WE
DEMOCRATIZE
ANALYTICS

FRONTLINE solvers

EXCEL USERS
WEB USERS
DEVELOPERS

*Data Visualization * Data Mining * Simulation / Risk Analysis * Decision Trees * Conventional / Stochastic Optimization*

DATA MINING

REVIEW BASED ON

MANAGEMENT SCIENCE

The Art of Modeling with Spreadsheets

Using Analytic Solver Platform

FrontlineSolvers

What We'll Cover Today

- Introduction
  - Session II beta training program goals
  - Brief overview of XLMiner
- Overfitting problem
- Partitioning the data
- Supervised learning – classification

Session II Online Beta Training Goals

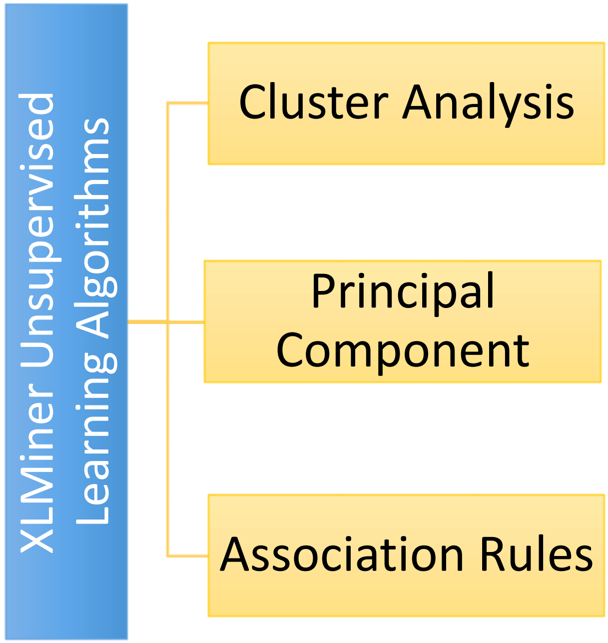
- To empower you to achieve success
  - State of the art tools
  - Online educational training
  - Training documents and demos
- To familiarize you with the following concepts:
  - Understanding the ideas behind the classification techniques
  - Fitting classification models to data
  - Assessing the performance of methods
  - Applying the models to predict unseen test cases

Data Mining Steps

Identify Opportunity → Collect Data → Explore, Understand, and Prepare Data → Identify Task and Tools → Build and Evaluate Models → Deploy Models

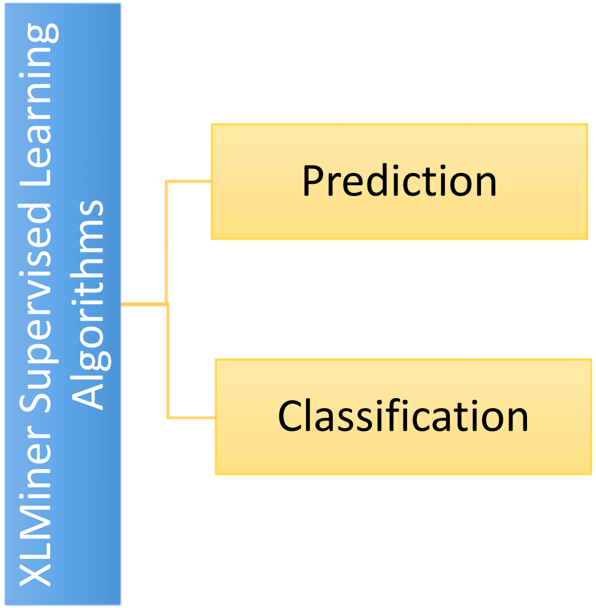
Unsupervised Learning Algorithms

- No outcome variable in the data set, just a set of variables (features) measured on a set of samples.
  - Market basket analysis.
  - Social network analysis.

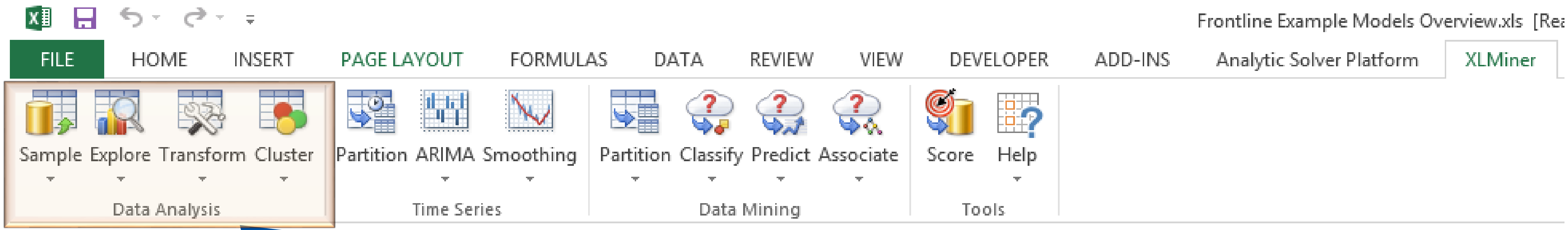


Supervised Learning Algorithms

- For each record:
  - Outcome measurement y (dependent variable, response, target).
  - Vector of predictor measurements x (feature vector consisting of independent variables).
- Prediction:
  - Housing market: Price.
  - Product: Demand.
- Classification:
  - Online Transactions: Fraudulent (Yes / No)?
  - Email: Spam / Not Spam?
  - Insurance Applicant: High / Medium / Low Risk?



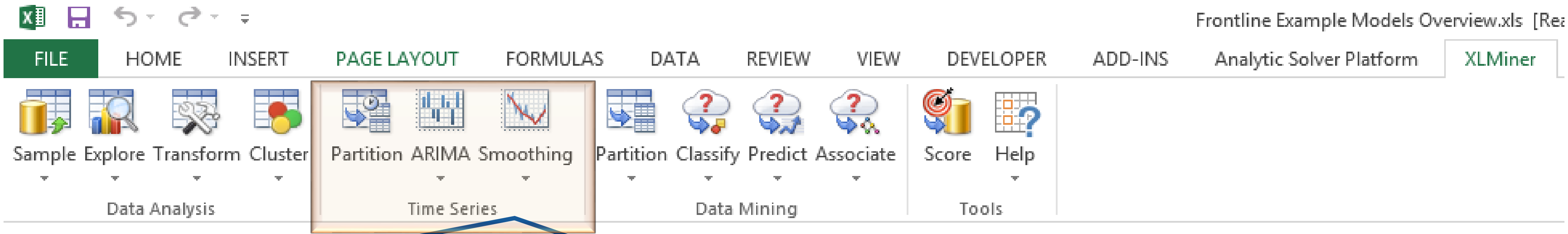
Brief Overview of XLMiner



Data Analysis

- Draw a sample of data from a spreadsheet, or from external database (MS-Access, SQL Server, Oracle, PowerPivot)
- Explore your data, identify outliers, verify the accuracy, and completeness of the data
- Transform your data, define appropriate way to represent variables, find the simplest way to convey maximum useful information
- Identify relationships between observations, segment observations

