**Submitted by: Lai Minxian, G2103336A**

**First**, examine the dataset to see if you can spot any systematic differences in the actions taken or demographics of the populations that are served by the different financial institutions. You can use the LEI column as the identifier of the reporting mortgage origination organization.

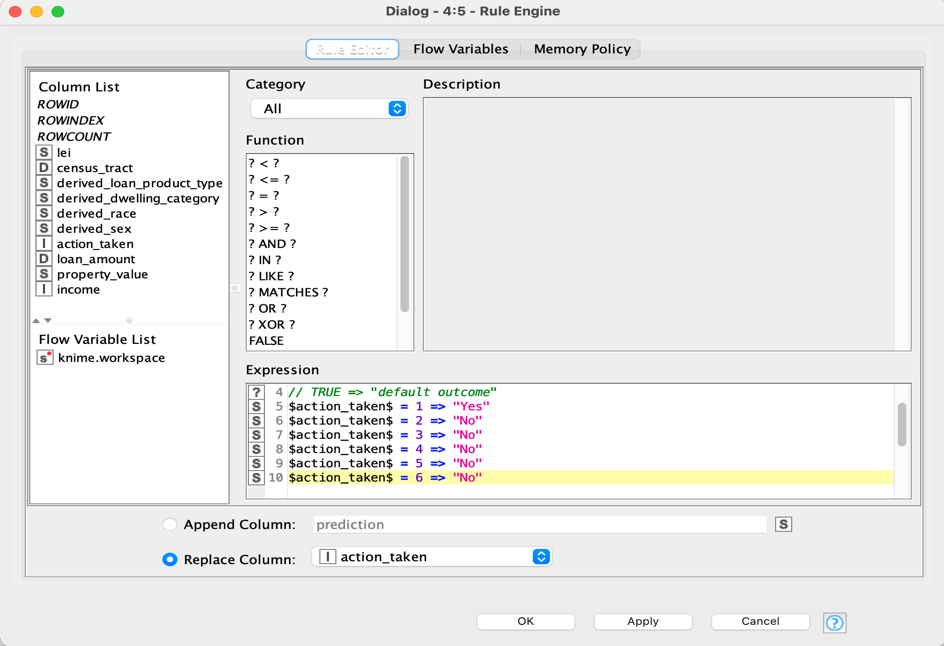
* In terms of systematic differences, such as ascribed to applicants’ race, demographics and other attributes in relation to the actions taken for the loan outcomes, there are some notable observations as follows derived from using:

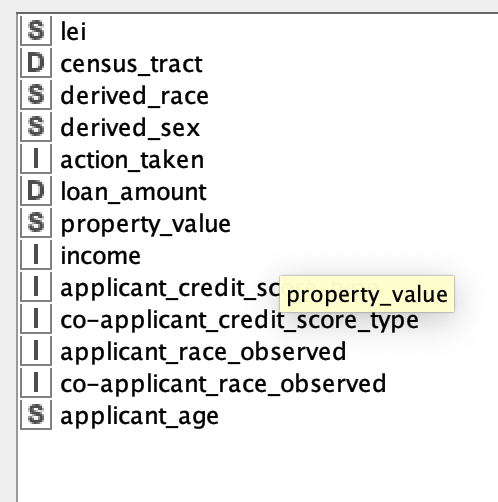
Statistics node on KNIME after converting the file to CSV format

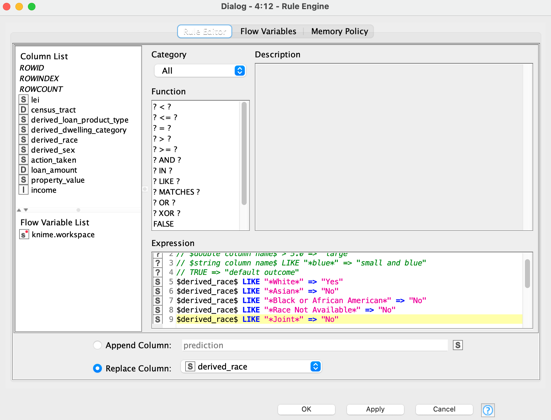
Observations:

