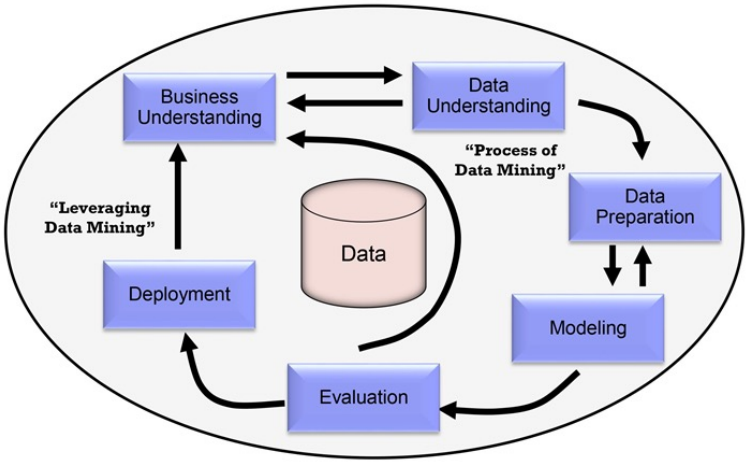
KNIME Project Example

MortgCo’s
Mortgage Refinancing Business Expansion

Created by Michael Kim

Phase 1. Business Understanding



6 Phases of CRISP-DM

- MortgCo's business problem is *to find informative & descriptive attributes for profiling households and to build a predictive model* which will discriminate between:
  - Households which own house with mortgage or loan (include home equity loans)
  - Households that own their homes free and clear
- I have deployed the ***CRISP-DM*** *approach with six phases* that is widely used on market.

Phase 2 & 3. Data Understanding & Preparation

- 2012-2016 Public Use Microdata Sample (PUMS) files have been used for Data Mining. PUMS dataset are a set of *records from housing units in Pennsylvania.*
- I pre-processed PUMS dataset by *eliminating irrelevant data* (columns & rows) and *replacing missing values*.
- A total of 97,556 data have been used: *60% of data as a training set* and 40% of data as a test set**.

*** training set:** a subset to train a model.
*** test set:** a subset to test the trained model.

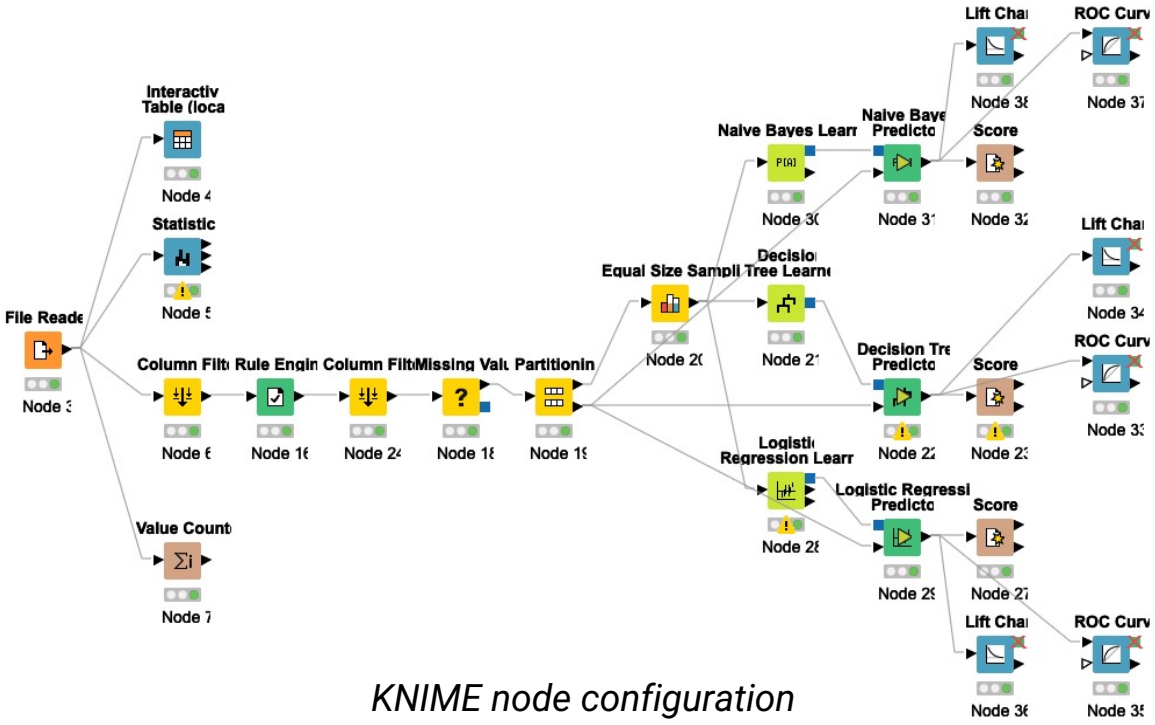
Phase 2 & 3. Data Understanding & Preparation (cont.)