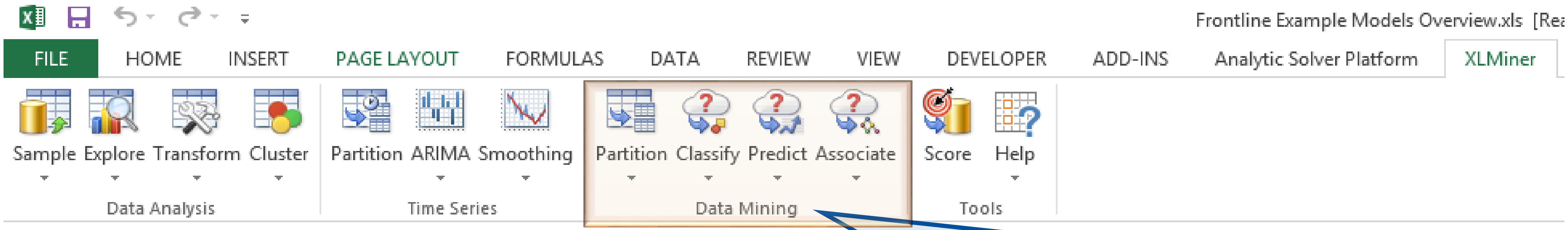
Brief Overview of XLMiner



Time Series

- Forecast the future values of a time series from current and past values
- Smooth out the variations to reveal underlying trends in data
  - Economic and business planning
  - Sales forecasting
  - Inventory and production planning

Brief Overview of XLMiner

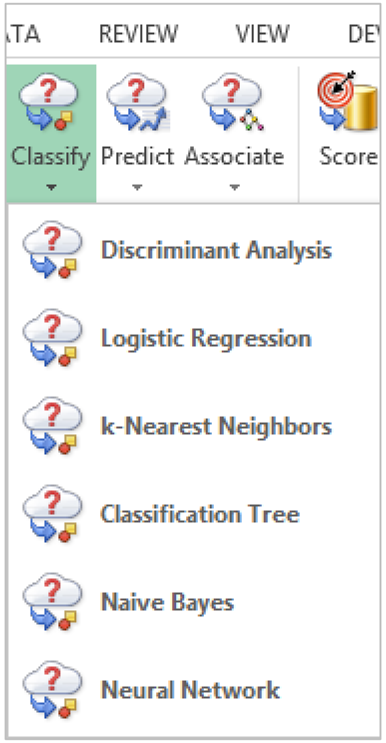


Data Mining

- Partition the data so a model can be fitted and then evaluated
- Classify a categorical outcome – good/bad credit risk
- Predict a value for a continuous outcome – house prices
- Find groups of similar observations – market basket analysis

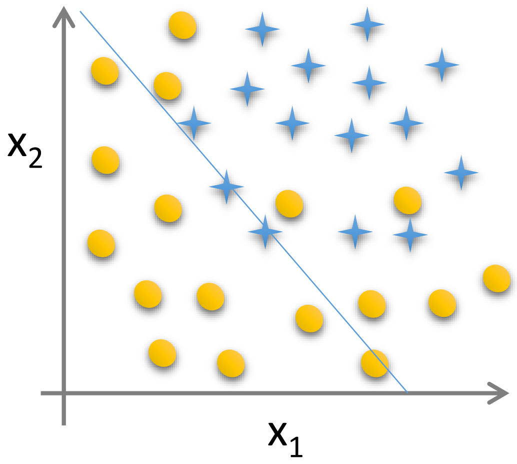
Chapter 6 - Part I
Classification Methods

Using XLMiner



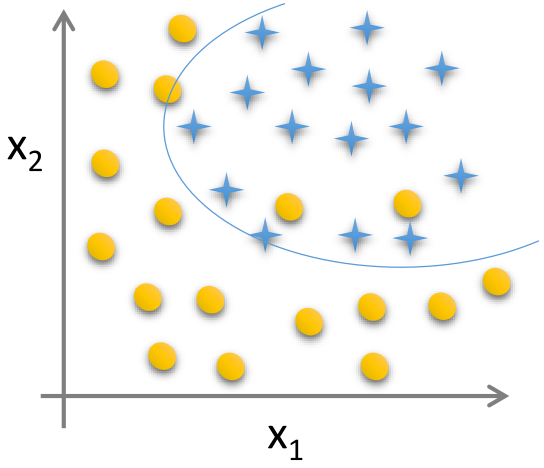
The Problem of Overfitting

- If we have a complicated model, the model may fit and explain the training data very well, yet fails to generalize to new data.

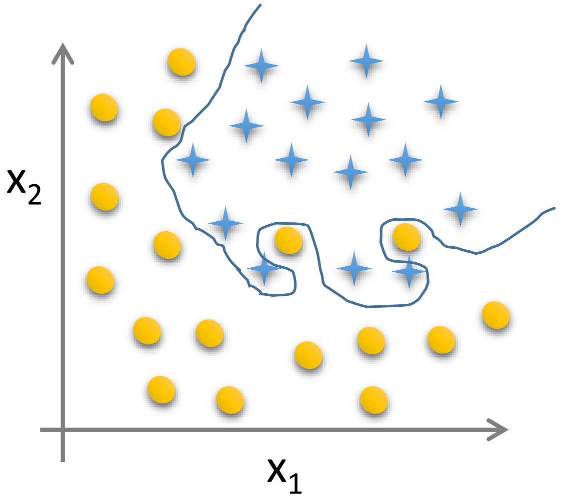


$f(\alpha_0+\alpha_1 x_1+\alpha_2 x_2)$

Underfit



$f(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_3 x_1{}^2$
$+\alpha_4 x_2{}^2+\alpha_5 x_1 x_2)$



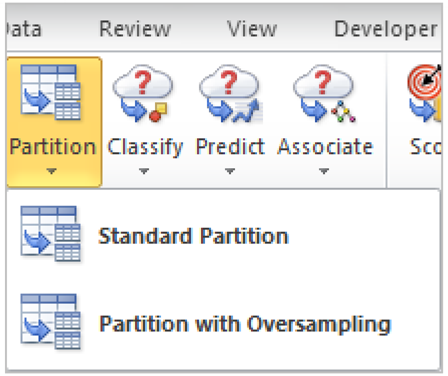
$f(\alpha_0+\alpha_1 x_1+\alpha_2 x_1{}^2+\alpha_3 x_1{}^2 x_2+$
$\alpha_4 x_1{}^2 x_2{}^2+\alpha_5 x_1{}^2 x_2{}^3+\alpha_6 x_1{}^3 x_2+\ldots)$

Overfit

Partitioning the Database

- Helps to avoid overfitting by testing the model on validation part.
- Partitioning is segmenting the data into following groups.
  - **Training set:** used for learning the parameters of model.
  - **Validation set:** used for evaluating the model error and tuning parameters.
  - **Test set (optional):** used for a final, independent test of the performance of the model on new data that was not part of the model building.

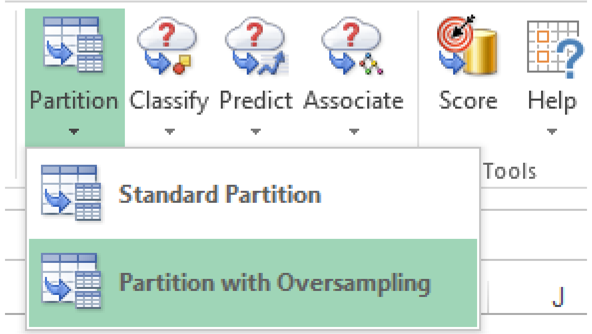
Partitioning the Database XLMiner



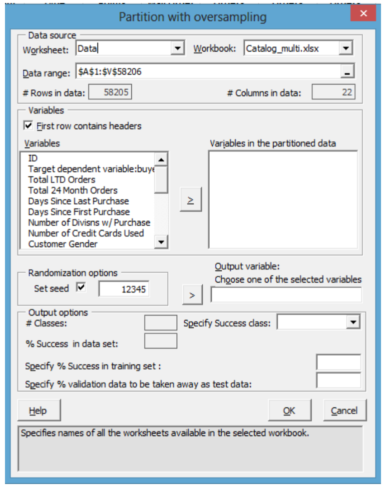
- Standard Partitioning
  - Random partitioning
  - User-defined Partitioning
- Partitioning with Oversampling
  - Use Oversampling when there are only two categories and the group of interest is rare.
  - **Example**: Universal Bank data – personal loans solicitations.

