ReservePrism's ENVISION

Predictive Modeling and Big Data Processing Platform

ENVISION is an enterprise level predictive modeling platform designed for sophisticated calculation and big data analysis. It is powered by open source R, parallel computing and grid scheduling techniques.

ENVISION Workflow:

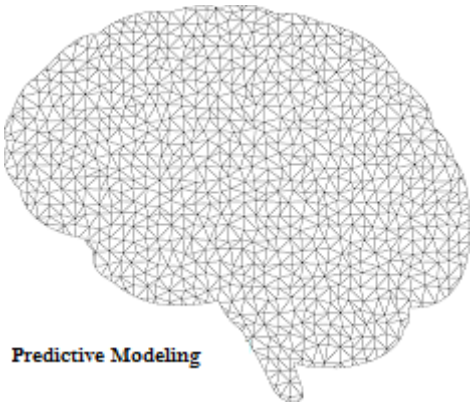
Data Processing Feature Extraction, Missing Data Treatment, Principal Component Analysis, Data Normalization, Collinearity Treatment, and Categorical to Dummy Mapping

Load Data ENVISION uses two types of data: (1) Historical data and (2) Application data, which is the data for which we want to predict results. ENVISION explores the relationships and patterns using Historical data. Then it performs predictions on the Application data based on the patterns found in the historical data.

Set Up Formula The formula defines the response variable (Y) you want to predict as a function of explanatory variables (X). $\mathbf{Y = f(X_1, X_2, X_3, \ldots, X_n)}$

Run Models Currently 12 models are available in Envision, covering both supervised and unsupervised learning.

Decision Making Models are compared based on statistical measures, and final prediction results are processed for your business decision-making. For classification analysis, models are validated by precision, recall and F-measure constructed from the confusion matrix. For regression problems, root-mean-square error (RMSE) and Kolmogorov–Smirnov (K-S) tests are used.

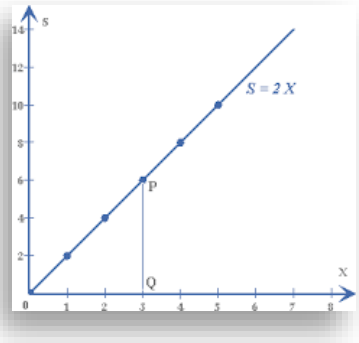


Predictive Modeling

ENVISION Predictive Models:

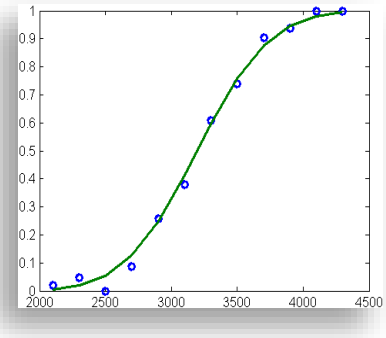
Supervised learning: We know the response variable Y.

Linear Regression Model



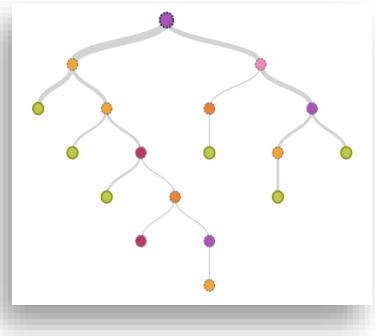
Linear regression is the most basic regression model used in the statistical world. It assumes a linear relationship between the response variable Y and one or multiple explanatory variables (X), with an error term that follows the normal distribution.

Generalized Linear Model



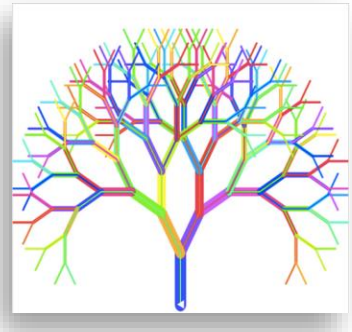
The Generalized Linear Model (GLM) extends the linear regression model by allowing the error function to belong to distributions other than normal distribution. The response variable Y can be transformed using a link function before being fed into the linear equation. A famous form of GLM is Logistic regression by assuming the error function following the Binomial distribution and a logit link function.

Classification and Regression Trees



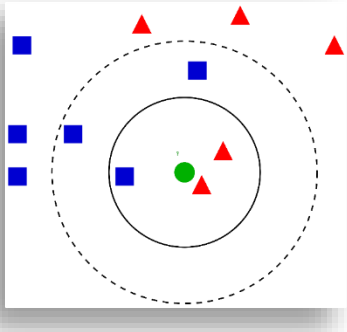
Classification and Regression Trees (CART) model is an extension of decision tree model. It constructs a tree to split the data using explanatory variables (X). The splitting is determined to maximize the accuracy improvement. CART can be used for both classification and regression (taking the mean of all cases in the terminal nodes).

