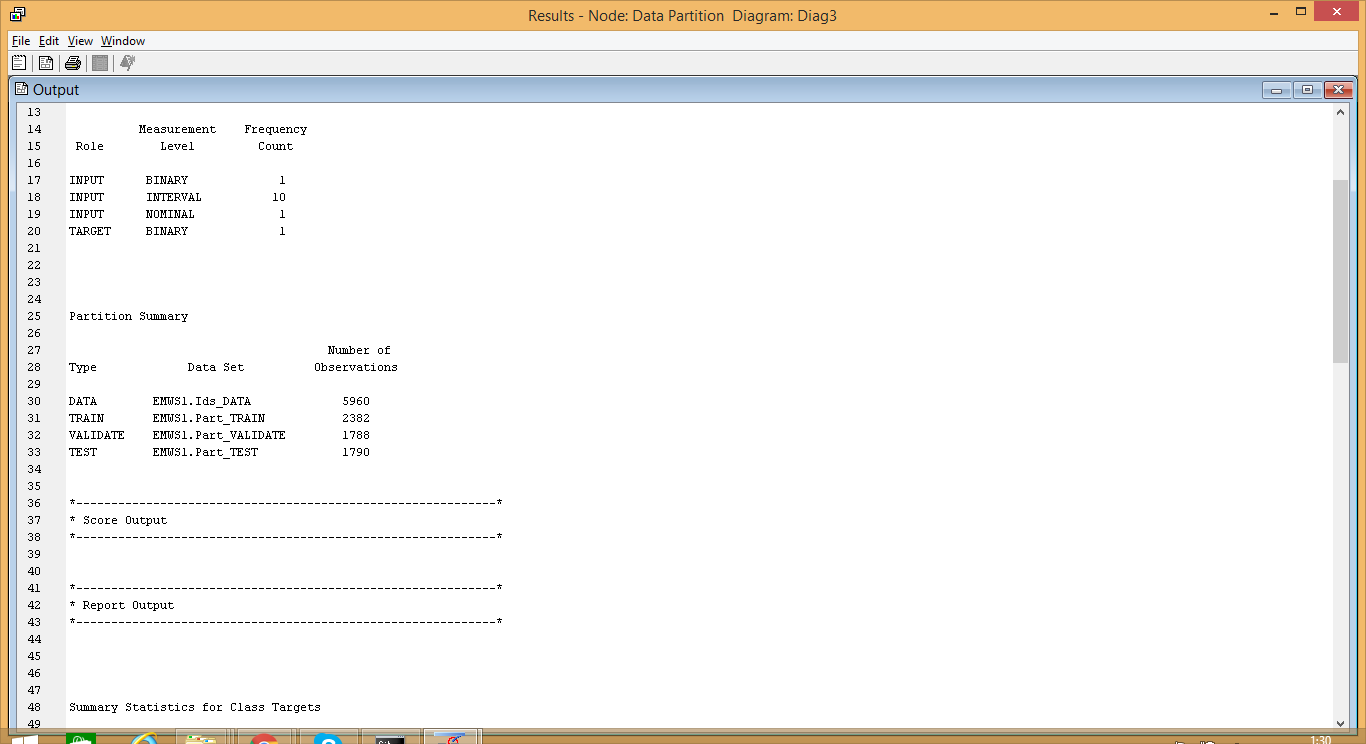
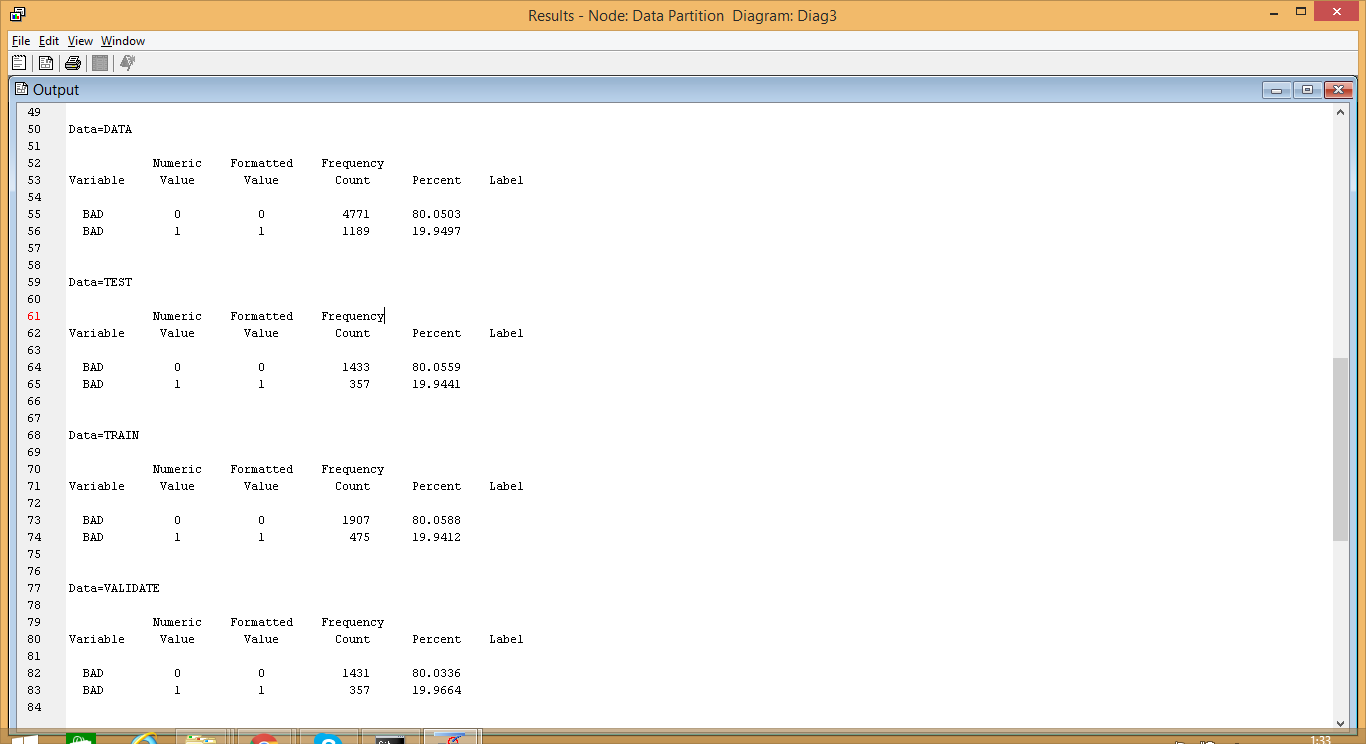
***Формирование обучающих и контрольных выборок***

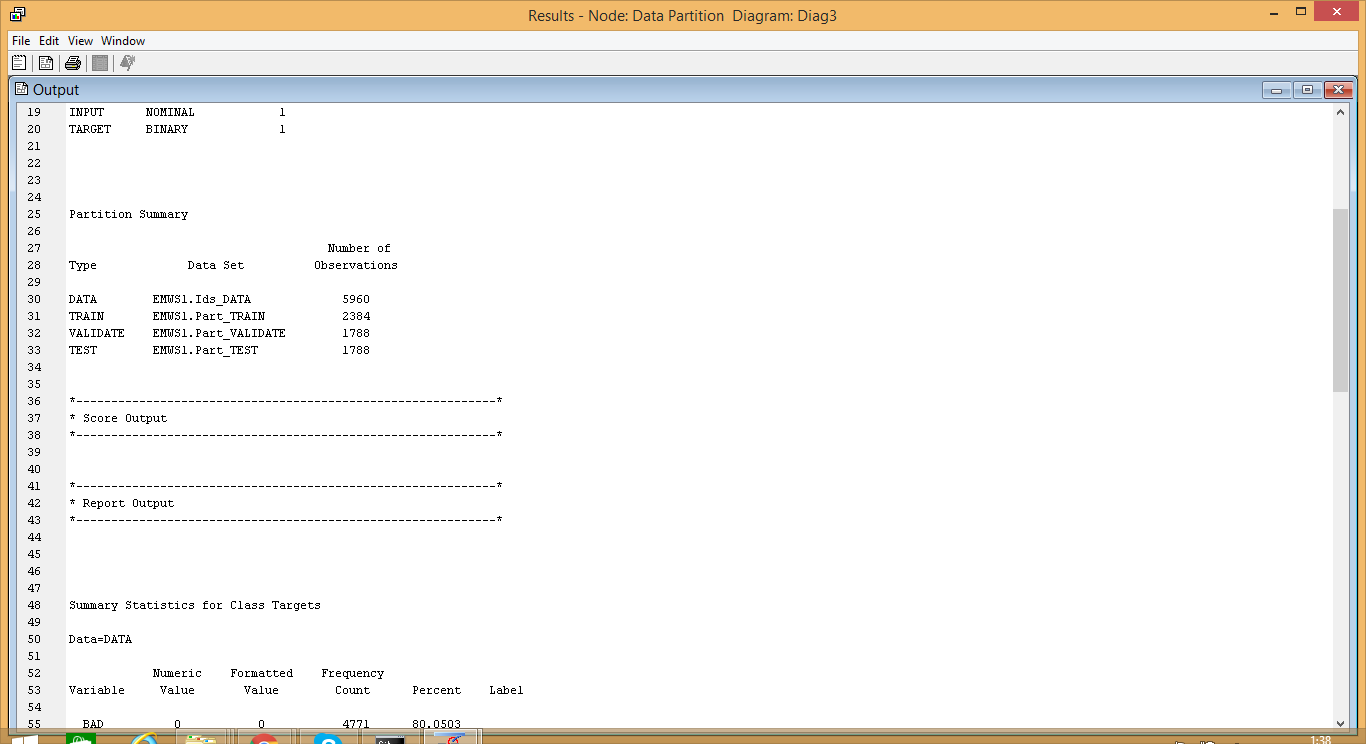
* *Сколько наблюдений попало в каждый из наборов? Почему?*

В SAS Enterprise Miner метод разделения данных по умолчанию (default) для целевых переменных класса заключается в стратификации целевой переменной. Наблюдения случайным образом выбрасываются так, чтобы сохранить распределение переменной стратификации.

* *Какова пропорция положительных и отрицательных откликов в каждом? Почему?*

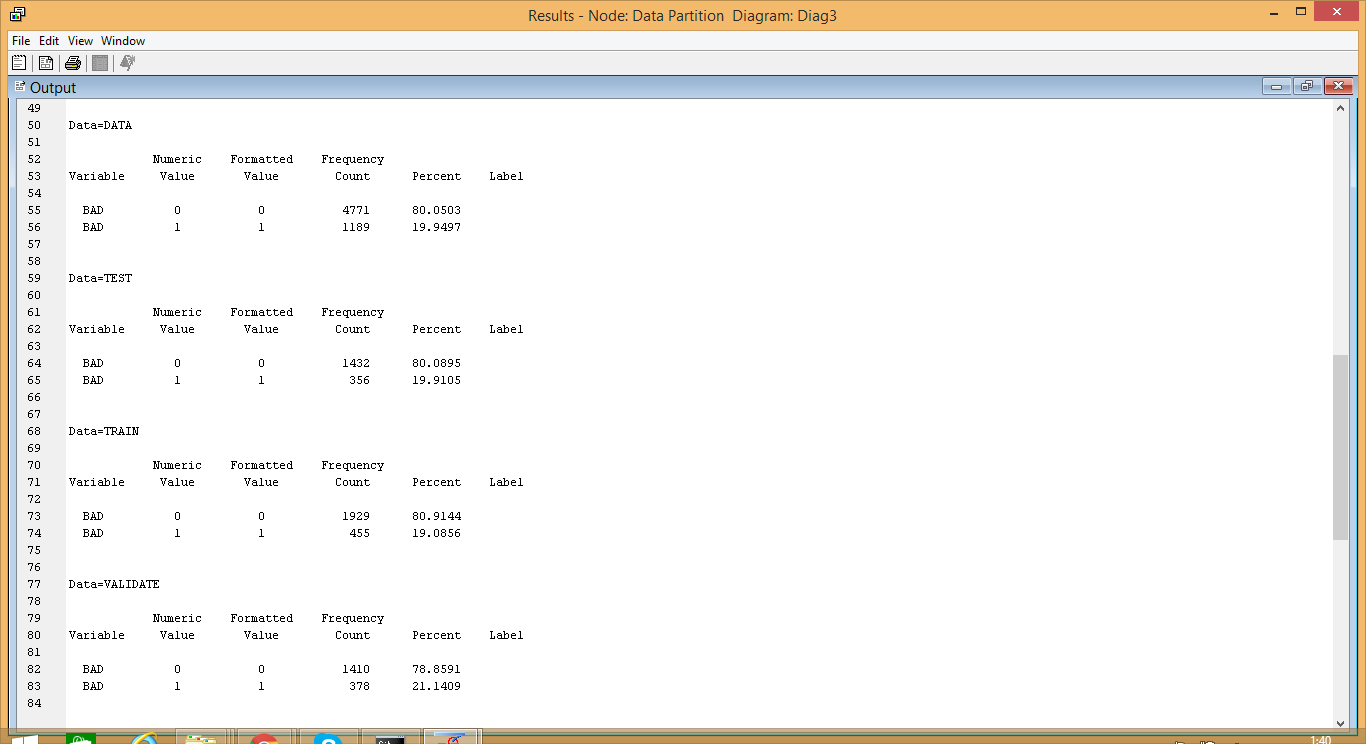
Относительно основного набора данных пропорция отклика в каждом из наборов сохраняется (тренировочный, тестовый и валидационный).