Summary- Partitioning with Oversampling Using XLMiner

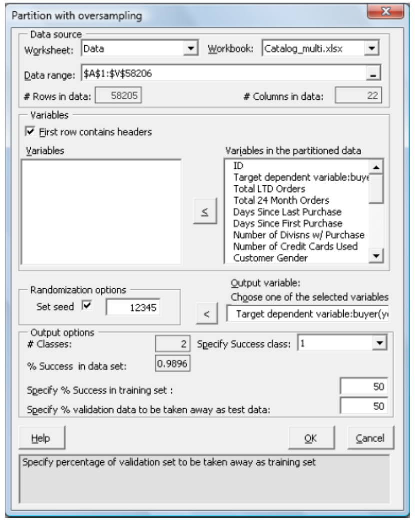
- Click any cell within the dataset, then click **Partition – Partition with Oversampling** (in the Data Mining section of the XLMiner ribbon).



- Select all variables in the **Variables** list box then click > to move all variables to the **Variables in the partitioned data** listbox.



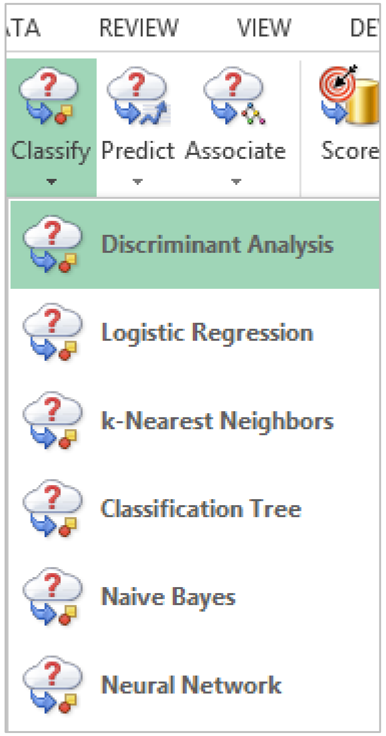
- Highlight the target variable in the *Variables in the partitioned data listbox* then click the > to the left of *Output variable* to designate this variable as the output variable, then click OK.



FrontlineSolvers
WE DEMOCRATIZE ANALYTICS

Classification Using XLMiner

- Discriminant Analysis
- Logistic Regression
- *k*-Nearest Neighbor
- Classification Tree
- Naïve Bayes
- Neural Networks

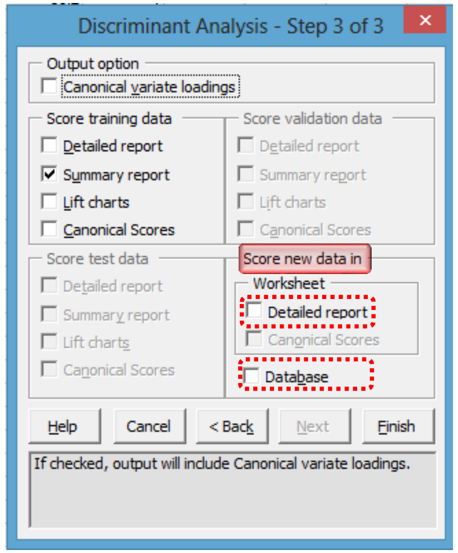


Discriminant Analysis (DA)

- Estimates the probabilities that a given record falls into one of the possible classes.
- Estimates means and covariance(s) of groups using training data.
- Models distribution of each group separately.
- Bayes theorem - posterior probabilities (adjusted with prior frequencies of classes).
- Independent variables are assumed to be normally distributed.
- Linear discriminant analysis (LDA) - linear decision boundaries.
- Quadratic discriminant analysis (QDA) - quadratic decision boundaries.

Scoring New Data

- XLMiner's dialogs for classification routines provide an option to score new data in a database or from worksheet.
- In the Discriminant Analysis – Step 3 of 3 dialog.
- Score new data in a database using XLMiner : MS-Access, SQL Server, Oracle.
    - Example: Scoring to MS-Access Database
- XLMiner's "Score" in the Tools group, will allow you to score new data after you have fitted your model. XLMiner produces Stored Worksheet with saved model.

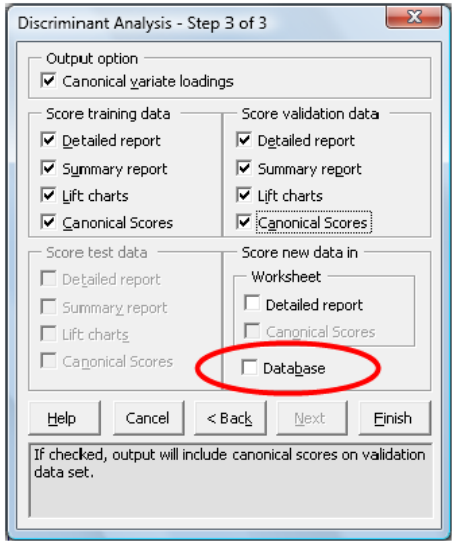


Score

FrontlineSolvers
WE DEMOCRATIZE ANALYTICS

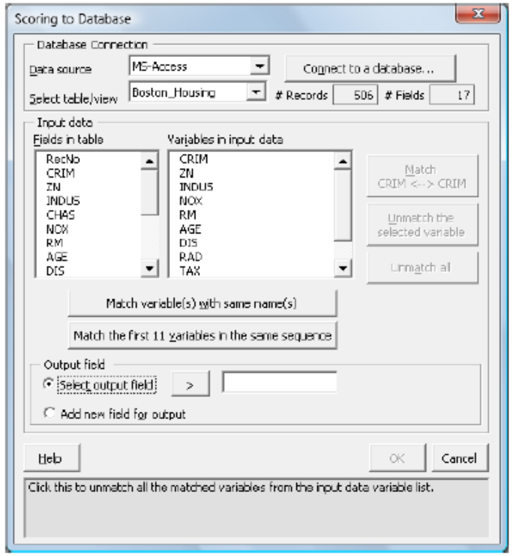
Summary-Scoring to a Database

- In the Discriminant Analysis method, this feature is found on the Step 3 of 3 dialog.



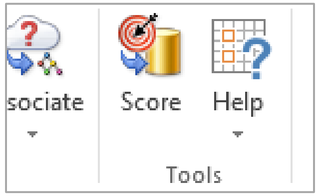
- In the *Score new data in* group, select **Database**. The *Scoring to Database*
- The first step on this dialog is to select the **Data source**.
- Once the *Data source* is selected, **Connect to a database…** will be enabled.
- Enter the appropriate details, then click **OK** to be connected to the database.

- Match variables in the dataset to variables in the database and click **OK**.