Random Forest



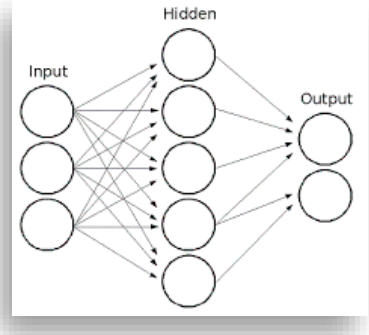
The Random Forest model is an ensemble method based on CART. The data is sampled into small subsets with a CART model fitted to each subset. Each CART model generates its own prediction. The final estimate is based on the average result of all CART models.

K-Nearest Neighbors



K-Nearest Neighbours (KNN) is a nonparametric model that use the distance between data points to determine their closeness. It can be used for classification based on the clusters and regression based on the representative value of the clusters (mean, median, etc.)

Artificial Neural Network



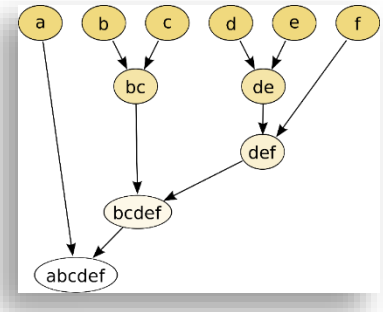
Artificial Neural networks (ANN) are one of the most fascinating machine-learning models. An ANN model mimics the network of neurons in brain and can approximate any relationships (linear or nonlinear). Multiple hidden layers also enable us to apply a deep learning algorithm for sophisticated relationship modeling.

Unsupervised learning: We do not know the response variable Y. The data can be used for clustering analysis, data dimension reduction, and relationship exploration.

Association Rules

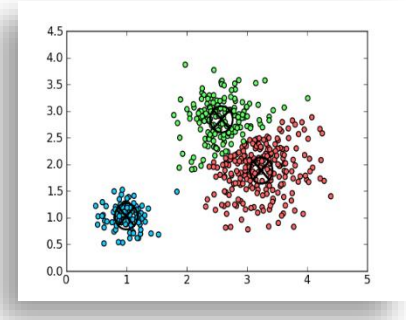
Association Rules is a learning method to discover relationships among item sets. For example, what are the usual combination of shopping items for supermarket customers? Is beer often sold with diapers? What product combinations are popular?

Hierarchical Clustering Model



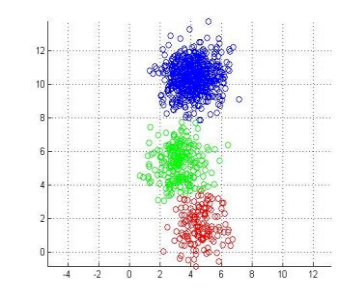
Hierarchical clustering builds a hierarchy of clusters like a tree and does not require that the number of clusters be specified at the start. Each data point starts as a cluster, and clusters are grouped together gradually when moving up the hierarchy.

K-Means Model



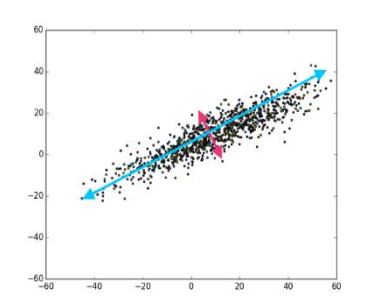
K-means clustering partitions the dataset into several clusters based on the distance between each data point and the center of each cluster. The center of a cluster is determined as the mean of all the data points belonging to that cluster.

K-Medoids Model



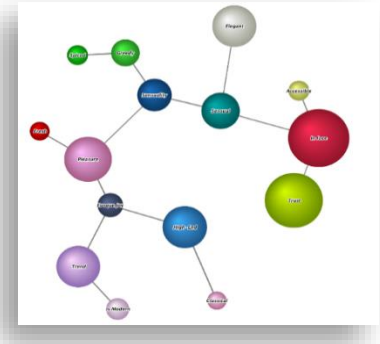
K-medoids clustering partitions the dataset into several clusters based on the distance between each data point and the center of each cluster. Unlike K-means, a cluster's center is a real data point rather than the mean of the data points in that cluster.

Principal Component Analysis



Principal component analysis transforms the dataset into orthogonal vectors (uncorrelated variables) that explain the variance of the dataset. Usually the first few principal components explain over 99% of the total variance. Principal components may be used as explanatory variables to reduce the dimension of the dataset.

Bayesian Network



A Bayesian network is a graphical model that studies the interdependency of the variables in a dataset. It will construct a directed acyclic graph with probability functions for each node. It can help discover cause-and-effect relationships.

