Classification

Decision Tree

Dataset - Chapter10DataSet_Training, Chapter10DataSet_Scoring

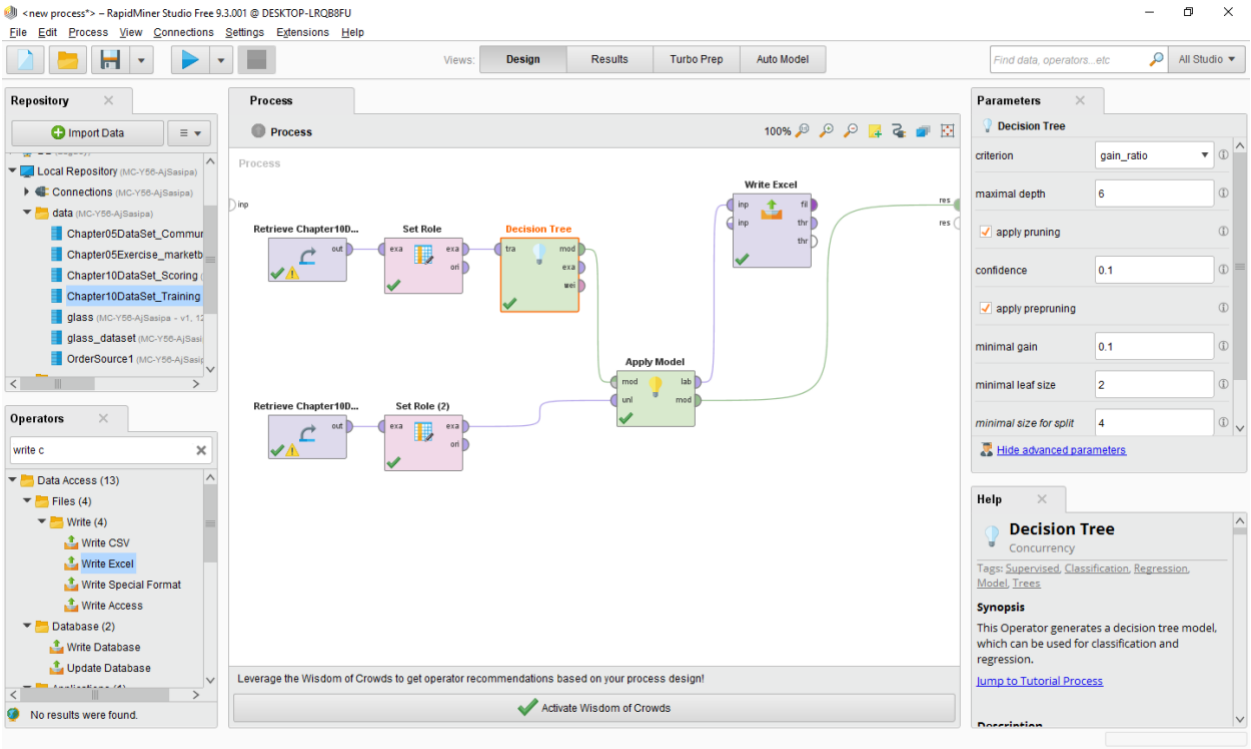
In analyzing his data set, Richard has found that customers' activity in the areas of digital media and books, and their general activity with electronics for sale on his company's site, seem to have a lot in common with when a person buys an eReader. With this in mind, we have worked with Richard to compile data sets comprised of the following attributes:

- **User_ID**: A numeric, unique identifier assigned to each person who has an account on the company's web site.
- **Gender**: The customer's gender, as identified in their customer account. In this data set, it is recorded a 'M' for male and 'F' for Female. The Decision Tree operator can handle non-numeric data types.
- **Age**: The person's age at the time the data were extracted from the web site's database. This is calculated to the nearest year by taking the difference between the system date and the person's birthdate as recorded in their account.
- **Marital_Status**: The person's marital status as recorded in their account. People who indicated on their account that they are married are entered in the data set as 'M'. Since the web site does not distinguish single types of people, those who are divorced or widowed are included with those who have never been married (indicated in the data set as 'S').
- **Website_Activity**: This attribute is an indication of how active each customer is on the company's web site. Working with Richard, we used the web site database's information which records the duration of each customers visits to the web site to calculate how frequently, and for how long each time, the customers use the web site. This is then translated into one of three categories: Seldom, Regular, or Frequent.
- **Browsed_Electronics_12Mo**: This is simply a Yes/No column indicating whether or not the person browsed for electronic products on the company's web site in the past year.

Bought_Electronics_12Mo: Another Yes/No column indicating whether or not they purchased an electronic item through Richard's company's web site in the past year.

- **Bought_Digital_Media_18Mo**: This attribute is a Yes/No field indicating whether or not the person has purchased some form of digital media (such as MP3 music) in the past year and a half. This attribute does not include digital book purchases.
- **Bought_Digital_Books**: Richard believes that as an indicator of buying behavior relative to the company's new eReader, this attribute will likely be the best indicator. Thus, this attribute has been set apart from the purchase of other types of digital media. Further, this attribute indicates whether or not the customer has *ever* bought a digital book, not just in the past year or so.
- **Payment_Method**: This attribute indicates how the person pays for their purchases. In cases where the person has paid in more than one way, the mode, or most frequent method of payment is used. There are four options:
    - Bank Transfer—payment via e-check or other form of wire transfer directly from the bank to the company.
    - Website Account—the customer has set up a credit card or permanent electronic funds transfer on their account so that purchases are directly charged through their account at the time of purchase.
    - Credit Card—the person enters a credit card number and authorization each time they purchase something through the site.
    - Monthly Billing—the person makes purchases periodically and receives a paper or electronic bill which they pay later either by mailing a check or through the company web site's payment system.
- **eReader_Adoption**: This attribute exists only in the training data set. It consists of data for customers who purchased the previous-gen eReader. Those who purchased within a week of the product's release are recorded in this attribute as 'Innovator'. Those who purchased after the first week but within the second or third weeks are entered as 'Early Adopter'. Those who purchased after three weeks but within the first two months are 'Early Majority'. Those who purchased after the first two months are 'Late Majority'. This attribute will serve as our label when we apply our training data to our scoring data.

