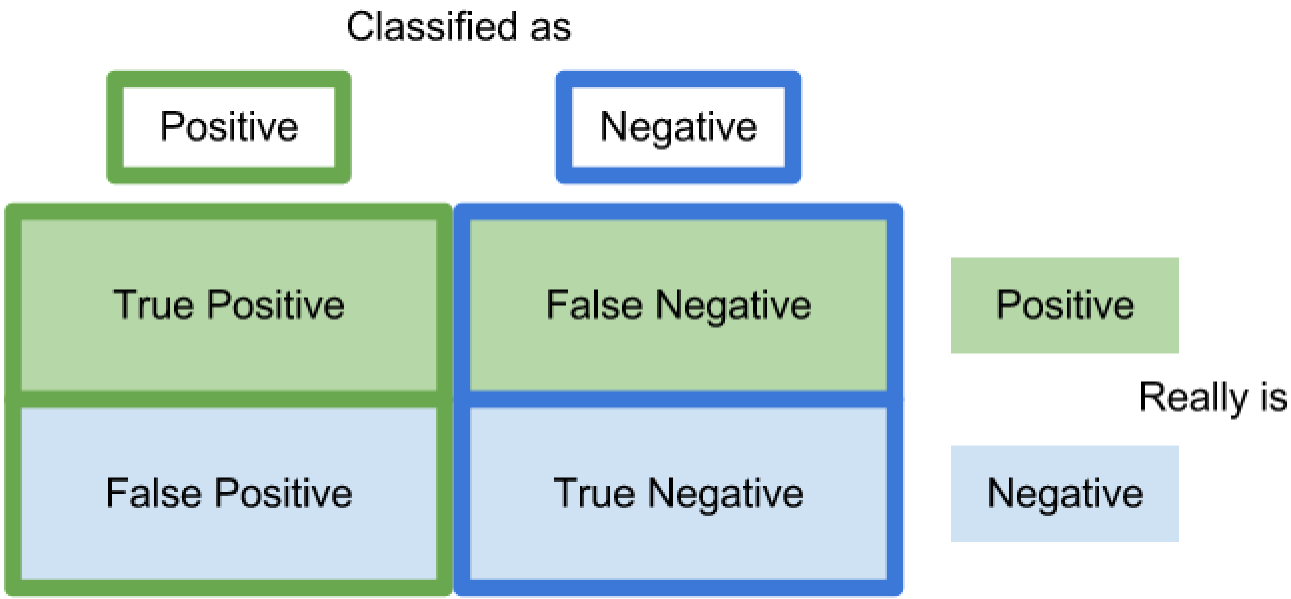
- Focus on the predictive capability of a model
  - Rather than speed, scalability, etc.
- Confusion Matrix:



Classified as

Positive	Negative	Really is
True Positive	False Negative	Positive
False Positive	True Negative	Negative

Accuracy as Performance Evaluation



	Classified as Positive	Classified as Negative	
Positive	True Positive	False Negative	Really is
Negative	False Positive	True Negative	

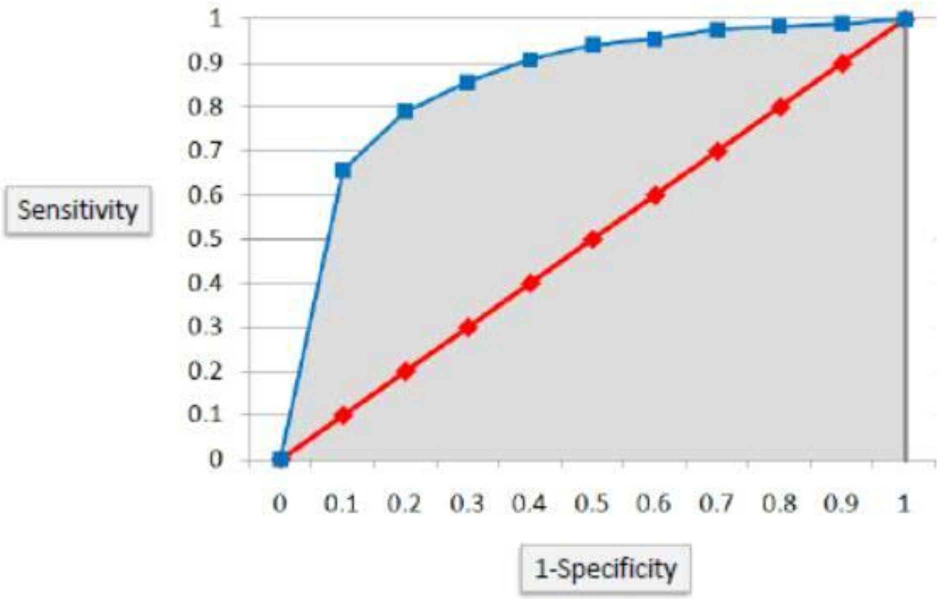
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Overcoming Limitations of Accuracy Measure