* *Как изменится число наблюдений в наборах, если метод разбиения поставить Simple Random? Почему?*

**

Так как при Simple Random - каждое наблюдение в наборе имеет одинаковую вероятность попасть в любой набор.

* *Как изменится пропорция отклика, если метод разбиения поставить Simple Random? Почему?*

**

Пропорция отклика изменилась несущественно, так, как метод разделения Simple Random поделил исходный набор данных ровно на 40-30-30.

Так, общее количество откликов в каждом из наборов поменялось.

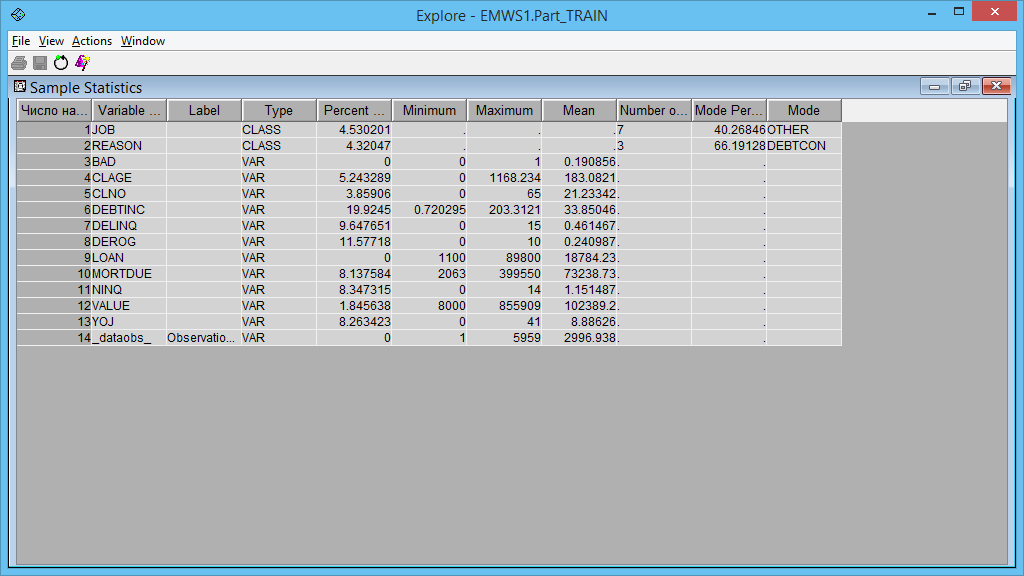
***Поиск и удаление выбросов***

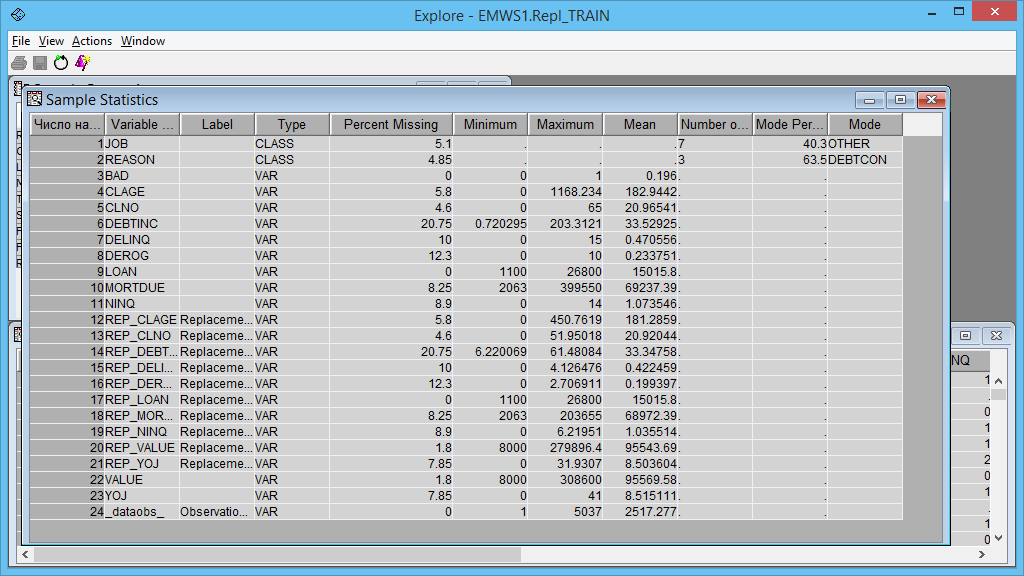
* *Как будет формироваться интервал допустимых значений для числовых переменных при такой настройке?*

Standard Deviation From the Mean - устраняет значения, превышающие n стандартных отклонений от среднего.

* *В списке “Imported Data” и “Exported data” для роли train посмотрите список переменных. Что изменилось? Как изменилась роль исходных переменных?*

**Imported Data**

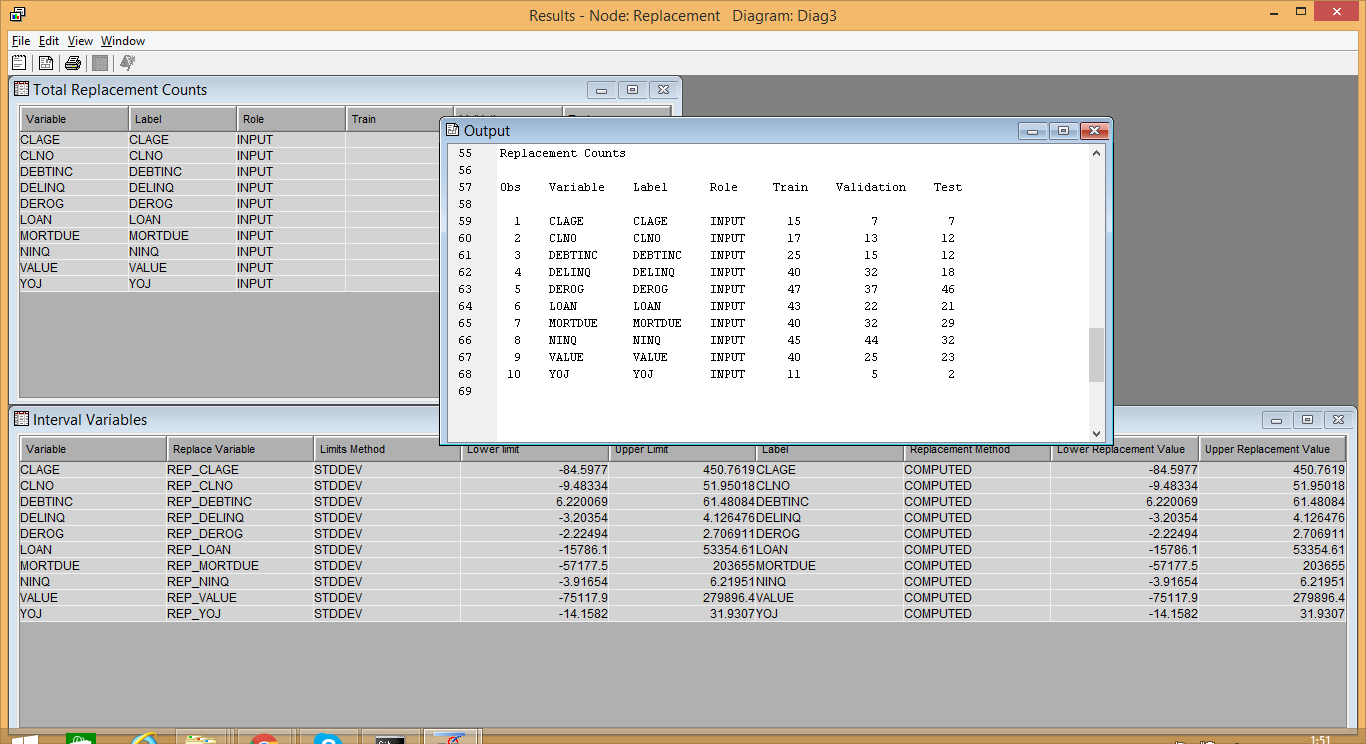


**Exported data** 

Число переменных увеличилось.

* *Как изменилось число пропусков для переменных с префиксом REP\_?*

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | REP\_ | Percent Missing | REP\_Percent Missing |
| DEBTINC | REP\_DEBTINC | 20,75 | 20,75 |
| DEROG | REP\_DEROG | 12,3 | 12,3 |
| DELINQ | REP\_DELINQ | 10,0 | 10,0 |
| NINQ | REP\_NINQ | 8,9 | 8,9 |
| MORTDUE | REP\_MORTDUE | 8,25 | 8,25 |
| YOJ | REP\_YOJ | 7,85 | 7,85 |
| CLAGE | REP\_CLAGE | 5,8 | 5,8 |
| CLNO | REP\_CLNO | 4,6 | 4,6 |
| VALUE | REP\_VALUE | 1,79 | 1,79 |
| LOAN | REP\_LOAN | 0,0 | 0,0 |



* *Как думаете, у переменных в тренировочном и тестовом наборе интервалы допустимых значений будут одинаковыми или нет? Почему?*

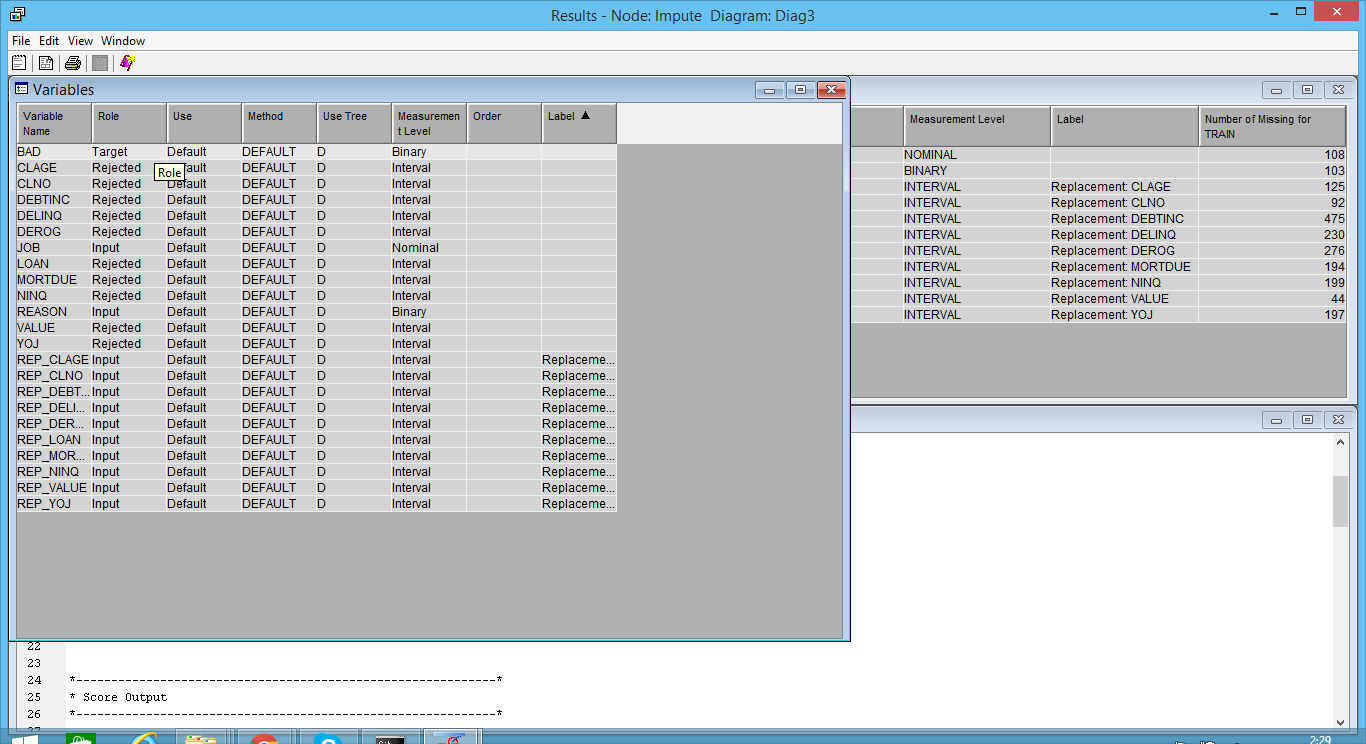
Интервал будет одинаковый. Так как Enterprise Miner по умолчанию использует образец из набора данных обучения, чтобы выбрать значения для замены.