ENVISION Big Data Processing

Grid Scheduling (Pull Technology)

Getting prediction results in a timely manner is important for exploiting the business value of your data. ENVISION Grid scheduling allows users to make runs at the same time using multiple machines for multiple models. In Grid Scheduling, a pool of jobs is created by the user. Then available computing resources take jobs from the pool until it is empty. It is a job pulling scheme instead of a pushing scheme. The results are then summarized, compared, and presented.

Random Ensemble on GLM, ANN and CART

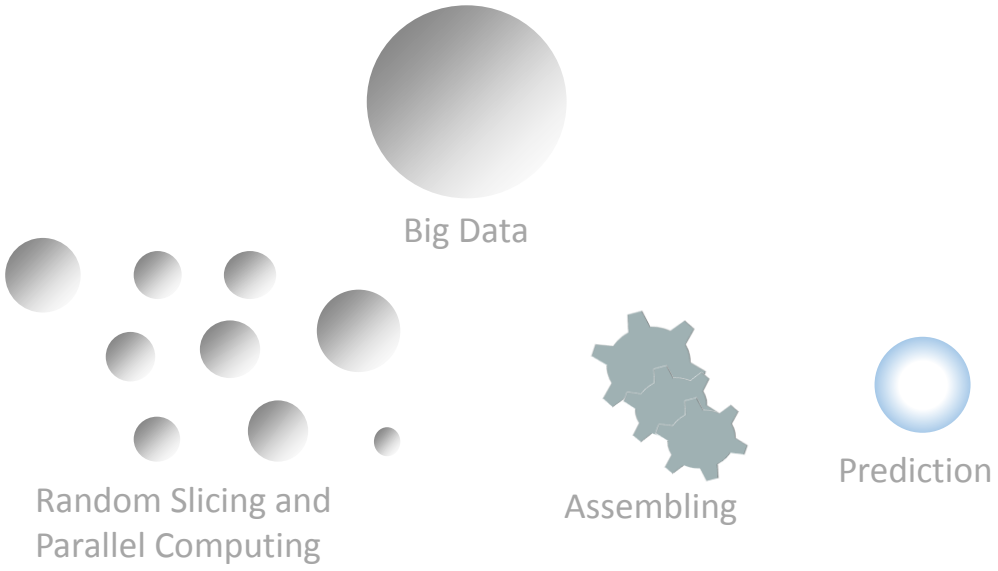
Random ensemble, constructed upon the Parallel Grid Scheduling technique, is our innovative solution to efficiently overcome the big data challenge. The model slices the big data into small random datasets, then each subset is used to train a supervised learning model. The final prediction results are based on the average results of trained models.

The size of the small sampled dataset can be set according to the memory and calculation power of the machines. It is recommended that each dataset has a size less than 500 MB. To make sure that most data records will be included in at least one random dataset, the number of datasets is large

all

enough so that, on average, one data record will be included in 3 sample datasets.

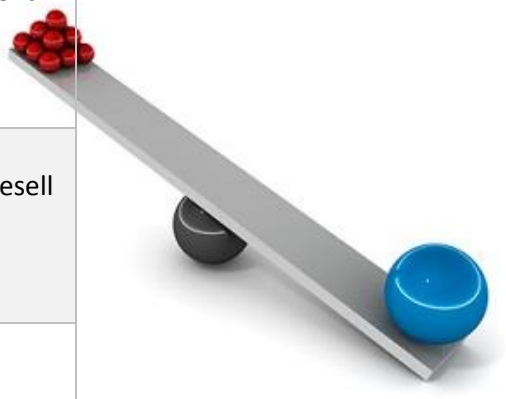
Random ensemble is available for **GLM, ANN and CART** models in ENVISION. In addition to random ensemble, other techniques such as stochastic gradient descent is applied in system by leveraging on the advanced **h2o** platform.



Random Ensemble Technique

ENVISION Insurance Applications

Pricing More pricing factors	**Reserving** Claim classification Case reserve adequacy assessment False claim identification Claim closure/reopen Claim size prediction
Underwriting High risk case identification Automatic underwriting	**Marketing** Customer retention, renew, or resell Personalized product recommendation Customer categorization
Risk Analysis Fraud detection Credit score	**Business Disruption** UBI Health/Fitness discount



- ENVISION has Real-world case studies that demonstrate the business application of predictive modeling.
- Using ENVISION as an enterprise level predictive modeling platform, people can easily share projects and streamline the analytical process within our customized support and solutions.
- ENVISION is a platform that enriches big data analysis by implementing a highly scaled database solution.
- Using ENVISION can shorten your predictive modeling learning curve and model building time. Your team can spend more time on creative business solutions. The module includes all aspects of predictive modeling, including data processing, model selection, model validation and prediction.
- Our team has been working on many real-world predictive modeling applications and research projects. Our experience will be helpful for your data analytics work.