- **Situation:**
  - Positive cases = 990
  - Negative = 10
- **If model predicts everything to be positive, accuracy is 990/1000 = 99 %**
  - **But:** model fails on negative cases
  - What is negative cases are really important and costly to overlook?
- **Other (cost sensitive measures):**
  - Precision = TP / (TP + FP)
  - Sensitivity = TP / (TP + FN)
  - Specificity = TN / (TN + FP)

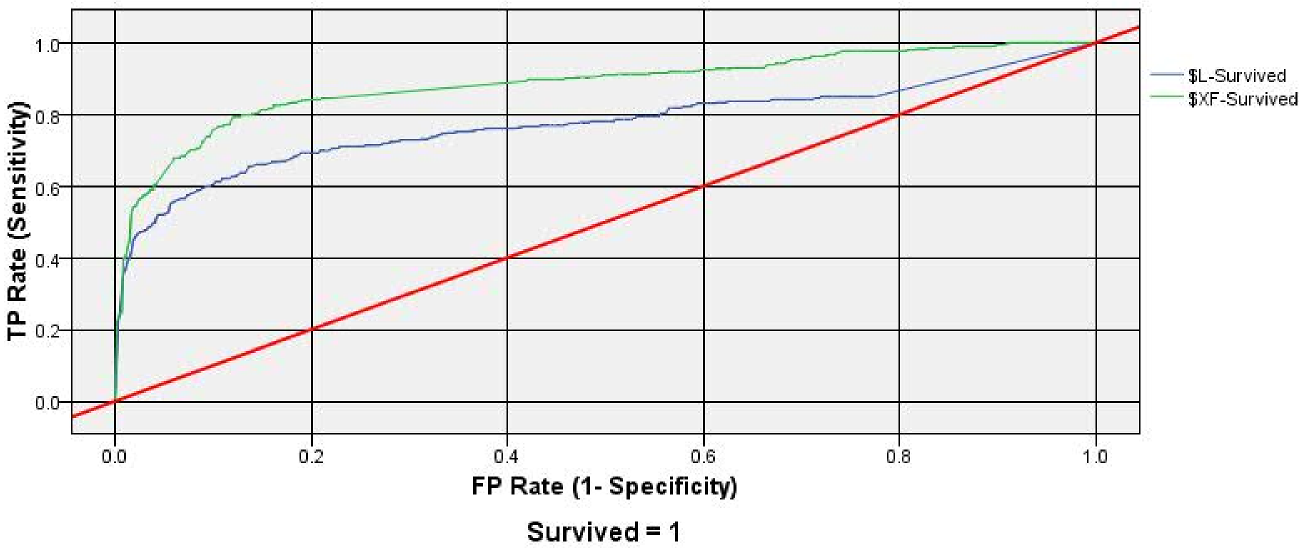
ROC Curves

- Receiver operator characteristic
- Summarize & present performance of a binary classification model
- Models ability to distinguish between false & true positives

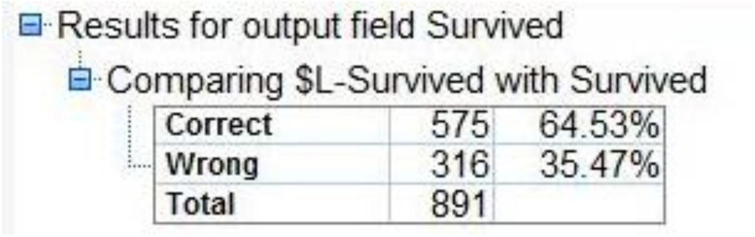


Tools for Model Evaluation

Evaluation



Analysis



Results for output field Survived

Comparing $L-Survived with Survived

Correct	575	64.53%
Wrong	316	35.47%
Total	891	

Lab 4:

- Evaluate how your model performs
- Compare models