***Подстановка пропущенных значений***

* *Переменные с какими префиксами добавились после работы этого узла?*

С префиксом IMP

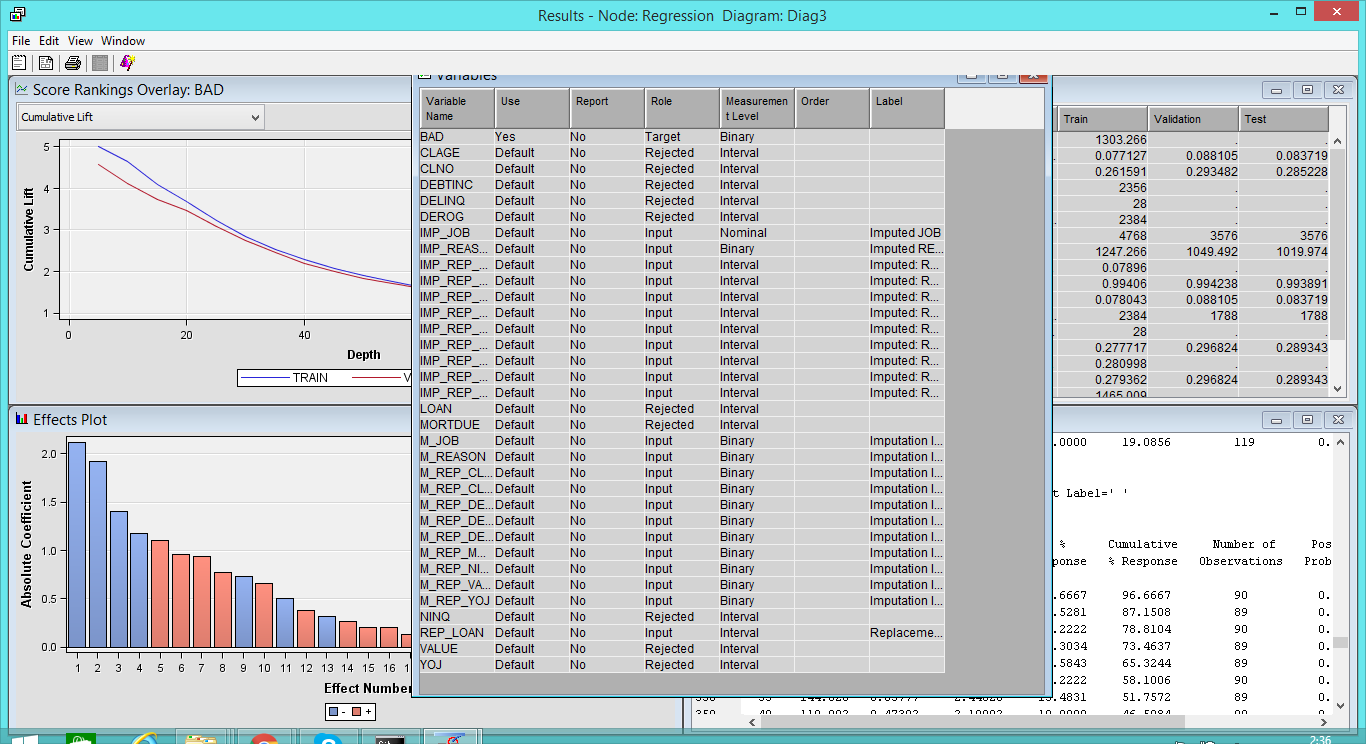
* *Как поменялась роль исходных переменных?*

**

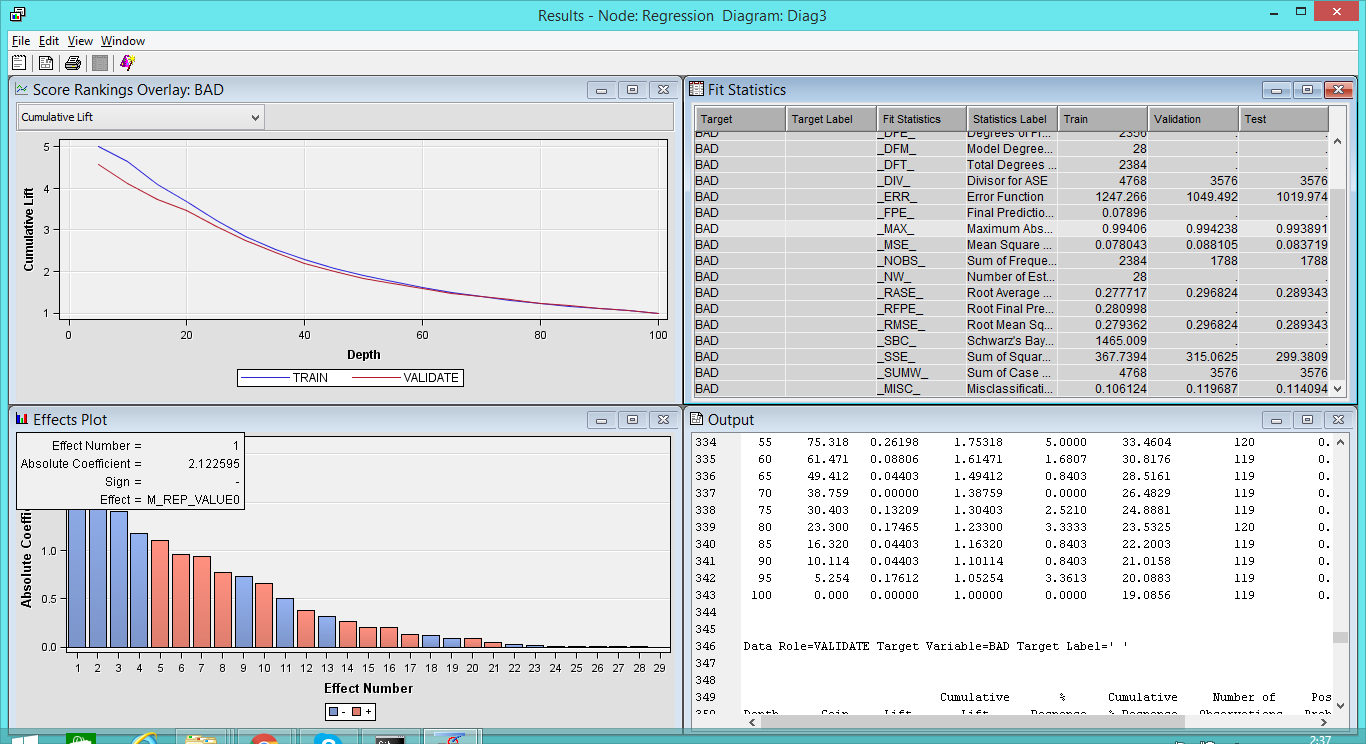
* *Содержат ли пропуски переменные с префиксом IMP\_?* Нет
* *Чему равно значение переменной M\_IMP\_JOB, если значение исходной переменной IMP\_JOB было “Office”?* M\_IMP\_JOB=0

***Построение регрессионной модели***

* *Какие переменные вошли в модель?*

**

* *У какой переменной самый большой коэффициент?*

**

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

*Установите метод выбора значимых переменных в разделе Model Selection->Selection*

*model:*

*Вариант I: Forward*

***Вариант II: Backward***

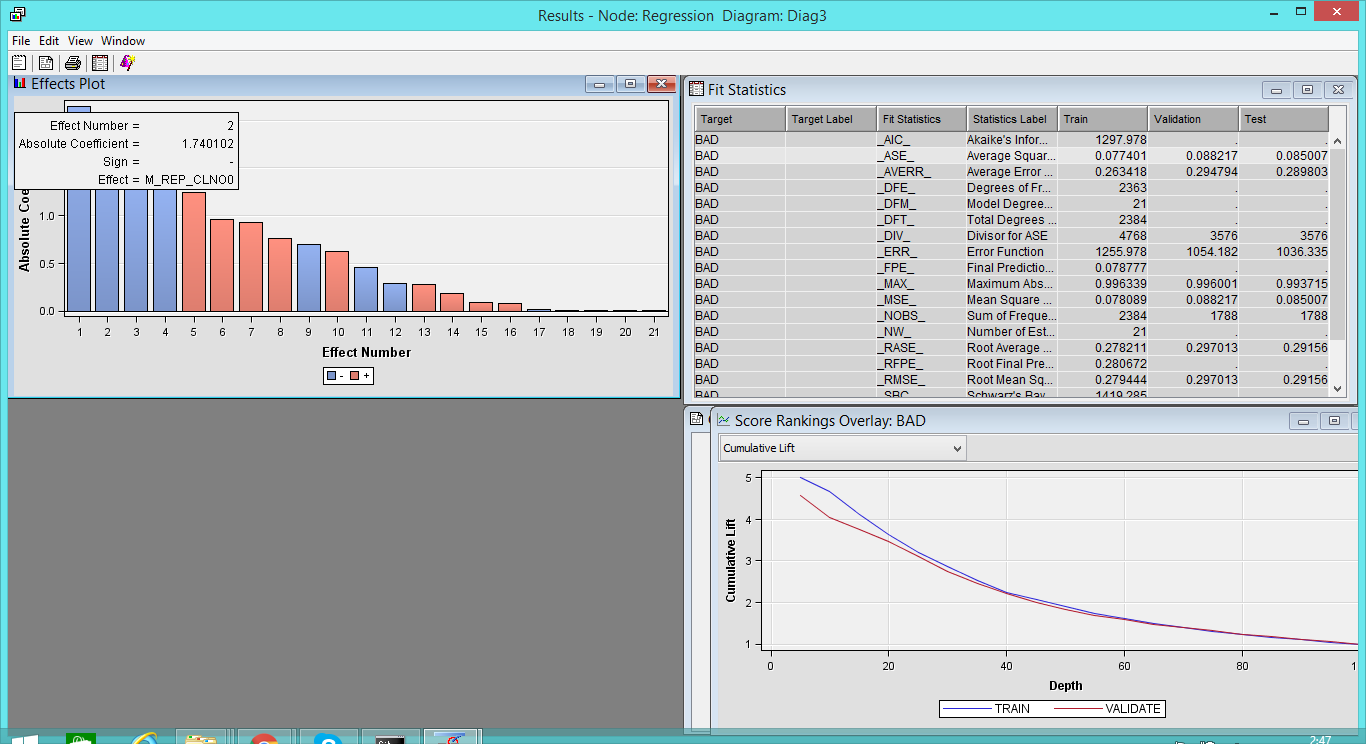
* *Какие переменные вошли в модель? В каком порядке? (для этого посмотрите либо*