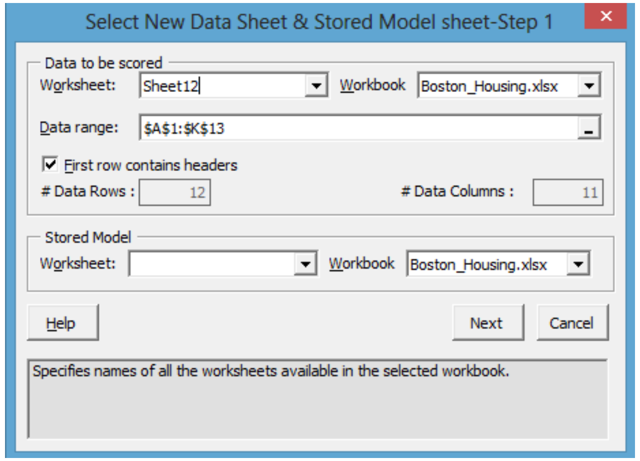


Summary-Score Test Data Using DA Model

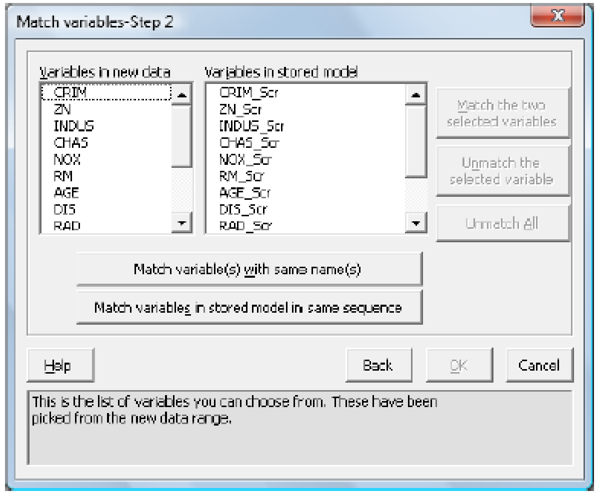
- Click **Score** on the XLMiner ribbon.



- Select the new data and the Stored Model worksheets.



- Click **Next**. XLMiner will open the *Match variables – Step 2* dialog.
- Match the Input variables to the New Data variables using **Match variable(s) with same names(s)** or **Match variables in stored model in same sequence.**
- Then click **OK**.



Strengths and Weaknesses of Discriminant Analysis

Strengths:

- Very fast even for large data.
- Useful and well-interpretable – number of features is not large.
- Perfect fit – normal group distributions.
- Stable model – well-separated groups.
- Multiclass learning – can explain data in lower dimensions.
  - Similar to PCA, but in a supervised way.

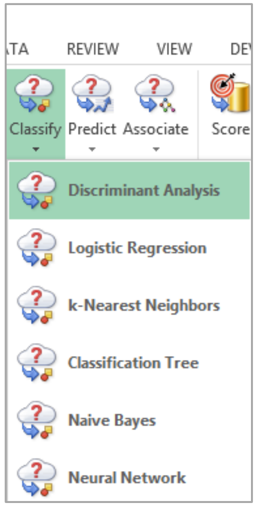
Strengths and Weaknesses of Discriminant Analysis

Weaknesses:

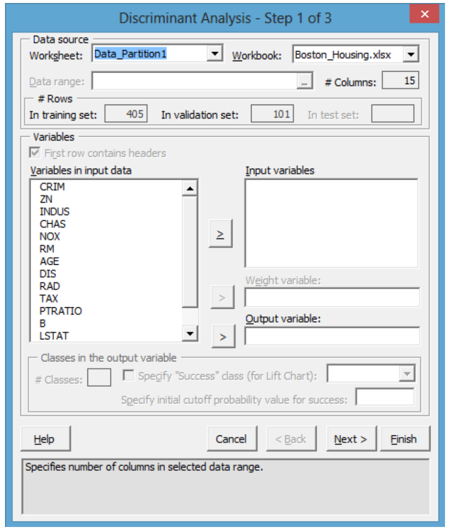
- Does not apply – number of features exceeds number of records.
- Overcomplicated and less stable – high-dimensional data.
- May fail to capture structure of the data – highly non-Normal distributions.

Summary-Discriminant Analysis

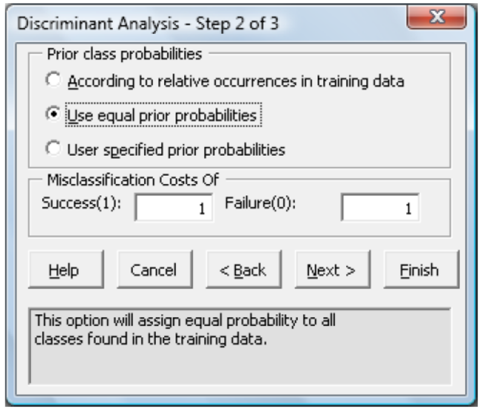
- Partition the data.
- Select a cell on the Data_Partition1 worksheet then click **Classify – Discriminant Analysis**.



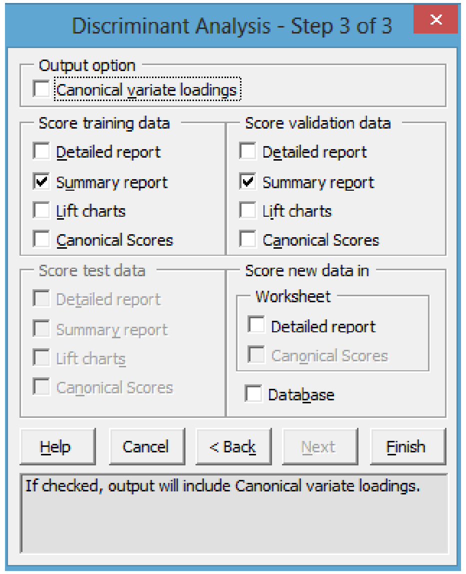
- Select the *Output variable and Input Variables.*



- Click **Next** and select the desired method of computing *Prior class probabilities.*

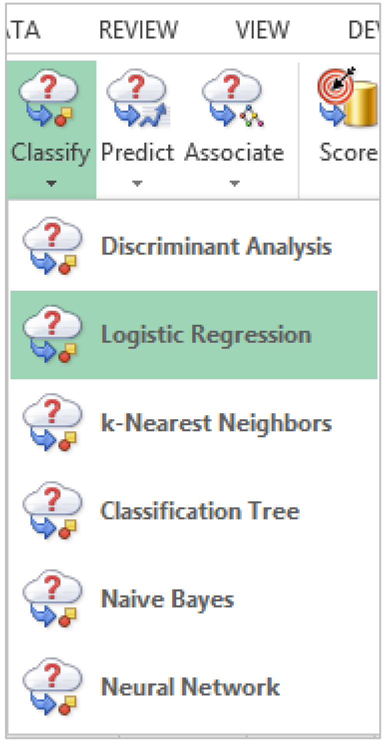


- Select the output and score training and validation data options.



Classification Using XLMiner

- Discriminant Analysis
- Logistic Regression
- k-Nearest Neighbor
- Classification Tree
- Naïve Bayes
- Neural Networks



Logistic Regression (LR)

- Extremely powerful and widely used.
- Extends Linear Regression.
- XLMiner – binary classification problems.
- Fitted parameters – estimate the probability of given records belonging to one of two possible groups.

Logistic Regression

- Models *Logit* transformation – linear combination of predictors:

$$Logit(P\{success|x\}) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

- LR – conditional probabilities (**generative learning**)
- DA – joint probabilities (**discriminative learning**)

Strengths and Weaknesses of Logistic Regression

Strengths:

- Very popular – 2 classes.
- No assumption – distribution of independent variables.
- Unlike Linear Regression – error terms are not assumed to be normally distributed.
- No assumption – linear relationship between independent and response variables.
- Performs well – data containing categorical predictors.
- Handles large high-dimensional datasets.