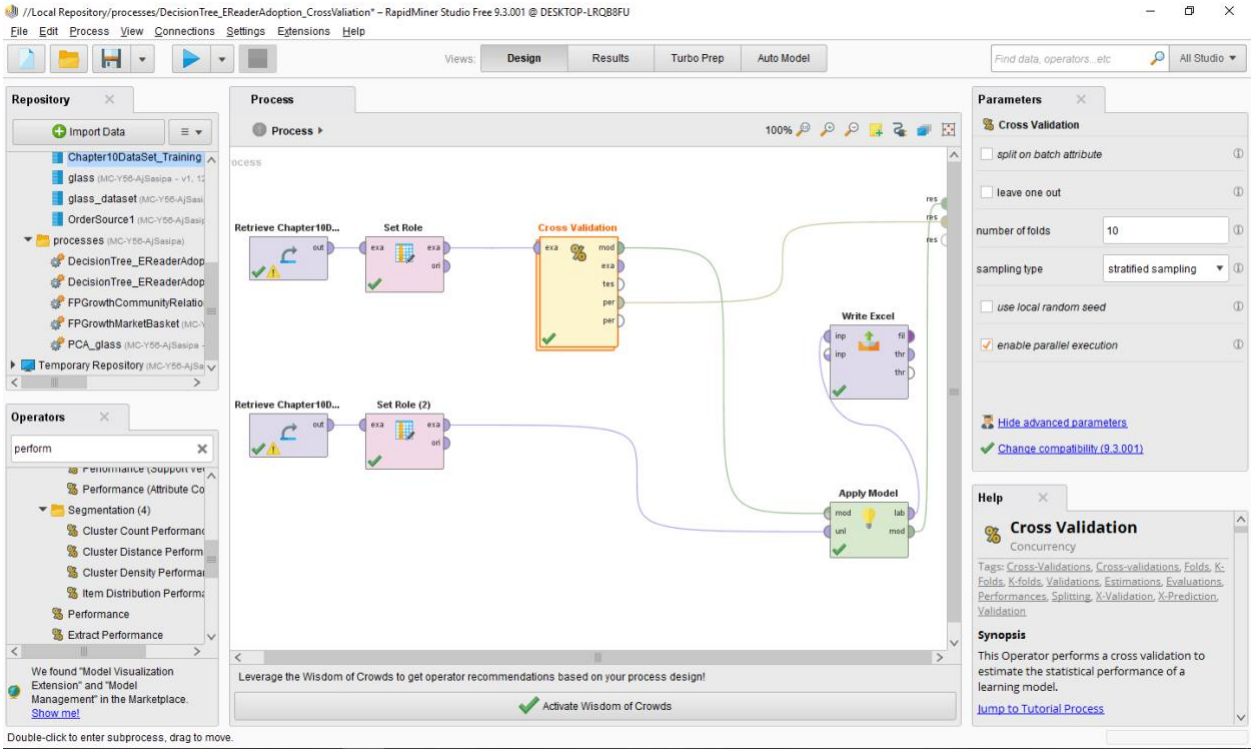
Set Role -> ID, Class label

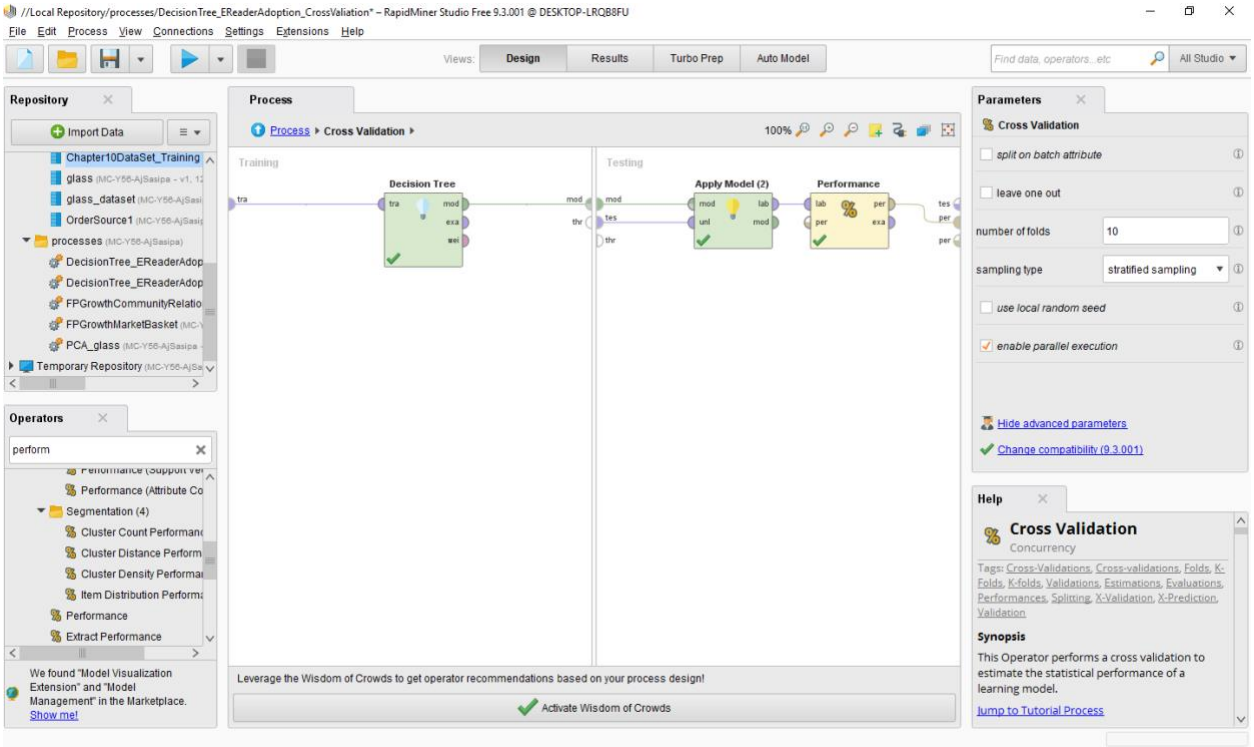


Save file ผลการทำนายเก็บไว้ด้วย

ทำต่ออีกหน่อย ด้วยการเพิ่มส่วนของ model evaluation

ให้ใช้ cross validation operator


//Local Repository/processes/DecisionTree_EReaderAdoption_CrossValiation* – RapidMiner Studio Free 9.3.001 @ DESKTOP-LRQB8FU
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc All Studio
Repository
Import Data
Chapter10DataSet_Training
glass
glass_dataset
OrderSource1
processes
DecisionTree_EReaderAdop
DecisionTree_EReaderAdop
FPGrowthCommunityRelatio
FPGrowthMarketBasket
PCA_glass
Temporary Repository
Operators
perform
Performance (Attribute Co
Segmentation (4)
Cluster Count Performanc
Cluster Distance Perform
Cluster Density Performa
Item Distribution Performa
Performance
Extract Performance
We found "Model Visualization Extension" and "Model Management" in the Marketplace.
Show me!
Process
Process
100%
Retrieve Chapter10D...
Set Role
Cross Validation
Write Excel
Retrieve Chapter10D...
Set Role (2)
Apply Model
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Cross Validation
split on batch attribute
leave one out
number of folds 10
sampling type stratified sampling
use local random seed
enable parallel execution
Hide advanced parameters
Change compatibility (9.3.001)
Help
Cross Validation
Concurrency
Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation
Synopsis
This Operator performs a cross validation to estimate the statistical performance of a learning model.
Jump to Tutorial Process
Double-click to enter subprocess, drag to move.


//Local Repository/processes/DecisionTree_EReaderAdoption_CrossValiation* – RapidMiner Studio Free 9.3.001 @ DESKTOP-LRQB8FU
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc All Studio
Repository
Import Data
Chapter10DataSet_Training
glass
glass_dataset
OrderSource1
processes
DecisionTree_EReaderAdop
DecisionTree_EReaderAdop
FPGrowthCommunityRelatio
FPGrowthMarketBasket
PCA_glass
Temporary Repository
Operators