*журнал работы компоненты, либо график View->model->Estimate Selection plot)*

*IMP\_JOB IMP\_REASON IMP\_REP\_CLAGE IMP\_REP\_CLNO IMP\_REP\_DEBTINC IMP\_REP\_DELINQ IMP\_REP\_DEROG IMP\_REP\_MORTDUE IMP\_REP\_NINQ IMP\_REP\_VALUE IMP\_REP\_YOJ M\_JOB M\_REASON M\_REP\_CLAGE M\_REP\_CLNO M\_REP\_DEBTINC M\_REP\_DELINQ M\_REP\_DEROG*

*M\_REP\_MORTDUE M\_REP\_NINQ M\_REP\_VALUE M\_REP\_YOJ REP\_LOAN*

* *У какой переменной самый большой коэффициент?* M\_Rep\_Value = 2.14

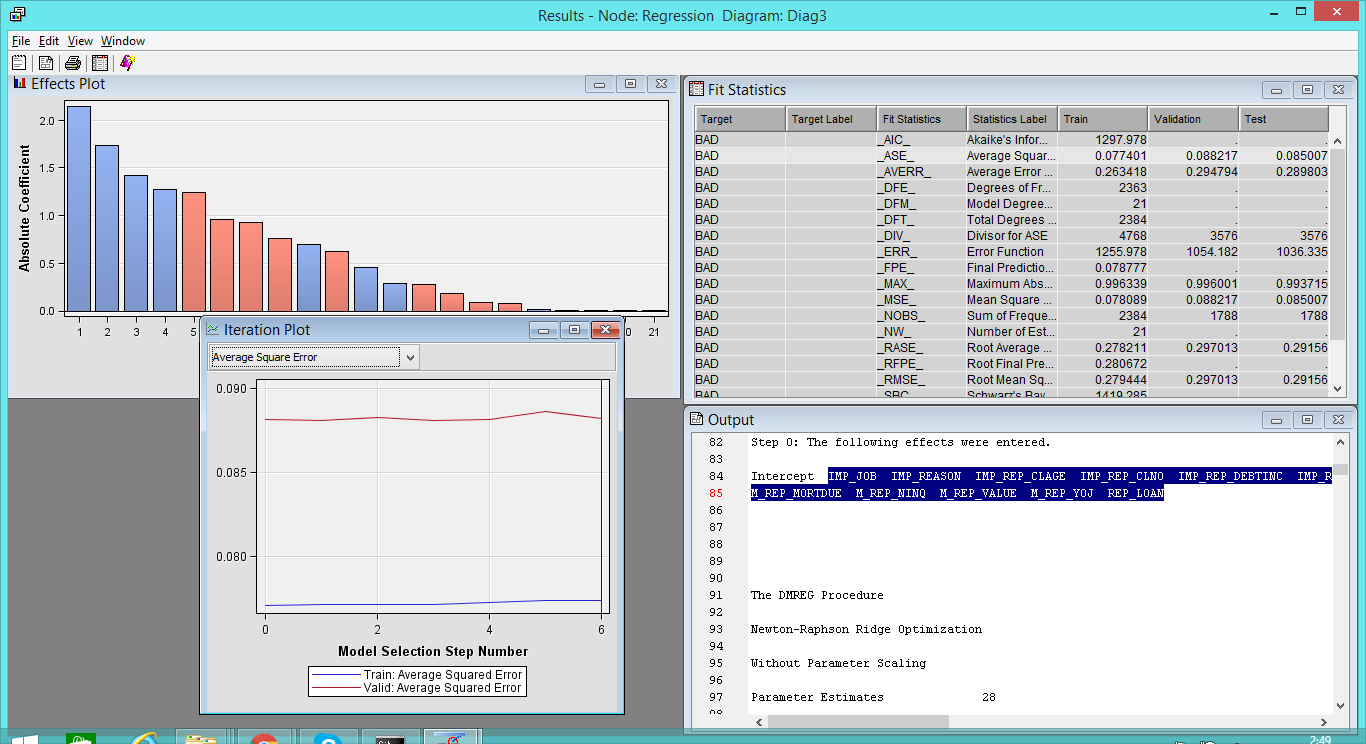
**

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и*

*валидационном наборах?*

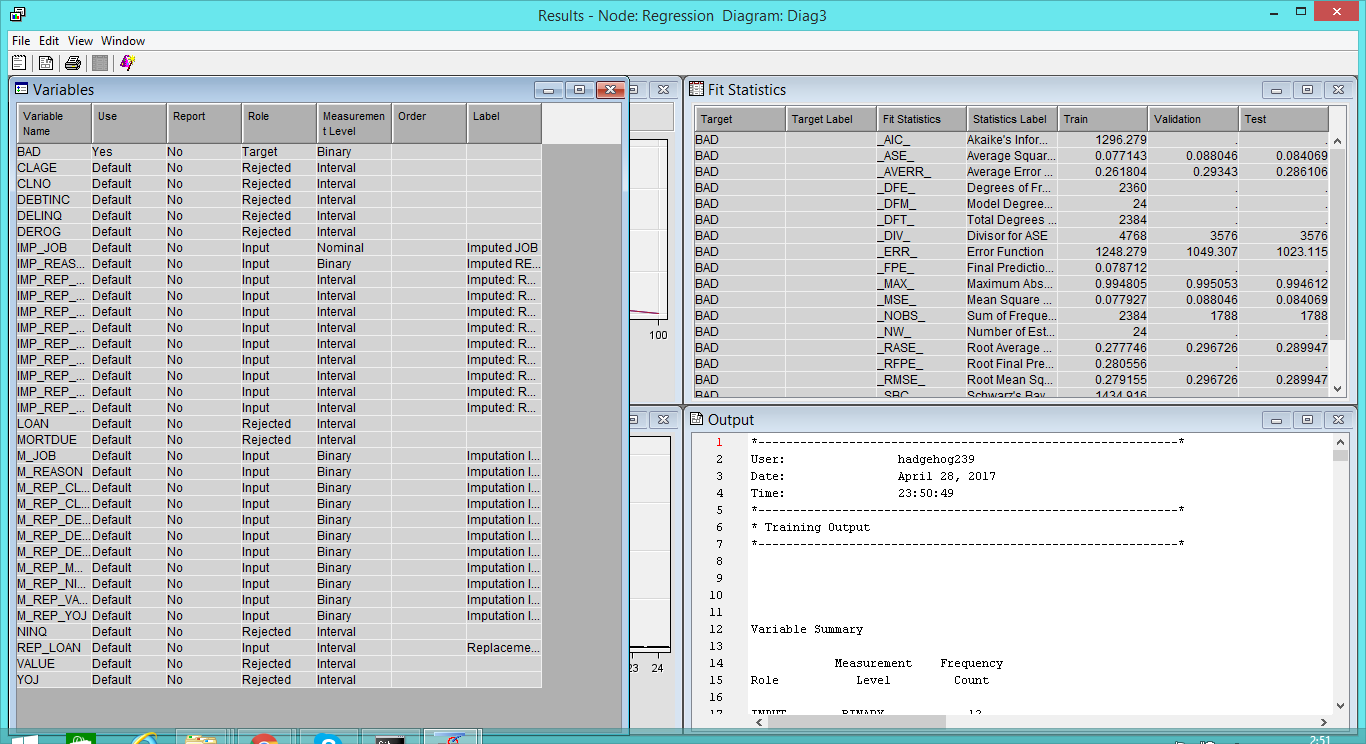
* *Использовался ли при выборе переменных валидационный набор?* Да
* *Использовался ли при выборе переменных тестовый набор?* Нет
* *Посмотрите на график view->model->iteration plot, выбрав в критериях “Error function”. Присутствует ли факт переобучения модели для вашего варианта?*

Да



* *Установите настройку в разделе Model Selection->Selection criteria: Validation Error.*

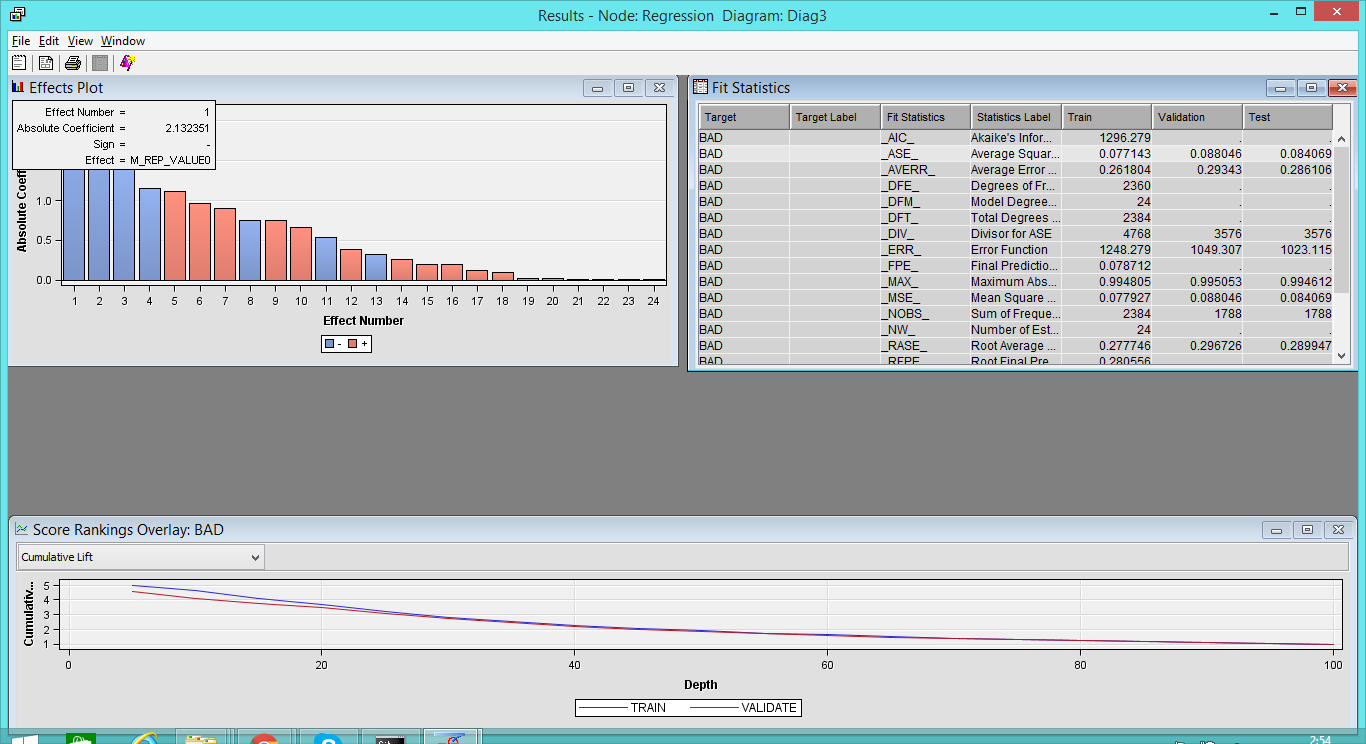
*Какие переменные вошли в модель? В каком порядке? (для этого посмотрите либо журнал работы компоненты, либо график View->model->Estimate Selection plot)*