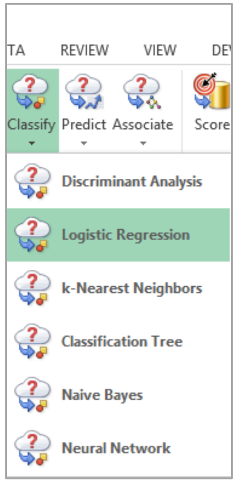
Strengths and Weaknesses of Logistic Regression

Weaknesses:

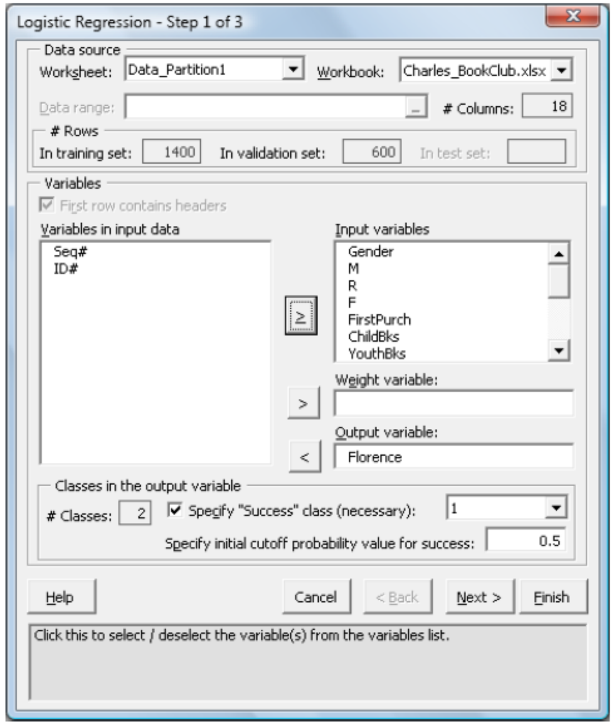
- Less stable – low dimensional data where classes are well-separated.
  - Discriminant Analysis.
- Less efficient – number of records are less than number of features and when collinearity is present.
  - XLMiner – **embedded variable selection** and **best subset**.

Summary- Logistic Regression

- Select a cell on the *Data_Partition1* output worksheet, then click **Classify – Logistic Regression** on the XLMiner ribbon.

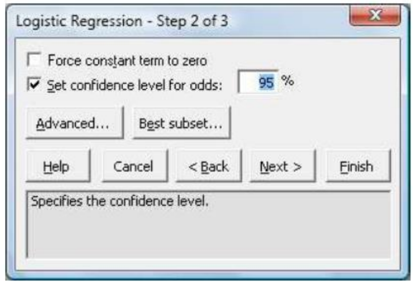


- Choose input and output variables.
- Choose the value that will be the indicator of "Success" by clicking the down arrow next to *Specify "Success" class (necessary)*.
- Specify the initial cutoff probability for success, and Click **Next.**

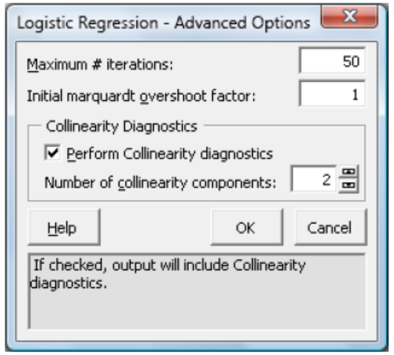


Summary- Logistic Regression

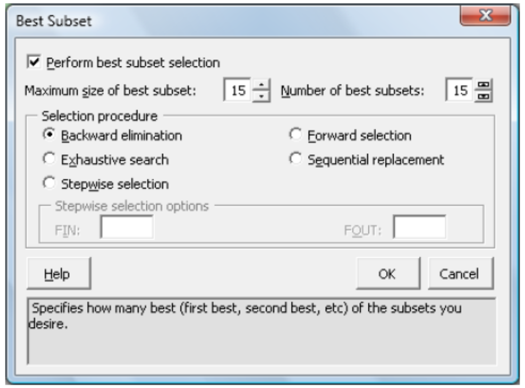
- Set confidence level and Click **Advanced.**



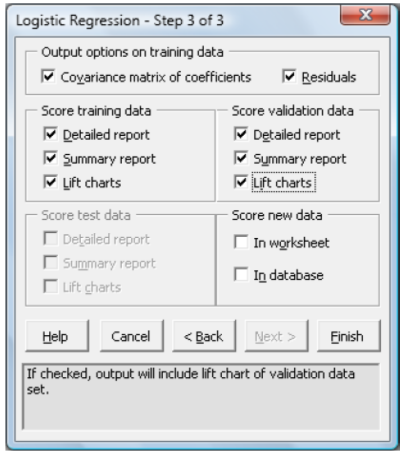
- Select the desired options and Click **OK** to return to the *Step 2 of 3* dialog.



- Click Best Subset and Select **Perform best subset selection**.
- Choose the desired selection procedures for selecting the best subset of variables.



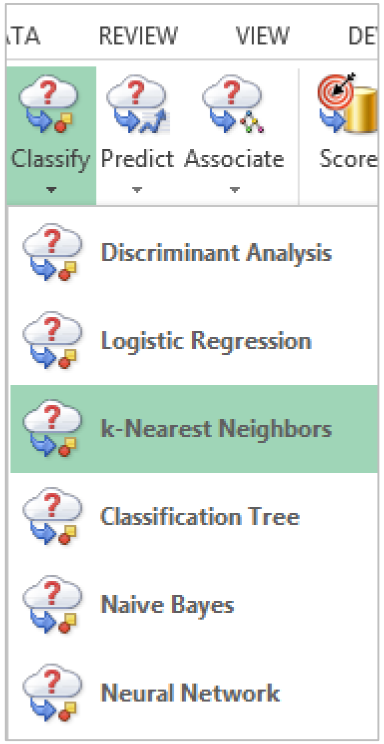
- Click **OK** to return to the *Step 2 of 3* dialog.
- Click **Next** to advance to the *Step 3 of 3* dialog.
- Select **Covariance matrix of coefficients, Residuals, reports, and Lift charts,** then Click **Finish**.



FrontlineSolvers
WE DEMOCRATIZE ANALYTICS

Classification Using XLMiner

- Discriminant Analysis
- Logistic Regression
- k-Nearest Neighbors
- Classification Tree
- Naïve Bayes
- Neural Networks



k-Nearest Neighbor

- Very simple powerful algorithm – classification decision based on information from neighboring records.
  - k observations – most similar.
  - Majority voting – most frequent group among the k nearest neighbors.
- No learning stage – training data is our model.
- Similarity measure – Euclidean Distance.
- Independent variables – scaled appropriately.
- Best model – assessing the classification error for various values of k.
- Less chance of overfitting – validation error.

Strengths and Weaknesses of the k-Nearest Neighbor Algorithm

Strengths:

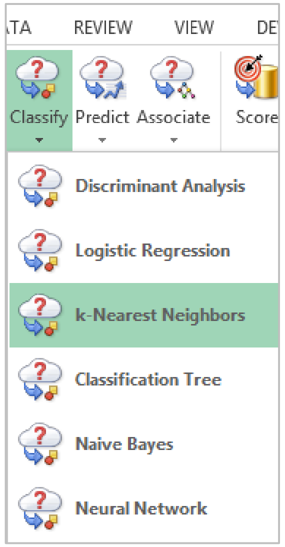
- Very often performs well in practice.
- Stable and easily interpretable results.