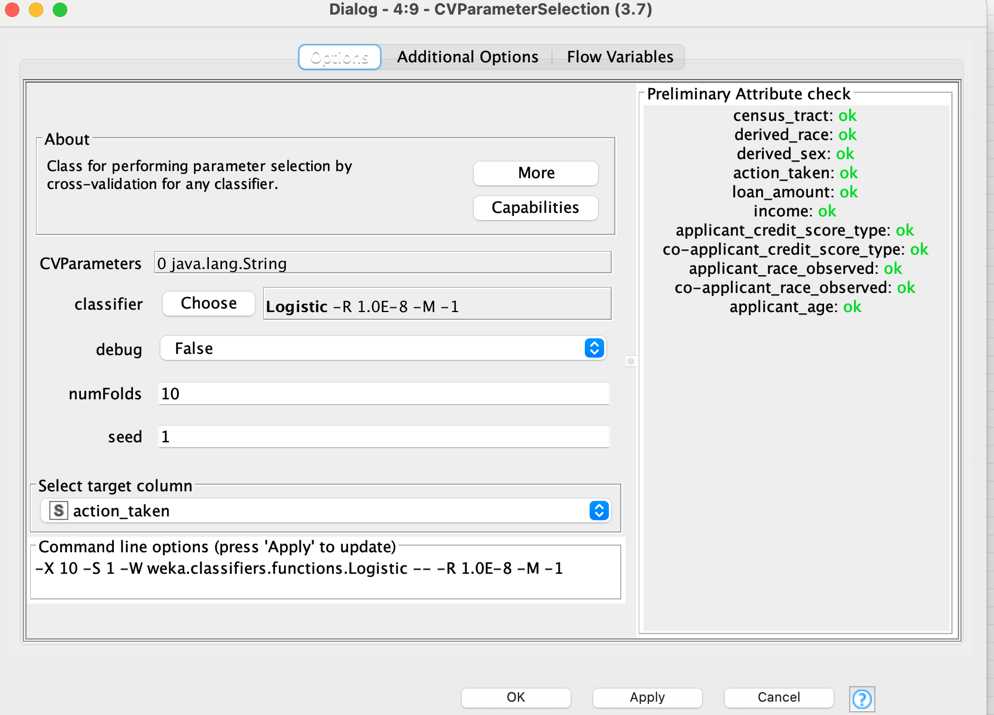
* Identifier/bank ‘KB1H1DSPRFMYMCUFXT09’ gives out the highest number of loans approved/originated at 6,866 (1. Loan Originated). Slightly over half of the successful borrowers consists of well-to-do Whites (52.5%) on joint applicant basis (30.4%), virtually all of whom live in Single Family (1-4 units) site-built landed properties (99.5%). Close to 100% of these successful borrowers have their conventional first lien mortgage loans already pre-approved, with most located at TN, 47157, 47157020900.
* By contrast, for identifier/bank ‘5493001GAASLCSP1PB64’ with up to 5,472 mortgage cases, it contains the highest number of purchased loans at 15.4% (6. Loans Purchased). A quarter of the applicants’ race is not identifiable; however, it has identified that they are mostly males (30.4%) living in Single Family (1-4 units) manufactured properties (0.003%), who applied for mostly FHA: First Lien mortgages (16.1%).
* The above suffices to paint a picture of a successful mortgage borrower, based on certain baseline attributes: race, residential address (by extension, income), and even the mode of mortgage repayment, which is overwhelmingly on balloon (re)payment basis (94.4%), interestingly.
* With some of these basic parameters or data fields mapped out and identified, the next steps is to employ these in a logistics regression model to train and test the data.