**

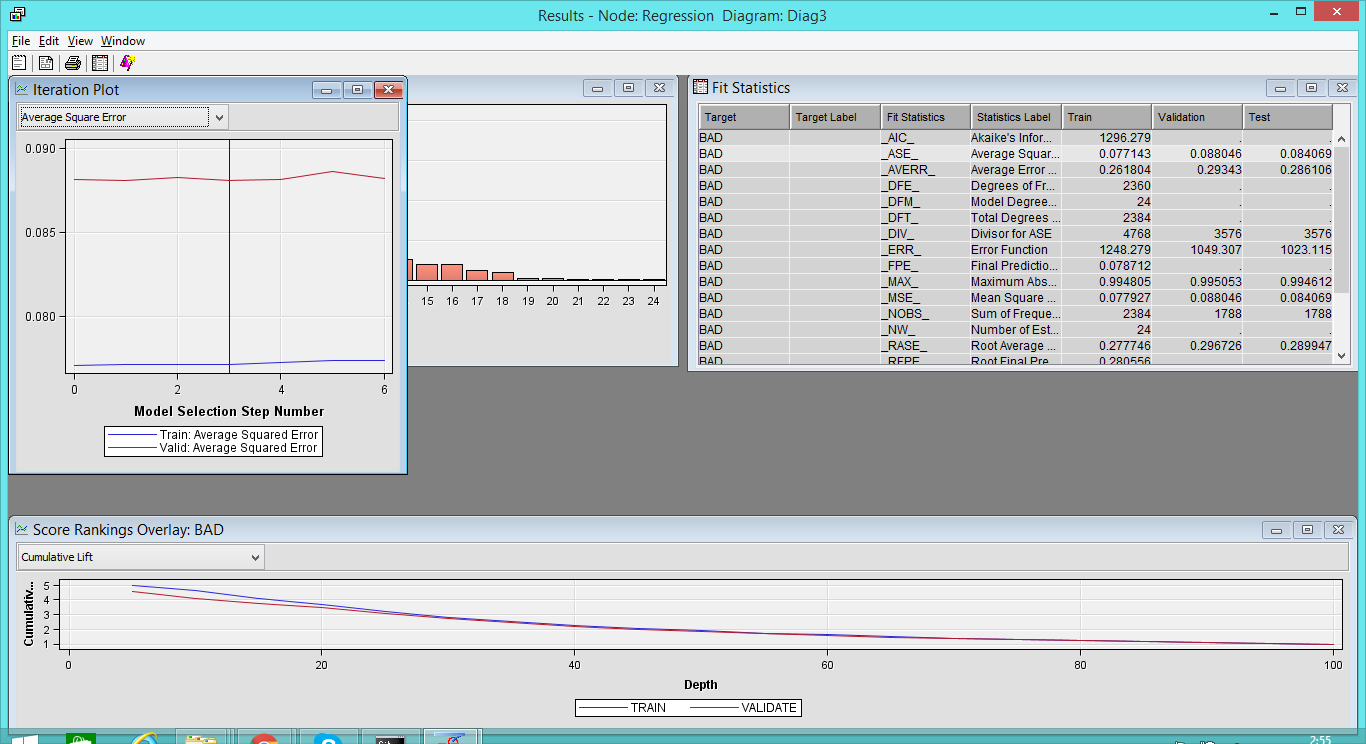
IMP\_JOB IMP\_REASON IMP\_REP\_CLAGE IMP\_REP\_CLNO IMP\_REP\_DEBTINC IMP\_REP\_DELINQ IMP\_REP\_DEROG IMP\_REP\_MORTDUE IMP\_REP\_NINQ IMP\_REP\_VALUE IMP\_REP\_YOJ M\_JOB M\_REASON M\_REP\_CLAGE M\_REP\_CLNO M\_REP\_DEBTINC M\_REP\_DELINQ M\_REP\_DEROG

M\_REP\_MORTDUE M\_REP\_NINQ M\_REP\_VALUE M\_REP\_YOJ REP\_LOAN

* *У какой переменной самый большой коэффициент?* M\_Rep\_Vaue = 2.13

**

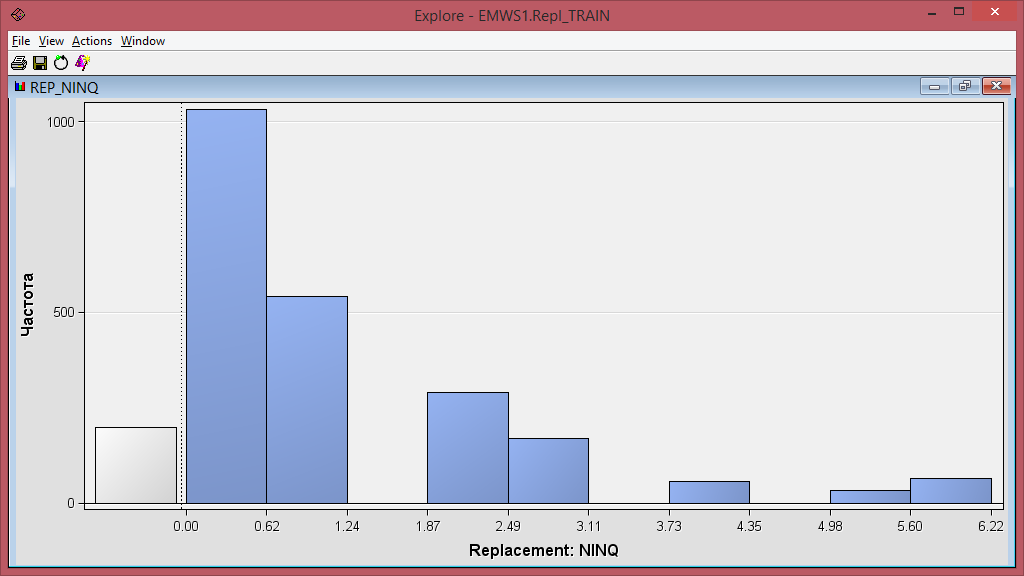
* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*
* *Использовался ли при выборе переменных валидационный набор?* Да
* *Использовался ли при выборе переменных тестовый набор?* Нет
* *Посмотрите на график view->model->iteration plot, выбрав в критериях “Error function”. Обратите внимание вертикальную синюю черту. На каком шаге была выбрана финальная модель?* На 3-ем шаге



***Преобразование переменных***

*Добавьте узел “Transform variables” между узлами “Replacement” и “Impute”. Постройте гистограмму распределения переменной REP\_NINQ (число заявок на кредит). Для этого выберите на узле “Transform variables” Edit Variables, выделите переменную REP\_NINQ и нажмите кнопку Explore.*

* *Что можно сказать про асимметричность распределения?*

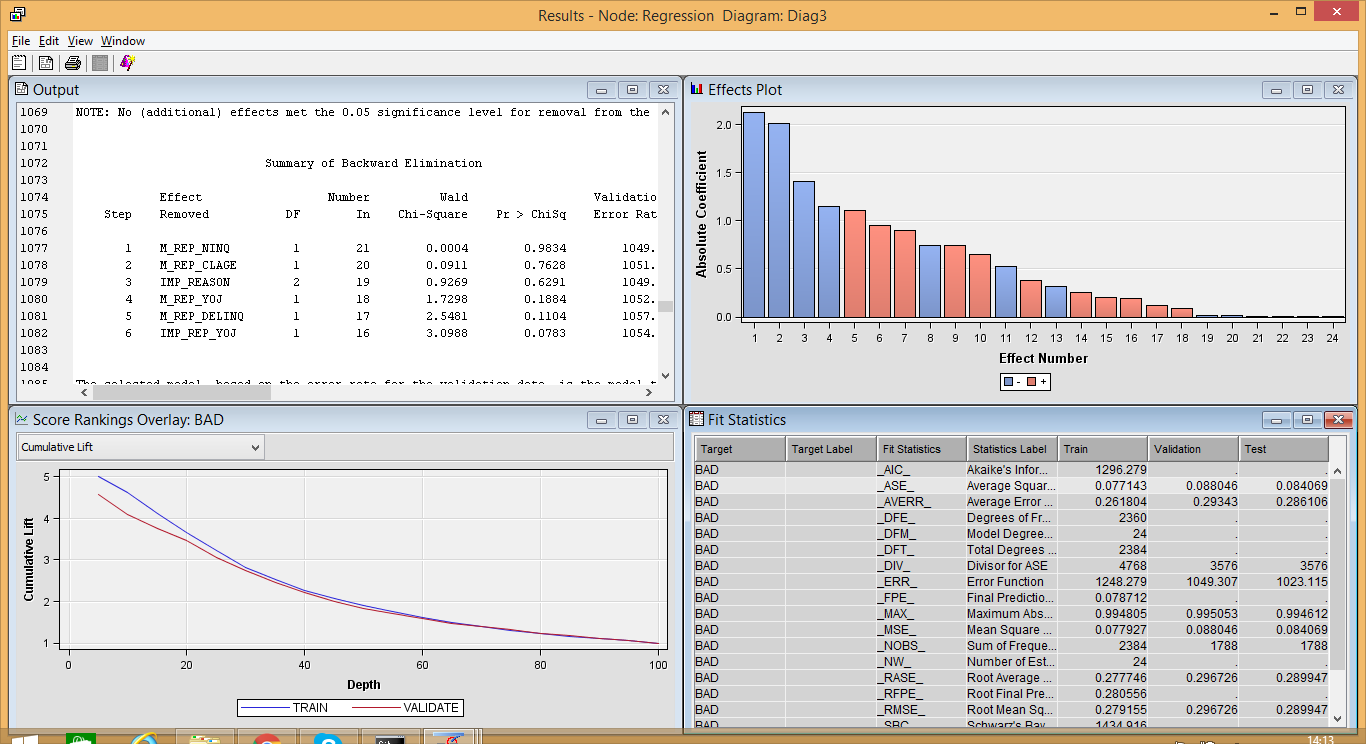
**

Присутствует левосторонняя асимметрия. Положительное значение коэффициента асимметрии указывает, что размер правого «хвоста» распределения больше, чем левого (относительно среднего).

***Вариант II:*** *Quintile (разбиение на интервалы с одинаковым числом наблюдений в каждом).*

*Обучите регрессионную модель.*

* *Вошла ли измененная переменная в модель?* Да.
* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

******

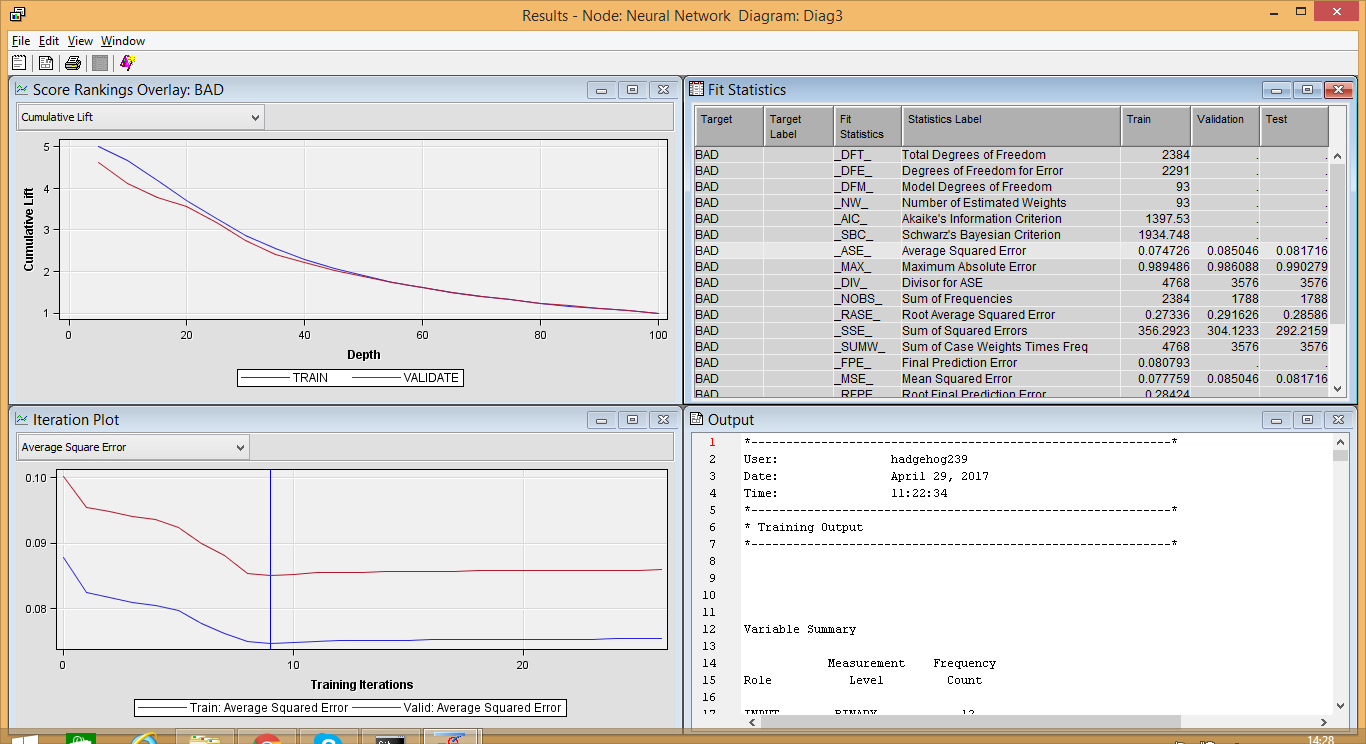
***Подключение нейронной сети***

*Подключите узел «Neural Networks» после «Impute», выберете Model Selection Criteria ->*

*Average Error, выберете архитектуру Network->architecture*

*Вариант II: “Normalized Radial Basis Network – unequal width and height“, обучите модель.*

* *Сколько степеней свободы получилось в сети (см. Fit Statistics)? 93 Почему?*

******

Количество степеней свободы показывает размерность вектора из случайных величин, количество «свободных» величин, необходимых для того, чтобы полностью определить вектор.