Weaknesses:

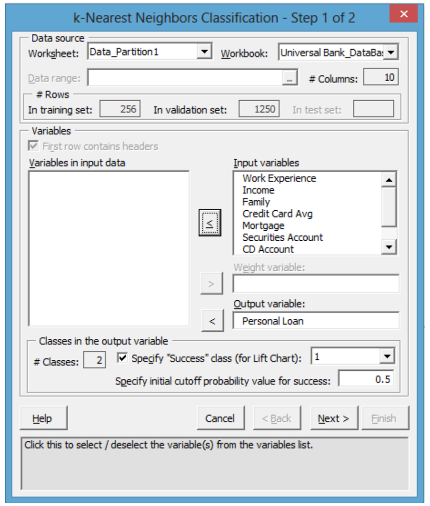
- Expensive – computationally.
- Focus – local structure.
  - Fails – global picture.
- “Curse of dimensionality.”
- Extremely sensitive – outliers and noise.
- Poor performance – undersampled/oversampled groups.

Summary-*k*-Nearest Neighbor

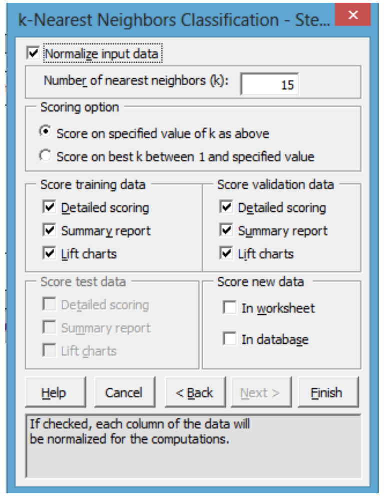
- Select a cell on the Data_Partition1 worksheet, then click **Classify – k-Nearest Neighbors** on the XLMiner ribbon.



- Select desired variables under *Variables in input data* then click > to select as input variables. Select the output variable or the variable to be classified.

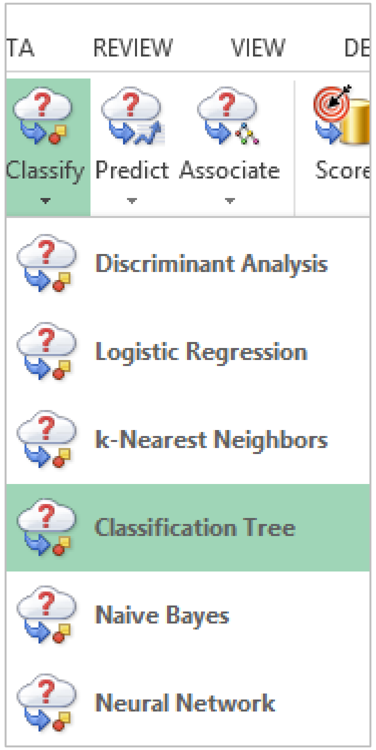


- *Specify "Success" class and the initial cutoff value*, and click **Next.**
- Select **Normalize input data** and the reports and input Number of nearest neighbors. Click **Finish**.

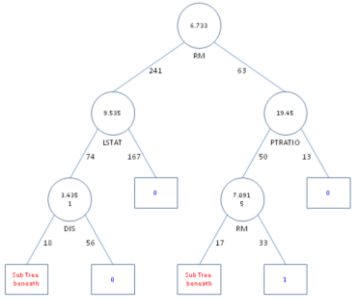


Classification Using XLMiner

- Discriminant Analysis
- Logistic Regression
- k-Nearest Neighbor
- Classification Tree
- Naïve Bayes
- Neural Networks

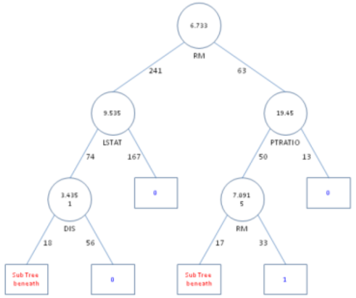


Classification Tree



- Splitting rules – partitions space of independent variables.
    - Tree – summarized and visualized process.
- “Best” splits – measure (e.g., Gini index, Information Gain).
- Internal node – for splitting.
- Branch – two subsets of possible values of parent node.
- Leaf nodes – value of response.

Classification Tree



- Fully grown classification tree – overfitting.
- Solution – *pruning*.
- Over-pruned tree – lose ability to capture structural information.
    - What is the optimal size?
- Optimal pruning techniques – reduce size without sacrificing predictive accuracy.

Strengths and Weaknesses of Classification Trees

Strengths:

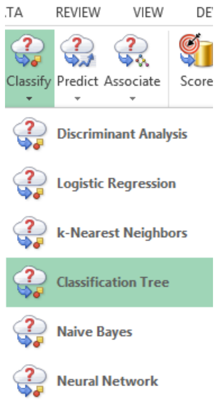
- Easily interpreted – if-then rules.
- Handles raw data.
- Implicit *feature selection*.
- No explicit assumptions – underlying relationships.

Weaknesses:

- Greedy heuristic approach – locally optimal solution.

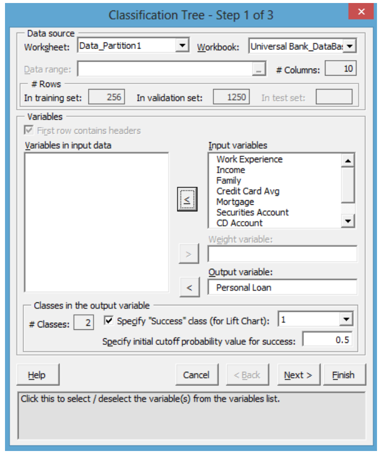
Summary-Classification Tree

- Select a cell on the Data_Partition1 worksheet, then click **Classify – Classification Tree** on the XLMiner ribbon.

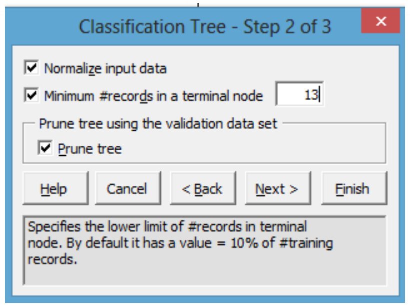


- Select *Output and Input variables*.

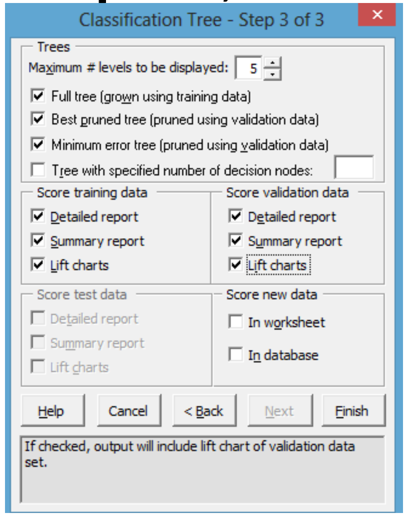
- Specify *"Success" class* and Specify *initial cutoff probability*, then click **Next.**



- Select **Normalize input data, Minimum #records in a terminal node**, and **Prune tree**, then click **Next.**