- First, data features that contains information *closely related to mortgage are eliminated*, otherwise they will affect the prediction accuracy for classifying targeting households. Removed features are MRGI, MRGP, MRGT, MRGX, SMP, SMX, FMRGIP, FMRGP, FMRGTP, FMRGXP, FSMP, FSMXSP, etc.. *Other irrelevant features are also eliminated*.

- Using *forward feature selection technique**, data features are carefully evaluated with *correlation coefficients* and I have chosen the following variables:
  i. MV – When moved into this apartment
  ii. BLD – Units in structure
  iii. RMSP – Number of rooms
  iv. BDSP – Number of bedrooms
  v. TYPE – Type of unit
  vi. NP – Number of persons associated with this housing record

  * **Forward feature selection:** an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
- Missing values have been replaced with *most frequent values*.

Phase 4. Modeling: 3 Data Mining Techniques developed



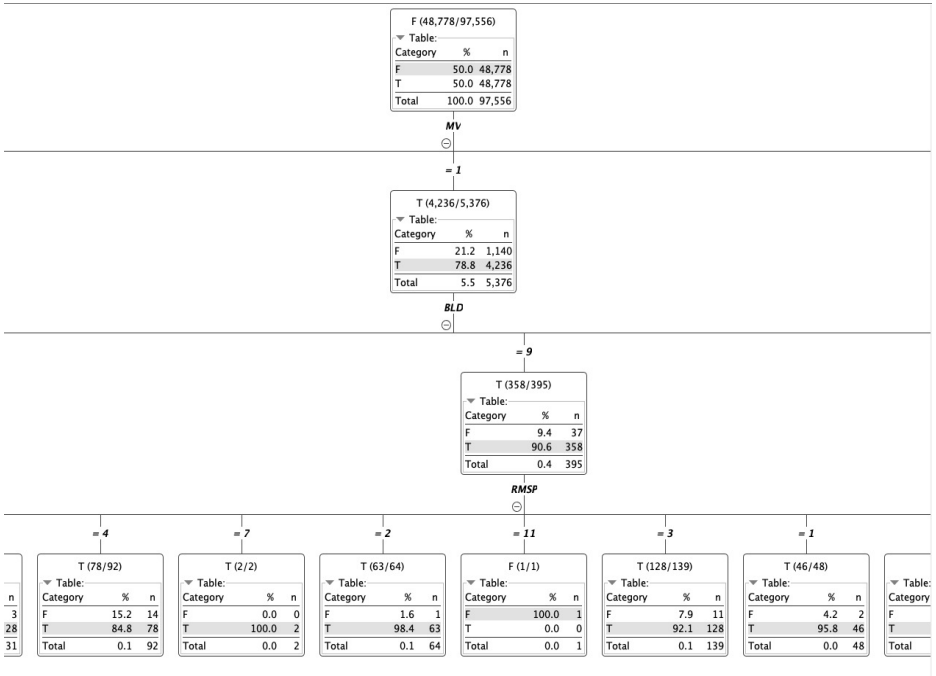
KNIME node configuration

- Three (3) *supervised learning classifiers* are developed.
- **Naive Bayes**
  Correct classified: 94,208
  Wrong classified: 32,759
  Accuracy: 74.2 %
- **Decision Tree**
  Correct classified: 94,763
  Wrong classified: 31,396
  Accuracy: 75.1 %
- **Logistic Regression**
  Correct classified: 97,157
  Wrong classified: 29,810
  Accuracy: 76.5 %

Phase 5. Evaluation: The Chosen Machine Learning Algorithm

- **What is a Decision Tree ?**
  One of the most popular machine learning algorithms. A decision tree will *identify which of the attributes or characteristic features has the highest predictive value*.

- **Why Decision Tree is chosen ?**
  - The performance of Decision Tree model is relatively lower than that of Logistic Regression in accuracy. However, *Decision Tree* model was *chosen because*
  - it is very *intuitive and easy to explain* to decision makers with *no complex formulas*.
  - It is *easy to understand* with only brief explanations.
  - A Decision Tree model follows the *same pattern of thinking* that humans use when making decisions.