**Second**, describe your methods as you construct a model that you deem best fit to predict the Action Taken by the reporting institutions using the data fields that you find sensible. It is likely that you will build multiple models on your way towards choosing the “best” by your own yardstick.

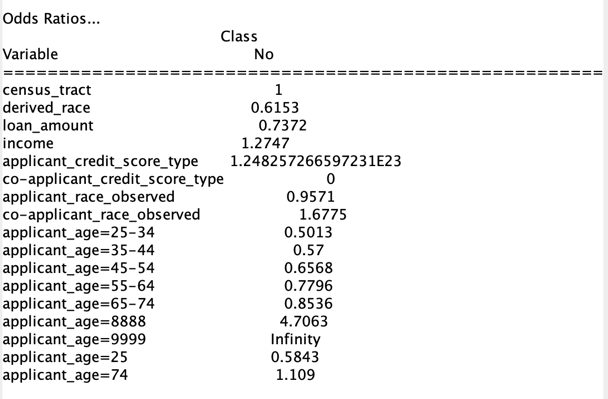
* In view of the above observations, it’s intuitive enough to allot the predictor variables through a binomial or binary logistic regression approach, ie, dealing with situations whereby the observed outcome for a response variable – which, in this case here, is the mortgage application outcome (ie, actions taken from 1-6) – can only have two possible types, "0" or “No” and "1" or “Yes”.
* Therefore, for the response variable:
* “1” (loan approved eventually) will be renamed as “Yes” using the Rule Engine node in KNIME
* “2-6” (loan not approved eventually) will be renamed as “No” using the Rule Engine node in KNIME
* 
* For the predictor variables, having done extensive trial and error through back-testing, they are as determined can be gleaned from the left side of the screenshot as follows, namely:



* As the rest of the data fields are more or less secondary derivatives of these and would serve to add unwanted “noise”, they won’t be used in the modelling.
* This is part of the first step, ie, data pre-processing.
* Next, the Race predictor variable will be binarised to either “White” (“Yes) or “Non-White” (“No”) using the Rule Engine node (see below)
* 
* Subsequently, the partition will be carried out as per the instruction: 90% of the data to be trained, and 10% to be tested and reported via the response variable, with a default random seed of 1234 for doing stratified sampling:
* Subsequently, the numeric values (non-string) parameters will be normalised using z-score normalisation. This is parsed and carried out by KNIME automatically, with a few self-inputted calibrations on the settings, such as selecting Logistics Regression from the various models available, and sticking to default settings as pre-filled accordingly.
* Next, having completed the z-score normalisation process, the predictor variables are checked through by the application to make sure everything is deemed ‘ok’ prior to running cross validation on the allotted 10%.



**Third**, consider if you can construct a model that can predictably find systematic examples of discrimination by financial institutions based on protected attributes. For example, is there a way to quite accurately predict action taken based on some subset of these attributes or combinations of them?

* From the Logistics Regression model development, the results of the trained model point to some level of discrimination based on attributes like age, income, race (to a lesser extent as initially thought).
* By far, the observable attributes found from the trained model linked to the mortgage loan outcome appear to be age (the higher the age of the applicant, the higher the odds to get a successful mortgage loan), income, credit score, loan amount, and derived race. 
* Somewhat against conventional thinking, the pure monetary attributes rather than those related to protected attributes like race seem to be given more weightage in the banks’ assessments towards whether mortgagee is able to repay back the loan, thereby, influencing their ultimate decision on the loan provision outcome.
* The accuracy of the model is 75.16%, after doing several trial and error and tweaks of the predictor variables (see Confusion Matrix below extracted from KNIME below).
* The Cohen’s kappa is 0.5%, which is acceptable in terms of intra-class correlation (ICC).
* In summary, using Binary/Binomial Logistics Regression, there is a way to quite accurately (75%) predict action taken based on some subset of these attributes or combinations of them.
* 