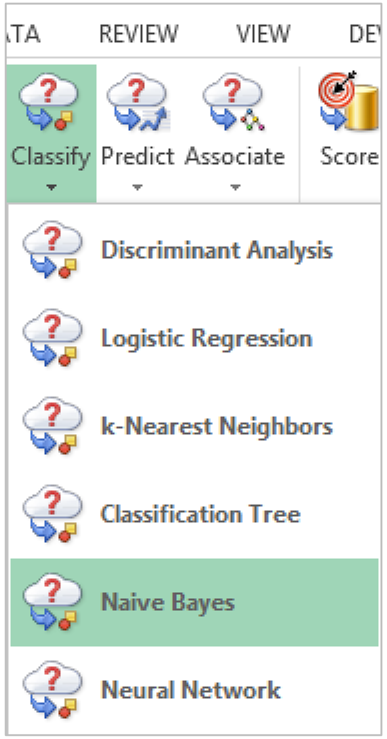
- Set *Maximum # levels to be displayed*, select **Full tree, Best pruned tree, Minimum error tree, and reports**, then click **finish.**



FrontlineSolvers
WE DEMOCRATIZE ANALYTICS

Classification Using XLMiner

- Discriminant Analysis
- Logistic Regression
- k-Nearest Neighbor
- Classification Tree
- **Naïve Bayes**
- Neural Networks



Naïve Bayes

- Bayes rule – posterior probabilities.
  - Assign classes – MAP (maximum a posteriori).
- Conditional independence of features.
- XLMiner – Multivariate Multinomial distribution.
  - XlMiner – Bin Continuous Data.
- “Naïve” assumptions – yet surprising efficiency.

Strengths and Weaknesses of the Naïve Bayes Algorithm

Strengths:

- Applicable – high-dimensional data.
- Parameter estimation – small training sample.
- Applicable – discrete and continuous data.
- Efficient – computationally.
- Robust with irrelevant features.
- Perfect classifier – independent features.

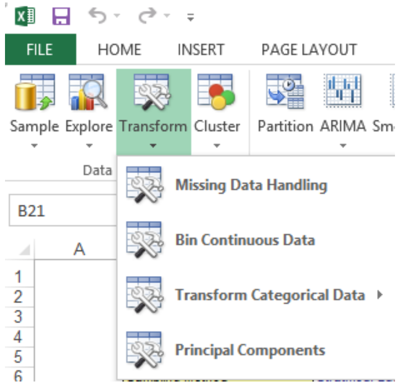
Strengths and Weaknesses of the Naïve Bayes Algorithm

Weaknesses:

- Independence assumption – strong.
- Multinomial model – must contain already observed values.

Naïve Bayes **Data Preparation:** Binning Continuous Data using XLMiner

- Click **Transform -- Bin Continuous Data** on the XLMiner ribbon.

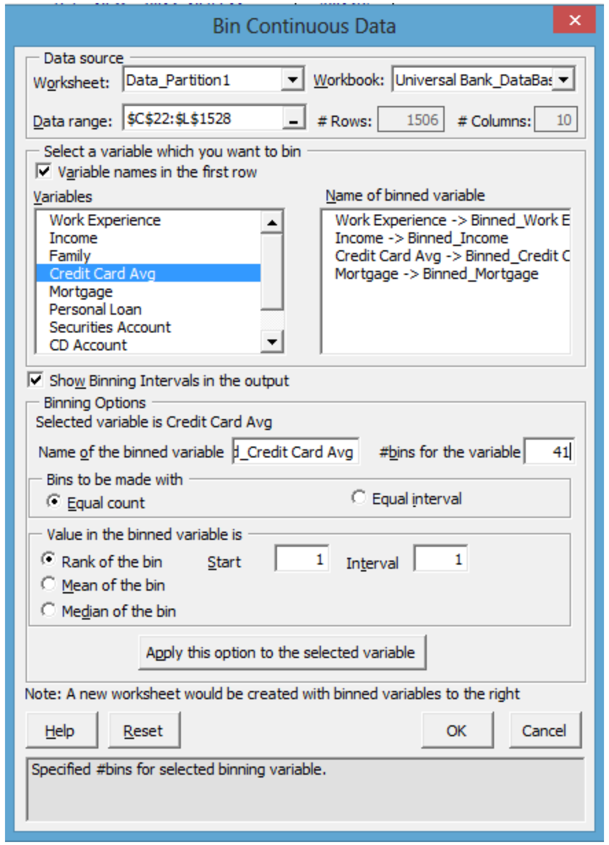


- Select the continuous variable and enter **#bins for the variable**.

- Select **Equal Count** for binning the variable.

- Select **Rank** to assign category label to bin intervals.

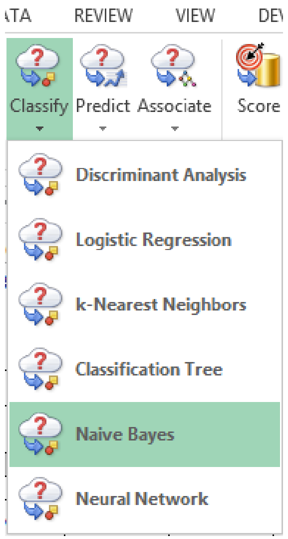
- Click on Apply this option and click on ok.



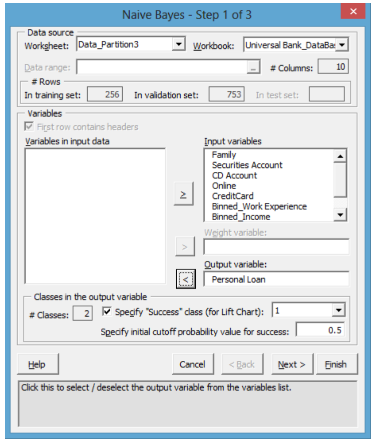
FrontlineSolvers
WE DEMOCRATIZE ANALYTICS

Summary- Naïve Bayes

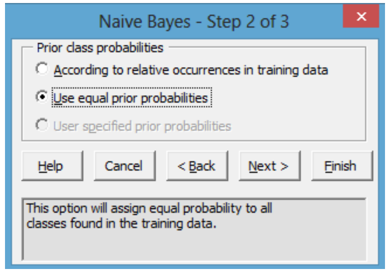
- Partition the Binneddata1.
- Select a cell on the *Data_Partition1* worksheet, then click **Classify – Naïve Bayes**.



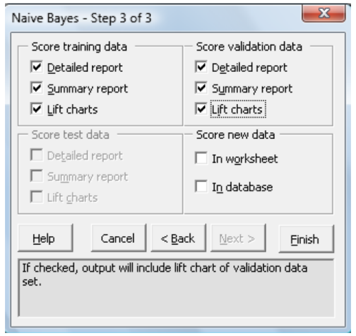
- Select Input and Output variables.
- *Specify "Success" class and* Enter a value between 0 and 1 for *Specify the initial cutoff probability for success*. Click **Next.**



- Select an option for *Prior class probabilities*. Then Click **Next**.