Screenshot of the Decision Tree Model

Phase 5. Evaluation (cont.):

Performance and Accuracy of Decision Tree Algorithm

MORTGAG...	T	F
T	68121	25447
F	5949	26642

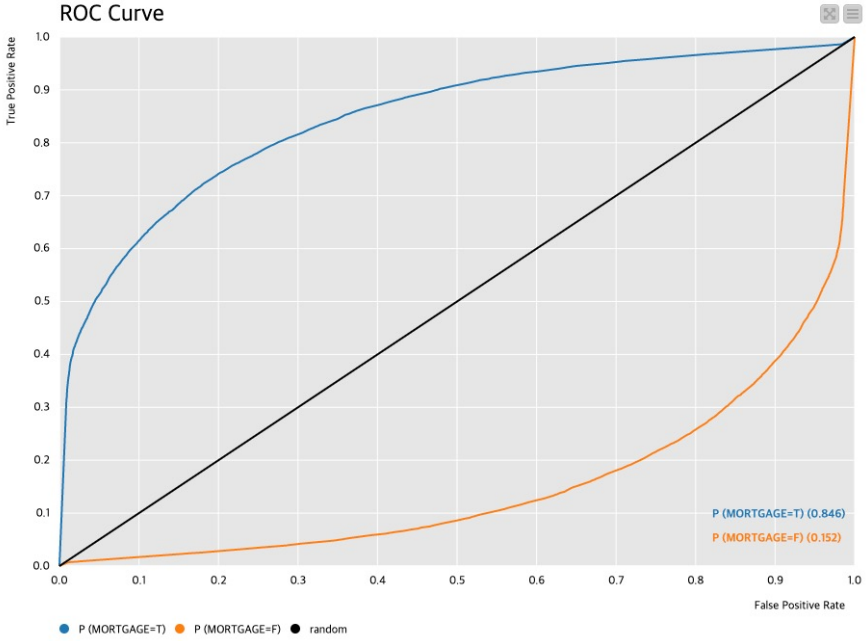
Correct classified: 94,763

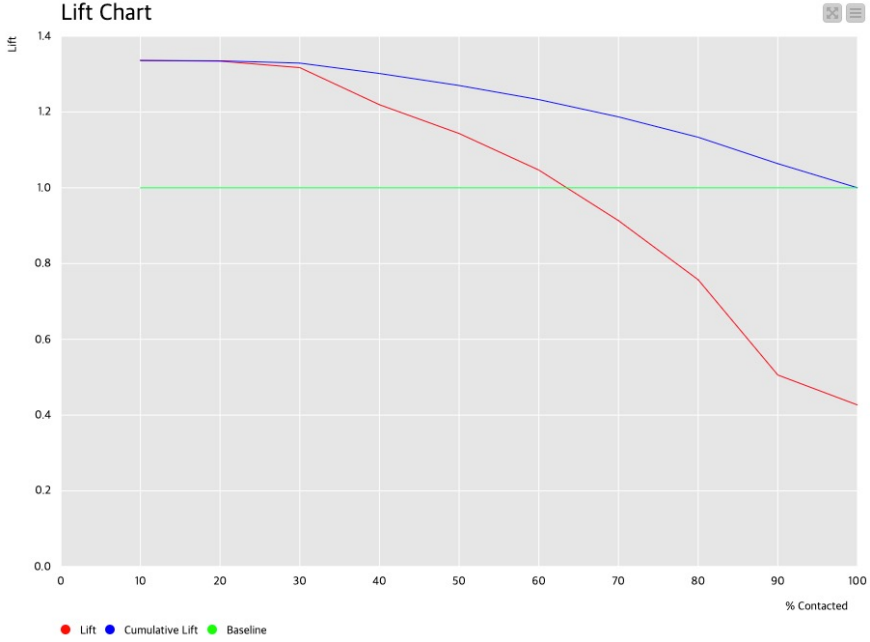
Wrong classified: 31,396

Accuracy: 75.114 %

Error: 24.886 %

Cohen's kappa (κ) 0.457





Phase 6. Deployment : Conclusions

- ***Initiate the data mining project to build a predictive model (Decision Tree classifiers).***

- Three (3) predictive models for *supervised-learning & classification* task have been created & evaluated – Naïve Bayes, Decision Tree, and Logistic Regression.
  - *Decision Tree model* has been chosen.

- The Decision tree model chosen above can predict the households that owns their houses with mortgage or loan in Pennsylvania with the *accuracy of around 75.1%.*

- MortgCo can discover the *important profiling features* applied to targeting households for mortgage refinancing business expansion.

- These attributes are such as *MV (When moved into this house), BLD (Units in structure), RMSP (Number of rooms), NP (Number of persons), BDSP (Number of bedrooms), TYPE (Type of unit),* etc.