perform
Performance (Attribute Co
Segmentation (4)
Cluster Count Performanc
Cluster Distance Perform
Cluster Density Performa
Item Distribution Performa
Performance
Extract Performance
We found "Model Visualization Extension" and "Model Management" in the Marketplace.
Show me!
Process
Process Cross Validation
100%
Training
Testing
Decision Tree
Apply Model (2)
Performance
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Cross Validation
split on batch attribute
leave one out
number of folds 10
sampling type stratified sampling
use local random seed
enable parallel execution
Hide advanced parameters
Change compatibility (9.3.001)
Help
Cross Validation
Concurrency
Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation
Synopsis
This Operator performs a cross validation to estimate the statistical performance of a learning model.
Jump to Tutorial Process

แบบฝึกหัด

Juan knows the business of athletic statistical analysis. He has seen how performance in one area, such as scoring, is often interconnected with other areas such as defense or fouls. The best athletes generally have strong connections between two or more performance areas, while more typical athletes may have a strength in one area but weaknesses in others. For example, good role players are often good defenders, but can't contribute much scoring to the team. Using league data and his knowledge of and experience with the players in the league, Juan prepares a training data set comprised of 263 observations and 19 attributes. The 59 prospective athletes Juan's team could acquire form the scoring data set, and he has the same attributes for each of these people. We will help Juan build a neural network, which is a data mining methodology that can predict categories or classifications in much the same way that decision trees do, but neural networks are better at finding the strength of connections between attributes, and it is those very connections that Juan is interested in. The attributes our neural network will evaluate are:

- Player_Name: This is the player's name. In our data preparation phase, we will set its role to 'id', since it is not predictive in any way, but is important to keep in our data set so that Juan can quickly make his recommendations without having to match the data back to the players' names later. (Note that the names in this chapter's data sets were created using a random name generator. They are fictitious and any similarity to real persons is unintended and purely conincidental.)
- Position_ID: For the sport Juan's team plays, there are 12 possible positions. Each one is represented as an integer from 0 to 11 in the data sets.
- Shots: This the total number of shots, or scoring opportunities each player took in their most recent season.
- Makes: This is the number times the athlete scored when shooting during the most recent season.
- Personal_Points: This is the number of points the athlete personally scored during the most recent season.
- Total_Points: This is the total number of points the athlete contributed to scoring in the most recent season. In the sport Juan's team plays, this statistic is recorded for each

point an athlete contributes to scoring. In other words, each time an athlete scores a personal point, their total points increase by one, and every time an athlete contributes to a teammate scoring, their total points increase by one as well.

- Assists: This is a defensive statistic indicating the number of times the athlete helped his team get the ball away from the opposing team during the most recent season.
- Concessions: This is the number of times the athlete's play directly caused the opposing team to concede an offensive advantage during the most recent season.
- Blocks: This is the number of times the athlete directly and independently blocked the opposing team's shot during the most recent season.
- Block_Assists: This is the number of times an athlete collaborated with a teammate to block the opposing team's shot during the most recent season. If recorded as a block assist, two or more players must have been involved. If only one player blocked the shot, it is recorded as a block. Since the playing surface is large and the players are spread out, it is much more likely for an athlete to record a block than for two or more to record block assists.
- Fouls: This is the number of times, in the most recent season, that the athlete committed a foul. Since fouling the other team gives them an advantage, the lower this number, the better the athlete's performance for his own team.
- Years_Pro: In the training data set, this is the number of years the athlete has played at the professional level. In the scoring data set, this is the number of year experience the athlete has, including years as a professional if any, and years in organized, competitive amateur leagues.
- Career_Shots: This is the same as the Shots attribute, except it is cumulative for the athlete's entire career. All career attributes are an attempt to assess the person's ability to perform consistently over time.
- Career_Makes: This is the same as the Makes attribute, except it is cumulative for the athlete's entire career.
- Career_PP: This is the same as the Personal Points attribute, except it is cumulative for the athlete's entire career.

- Career_TP: This is the same as the Total Points attribute, except it is cumulative for the athlete's entire career.
- Career_Assists: This is the same as the Career Assists attribute, except it is cumulative for the athlete's entire career.
- Career_Con: This is the same as the Career Concessions attribute, except it is cumulative for the athlete's entire career.
- Team_Value: This is a categorical attribute summarizing the athlete's value to his team. It is present only in the training data, as it will serve as our label to predict a Team_Value for each observation in the scoring data set. There are four categories:
    - Role Player: This is an athlete who is good enough to play at the professional level, and may be really good in one area, but is not excellent overall.
    - Contributor: This is an athlete who contributes across several categories of defense and offense and can be counted on to regularly help the team win.
    - Franchise Player: This is an athlete whose skills are so broad, strong and consistent that the team will want to hang on to them for a long time. These players are of such a talent level that they can form the foundation of a really good, competitive team.
    - Superstar: This is that rare individual who gifts are so superior that they make a difference in every game. Most teams in the league will have one such player, but teams with two or three always contend for the league title.

Dataset - Chapter11DataSet_Training, Chapter11DataSet_Scoring

ลองทำด้วย Neural Network และใช้ split validation operator แทน cross validation operator