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*
* *Обратите внимание на график Iteration plot: на каком шаге обучения выбрана оптимальная модель?*

******

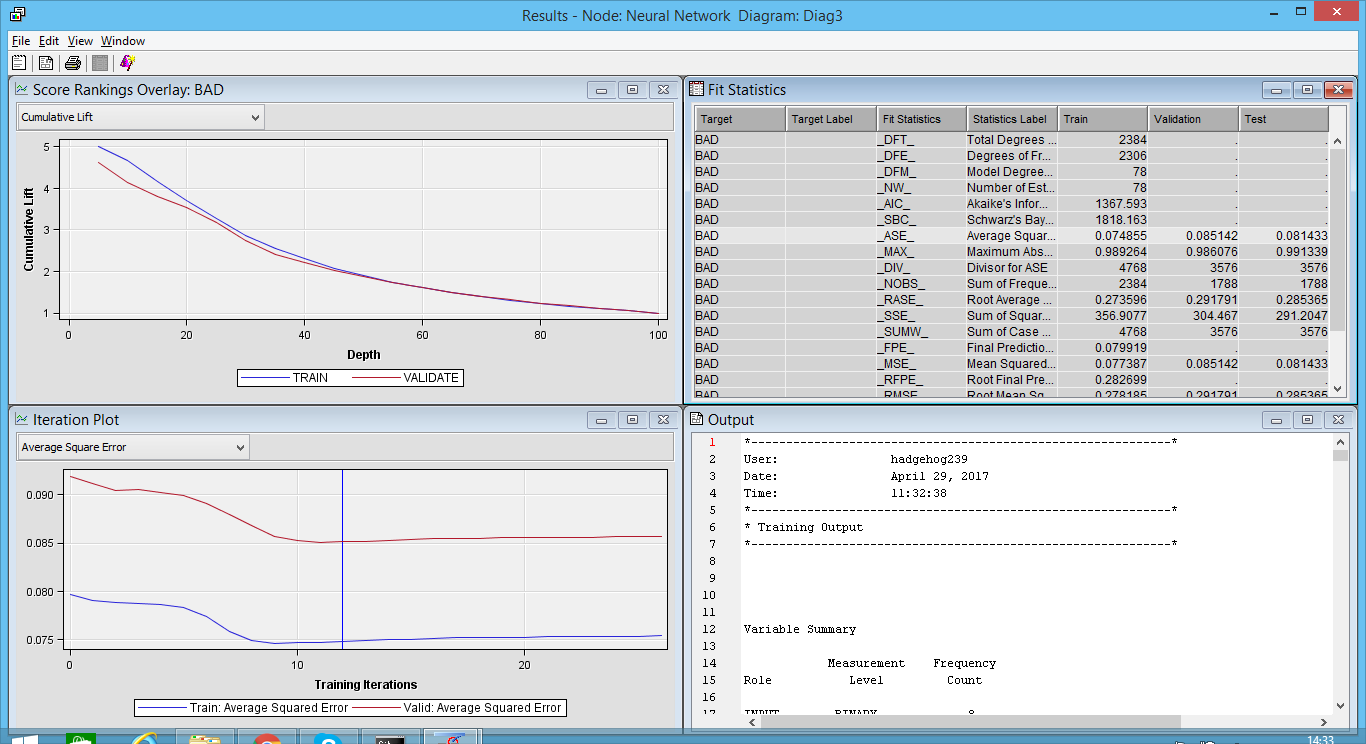
***Выбор значимых переменных***

* *Переместите ваш узел «Neural Networks» после узла “Regression”*
* *Сколько степеней свободы получилось в сети (см. Fit Statistics)? 78 Почему?*

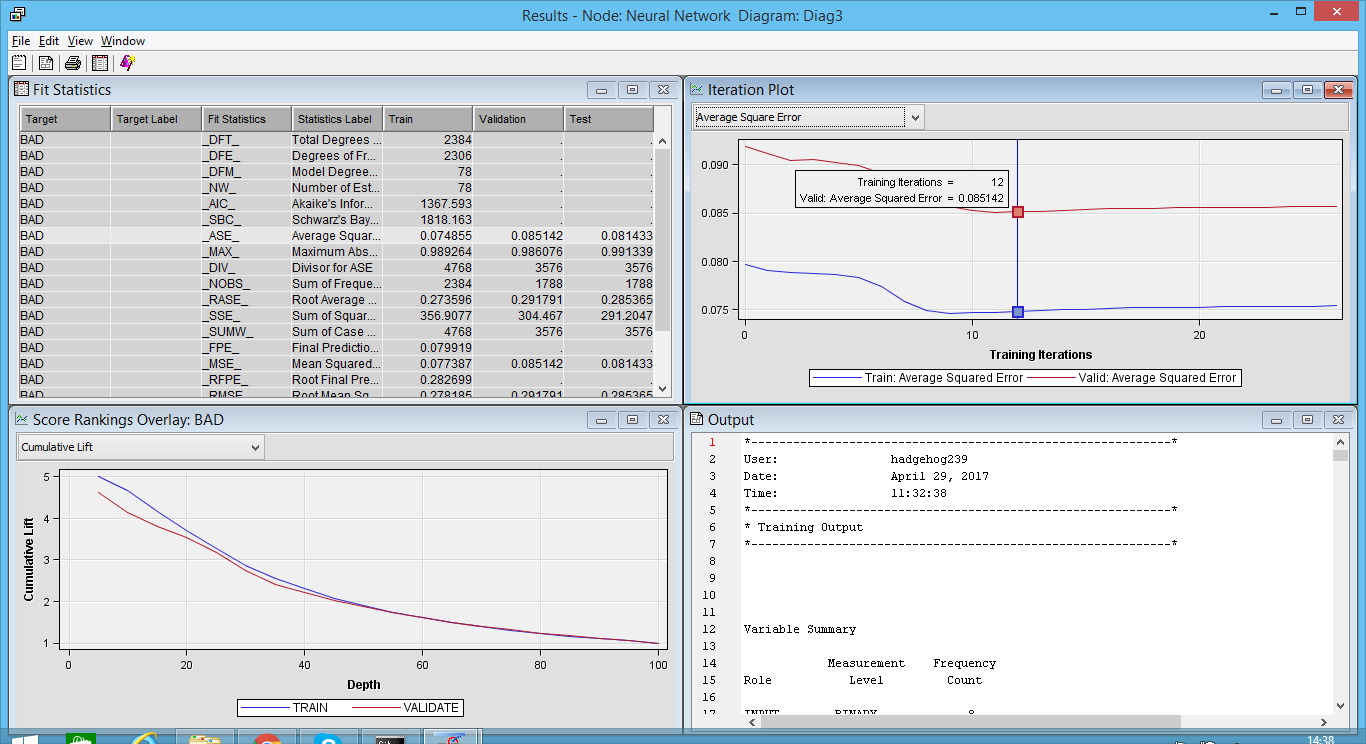
Число степеней свободы - это минимально необходимое число значений зависимой переменной, которых достаточно для получения искомой характеристики выборки и которые могут свободно варьироваться с учетом того, что для этой выборки известны все другие величины, используемые для расчета искомой характеристики.

Для получения остаточной дисперсии необходимы коэффициенты уравнения регрессии. В случае парной линейной регрессии коэффициентов два, поэтому в соответствии с формулой (принимая ) число степеней свободы равно .

Имеется в виду, что для определения остаточной дисперсии достаточно знать коэффициенты уравнения регрессии и только значений зависимой переменной из выборки. Оставшиеся два значения могут быть вычислены на основании этих данных, а значит, не являются свободно варьируемыми.

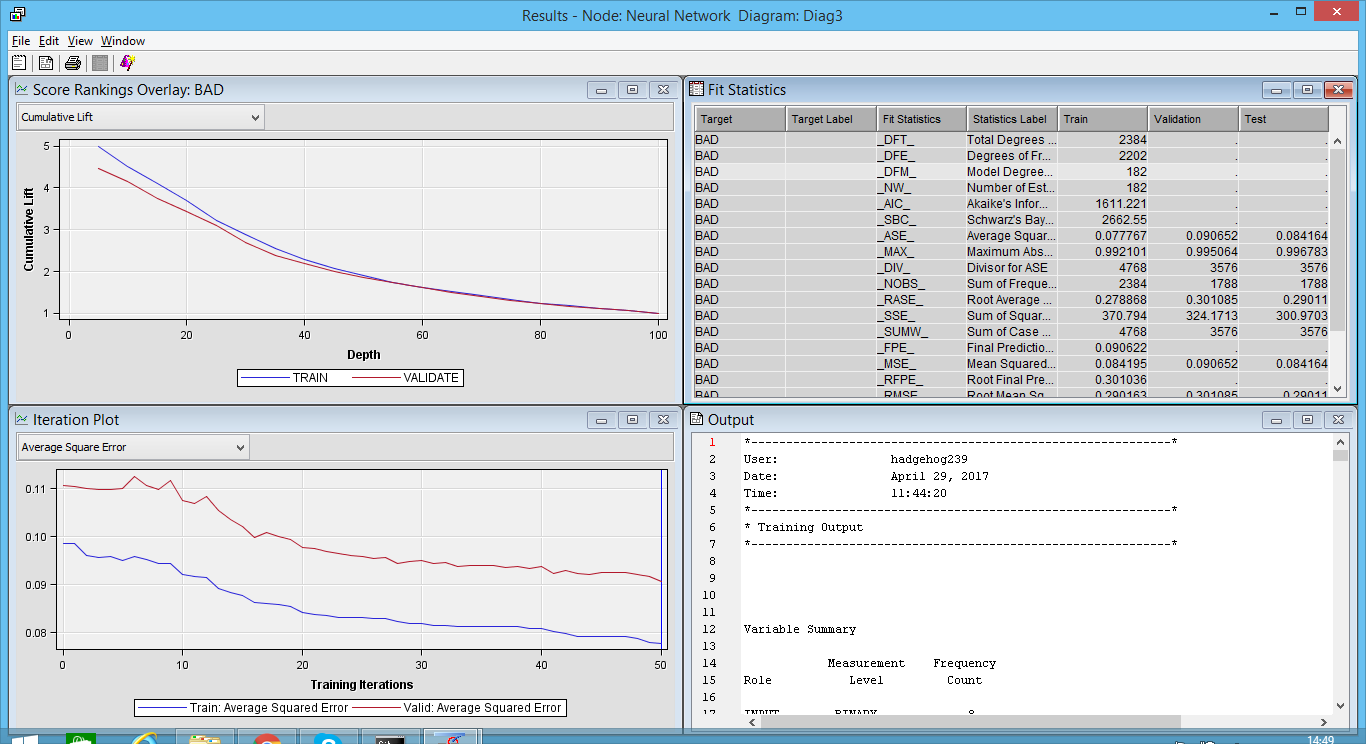
**

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*
* *Обратите внимание на график Iteration plot: на каком шаге обучения выбрана оптимальная модель?*