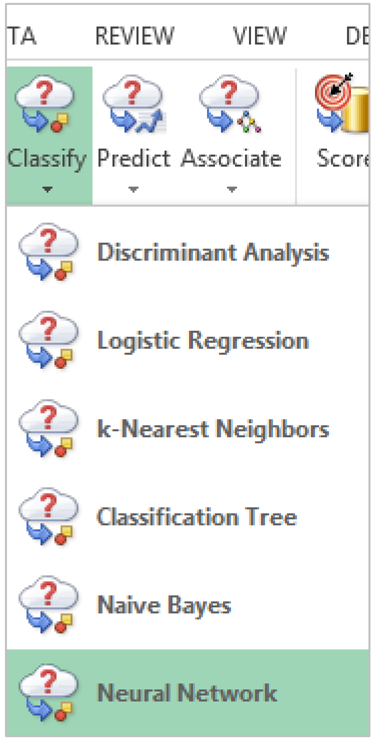
- Select **Detailed report**, **Summary report**, and **Lift charts.** Click **Finish.**



FrontlineSolvers
WE DEMOCRATIZE ANALYTICS

Classification Using XLMiner

- Discriminant Analysis
- Logistic Regression
- k-Nearest Neighbor
- Classification Tree
- Naïve Bayes
- Neural Networks

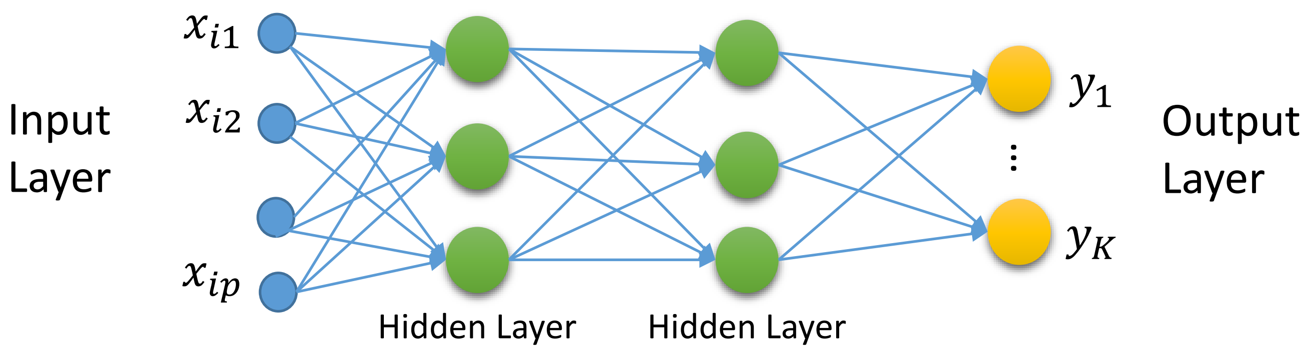


Neural Networks (NN)

- Powerful machine learning technique – structure of the human brain.
- XLMiner – feed-forward back-propagation.
- Interconnected neurons – organized in layers.
- Neurons – computational units.
- Internally feature extraction.
- Dependency – settings and architecture.

Neural Networks Key Components

- Input neurons – features.



- Output layer prediction – fed-forwarded information.
- Back-propagated errors – learning.
- Epoch – processing of all training observations.
- Desired predictive accuracy (training, cross-validation errors) – many learning epochs.

Strengths and Weaknesses of Neural Networks

Strengths:

- "Universal Approximators."
- Detects – independent and depended variables' nonlinear relationships.
- Detects – predictors' relationships.
- Automated Learning – less formal modeling.
- Robust model – large high-dimensional datasets.
- No strong explicit assumptions.

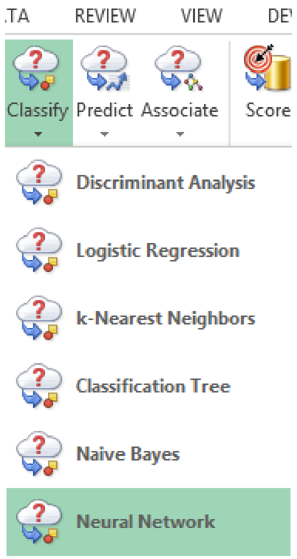
Strengths and Weaknesses of Neural Networks

Weaknesses:

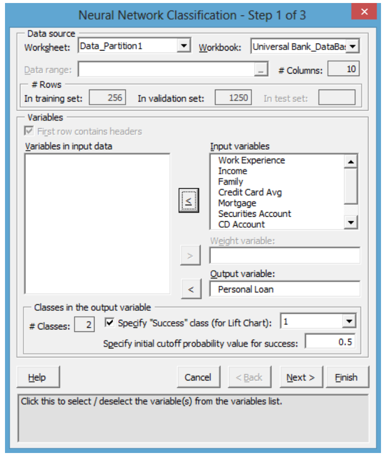
- “Black-box” learning.
- Expensive – computationally.
- Prone to overfitting.
- Dependency – architecture, parameters, choice of activation and error functions.
  - XLMiner – Automatic Network Architecture option.

Summary-Neural Networks

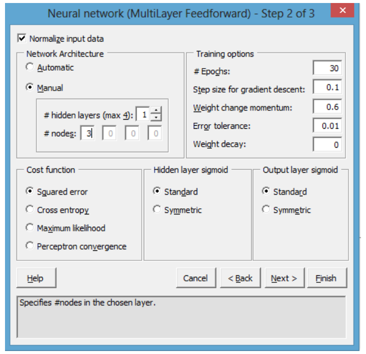
- Select a cell on the *Data_Partition1* worksheet, then click **Classify – Neural Network.**



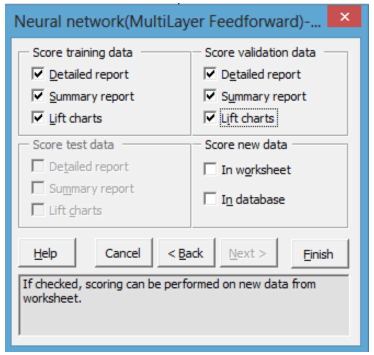
- Select Input and Output variables.
- Specify *"Success" class and* Enter a value between 0 and 1 for *Specify the initial cutoff probability for success*. Click **Next.**



- Select **Normalize input data.** Manfully adjust the **Network Architecture** and **Training options.**



- Select the **Reports** and click **Finish**.



Comments on Classification

- No perfect model – different predictive power and accuracy.
- Build several models – best overall performance.
- Fundamental problems:
  - **Overfitting**.
    - Choose simple – best.
    - Use cross-validation.
  - **Curse of dimensionality**.
    - Choose algorithm – consider dimensions.
    - Reduce data dimension – explicitly or use XLMiner's techniques.
- Final independent test – use test samples.

Summary

- Classification – whether a customer will buy a certain product.
- XLMiner classification techniques.
- Fitting classification models to data.
- Working with output of each method.
- Appling fitted models to classify new observations.

Summary

- Vital skill for business analysts – use data intelligently.
- Retrieve and combine data from from SQL databases to Web data sources – use Excel.
- Visualize and transform your data, apply supervised and unsupervised learning methods – use XLMiner in Excel.
- A complete toolset for descriptive, predictive and prescriptive analytics – use Analytic Solver Platform including XLMiner.

Contact Info

- Dr. Sima Maleki
- Best way to contact me: Consulting@Solver.com
- You may also download this presentation from our website.
- You can download a free trial version of XLMiner at http://www.solver.com/xlminer-data-mining

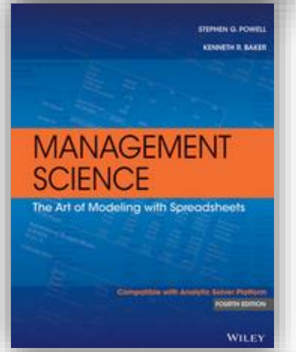
References

- **Spreadsheet Modeling and Decision Analysis: A Practical Introduction to Business Analytics, 7th Edition**

http://www.cengage.com/us/

- **MANAGEMENT SCIENCE-The Art of Modeling with Spreadsheets, 4th Edition**

http://www.wiley.com/WileyCDA/WileyTitle/productCd-EHEP002883.html



- **Essentials of Business Analytics, 1st Edition**

http://www.cengage.com/us/

Q & A

Thank You!