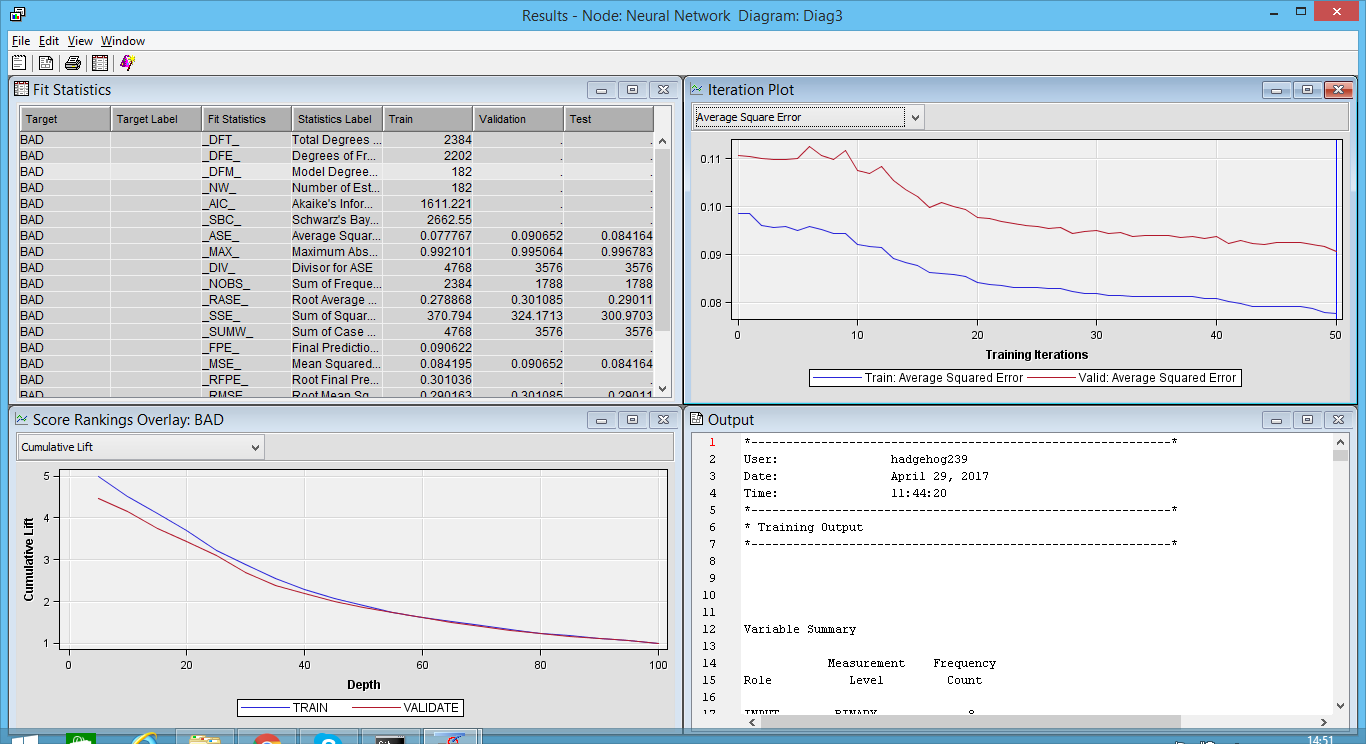
**

***Усложнение архитектуры***

* *Сколько степеней свободы получилось в сети (см. Fit Statistics)? 182 Почему?*

******

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*
* *Обратите внимание на график Iteration plot: на каком шаге обучения выбрана оптимальная модель?* 50

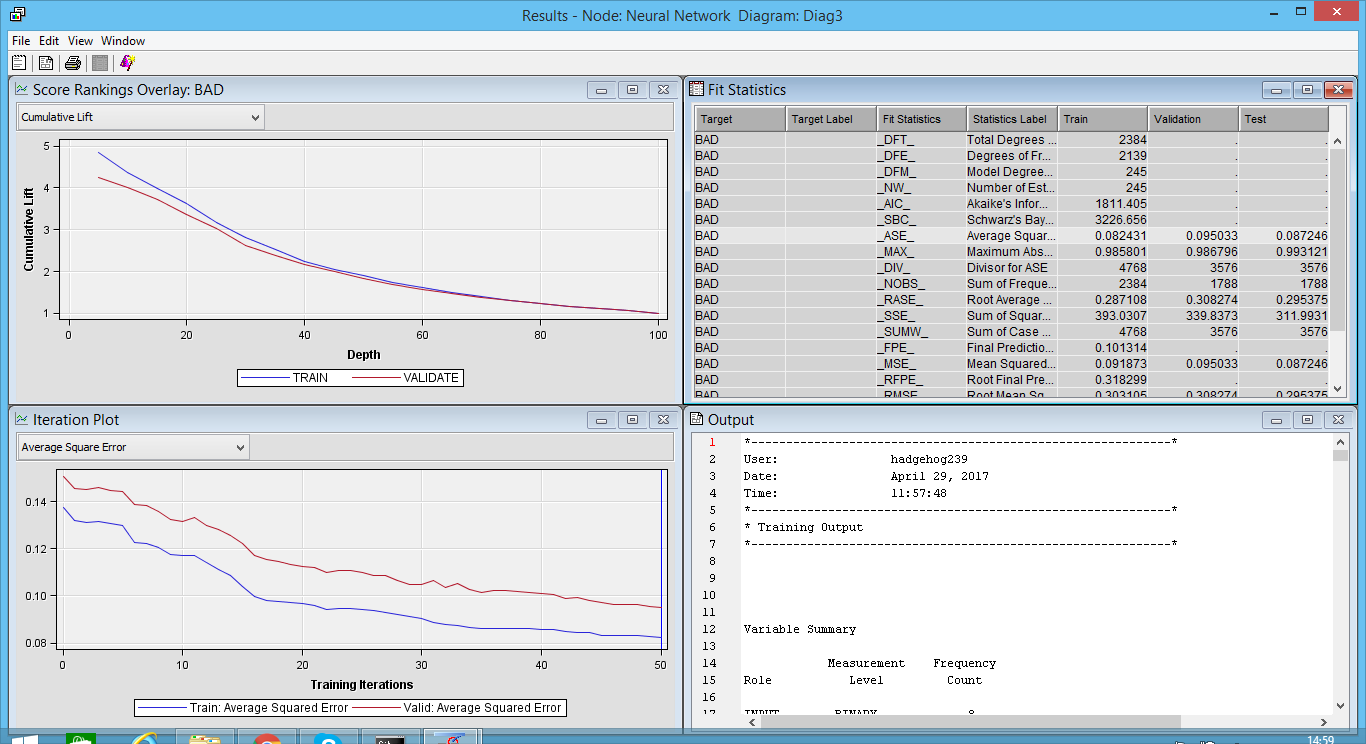
**

***Нелинейное преобразование входных признаков***

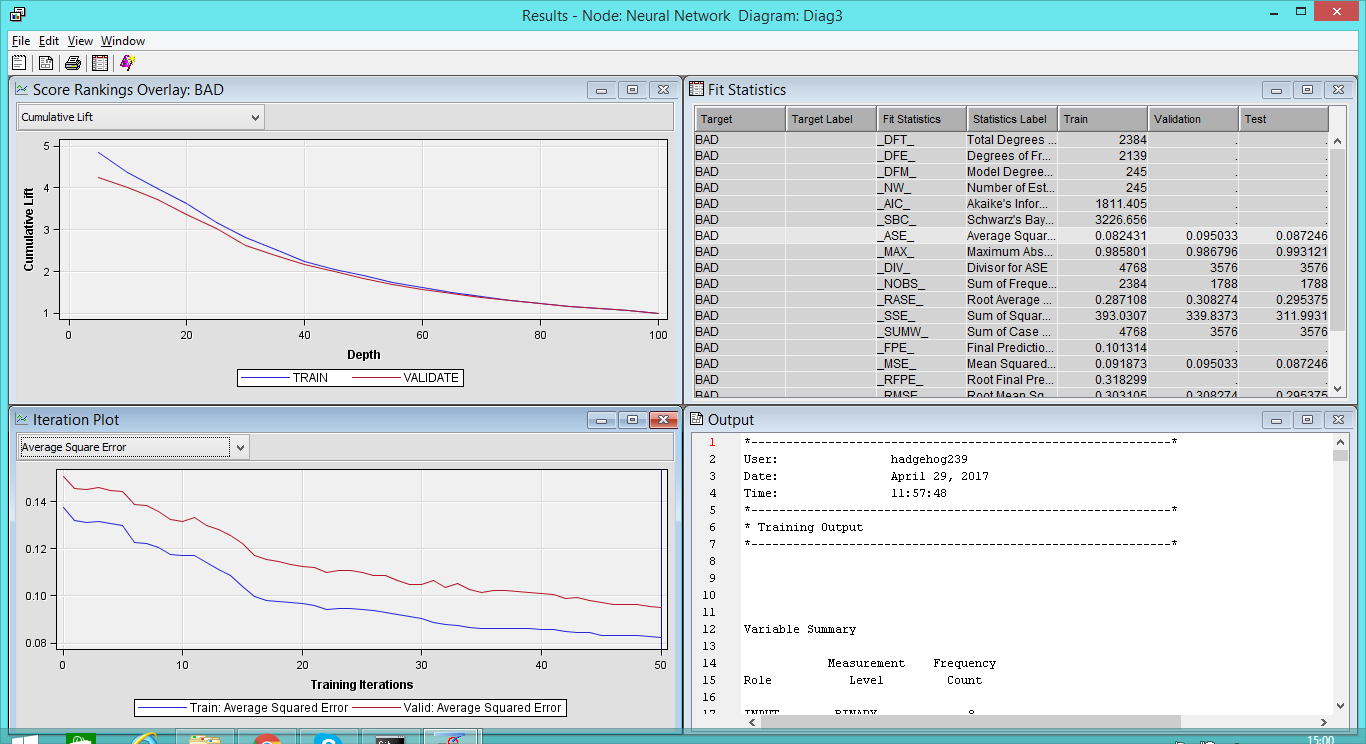
*Подключите узел «SOM» после «Impute», и подключите вашу нейронную сеть после узла*

*SOM. Обучите модель.*

* *Сколько степеней свободы получилось в сети (см. Fit Statistics)? Почему?*
* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

**

* *Обратите внимание на график Iteration plot: на каком шаге обучения выбрана оптимальная модель?* 50

**

***Максимальное дерево решений***

*Подключите узел «Decision tree» после узла «data partition», выберете в качестве nominal target criteria*

***Вариант II: gini***

*В разделе Subtree->Method выберете Largest, в разделе и обучите модель.*

* *Почему в отличии от регрессионных и нейросетевых моделей дерево решений можно подключить сразу после «data partition»?*

Потому что дерево решений используется как вспомогательный инструмент например для замены пропущенных значений, что нельзя сказать о регрессии или о нейронной сети.

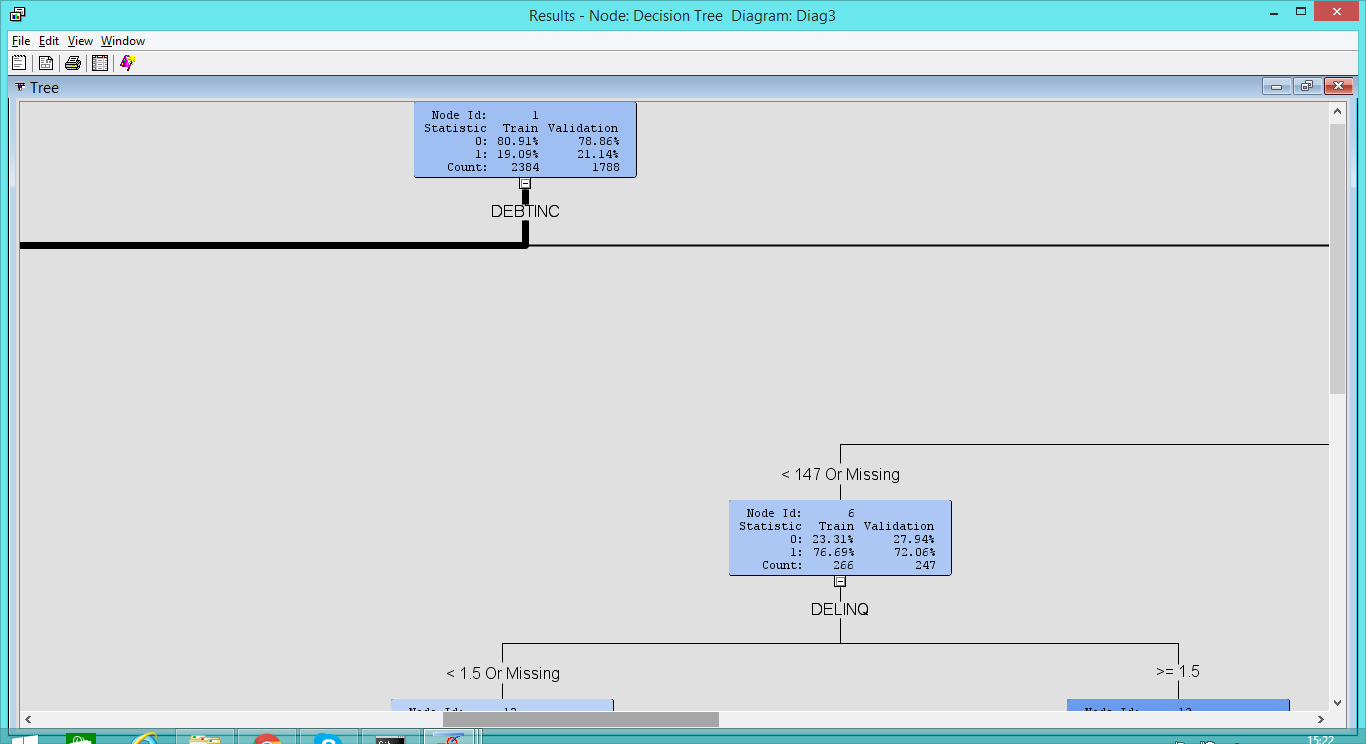
* *Сколько листьев получилось в дереве?*

32

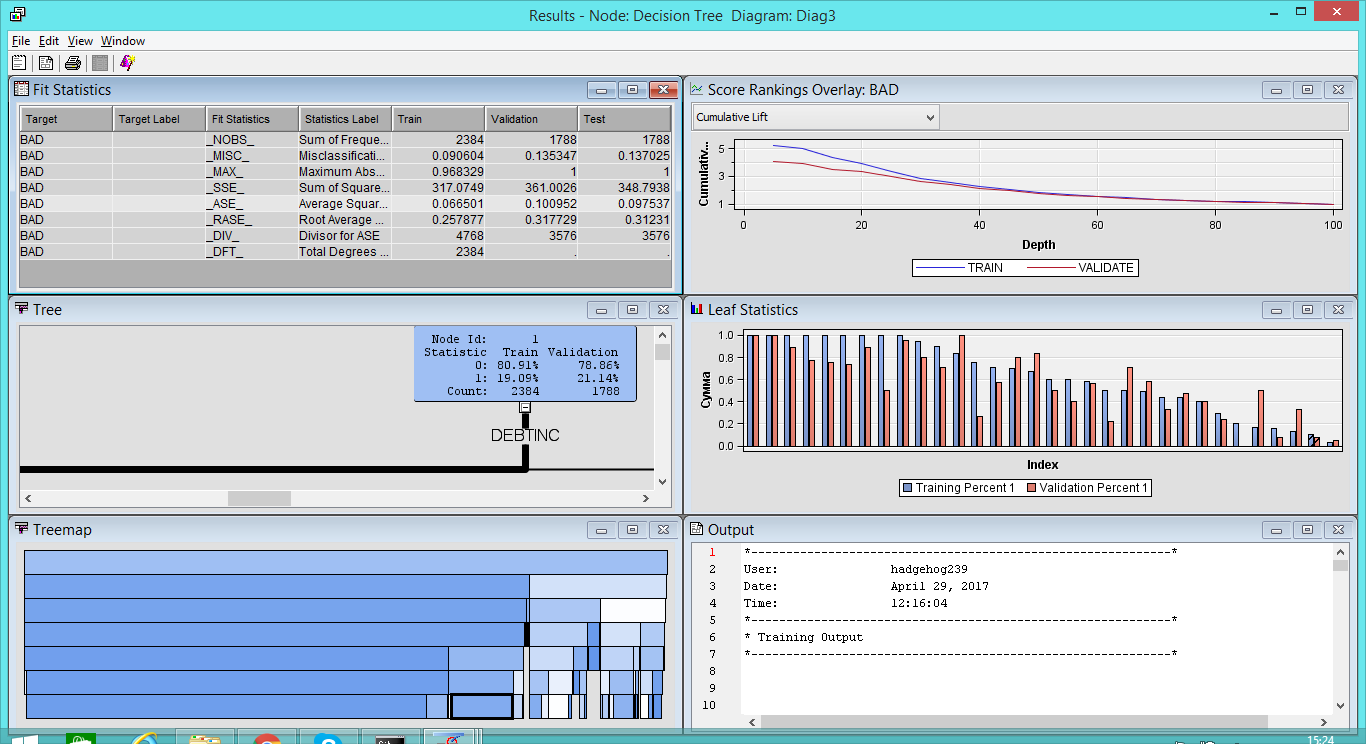
* *Какие переменные вошли в модель?*

DEBTINC DELINQ VALUE CLAGE DEROG MORTDUE LOAN JOB CLNO YOJ NINQ

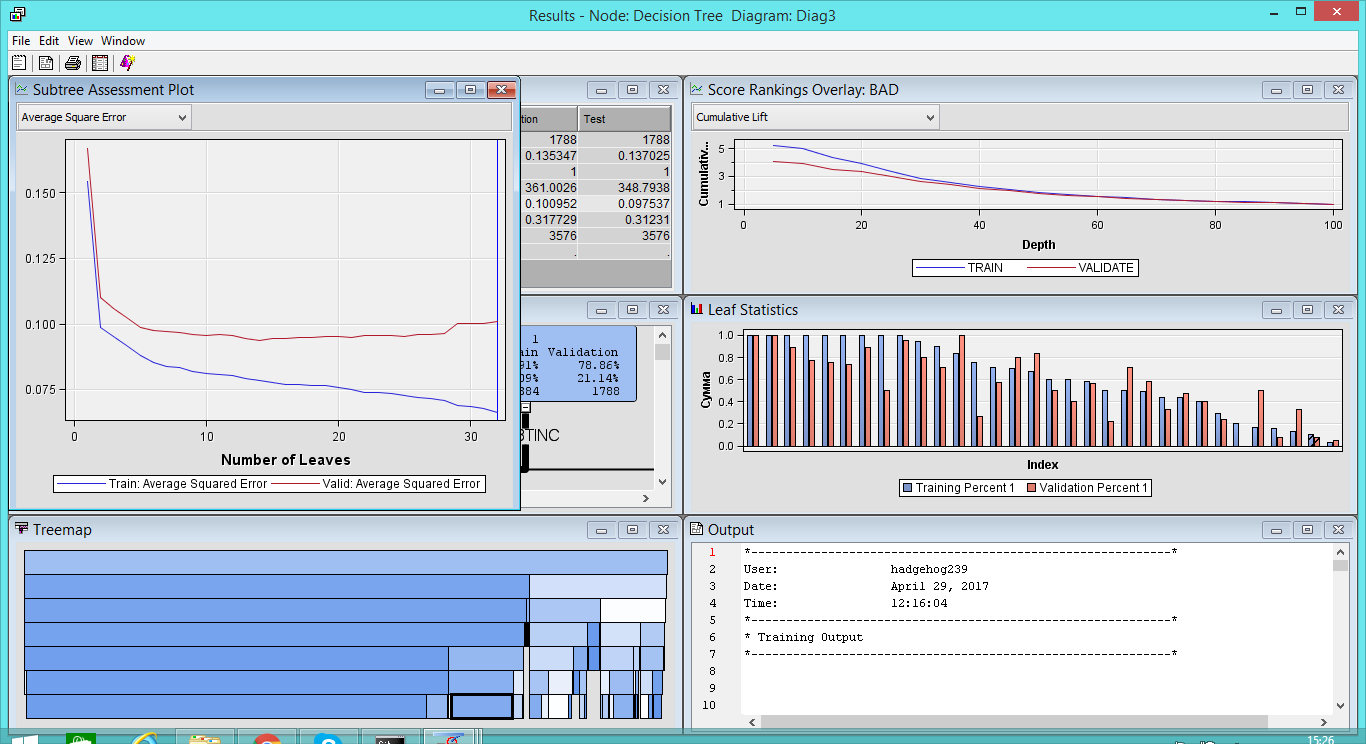
* *По какой переменной произошло первое разбиение?* DEBTINC

**

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

**

* *Посмотрите на график View->Model->subtree assessment plot. Как думаете какой размер дерева (при критерии ASE) должен быть выбран?* 9 - 10

******

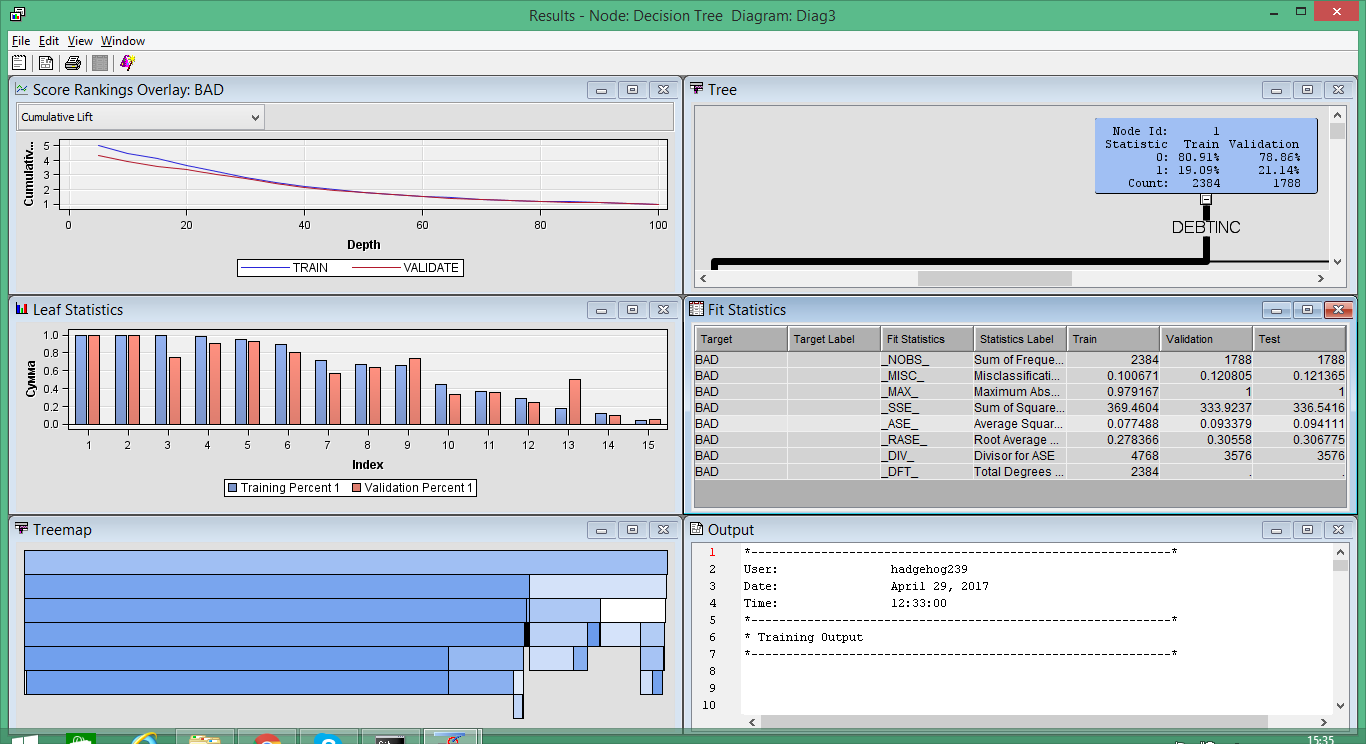
***Обрубание дерева решений***

*В разделе Subtree->Method выберете Assessment, Assessment measure выберете Average Square Error и обучите модель.*

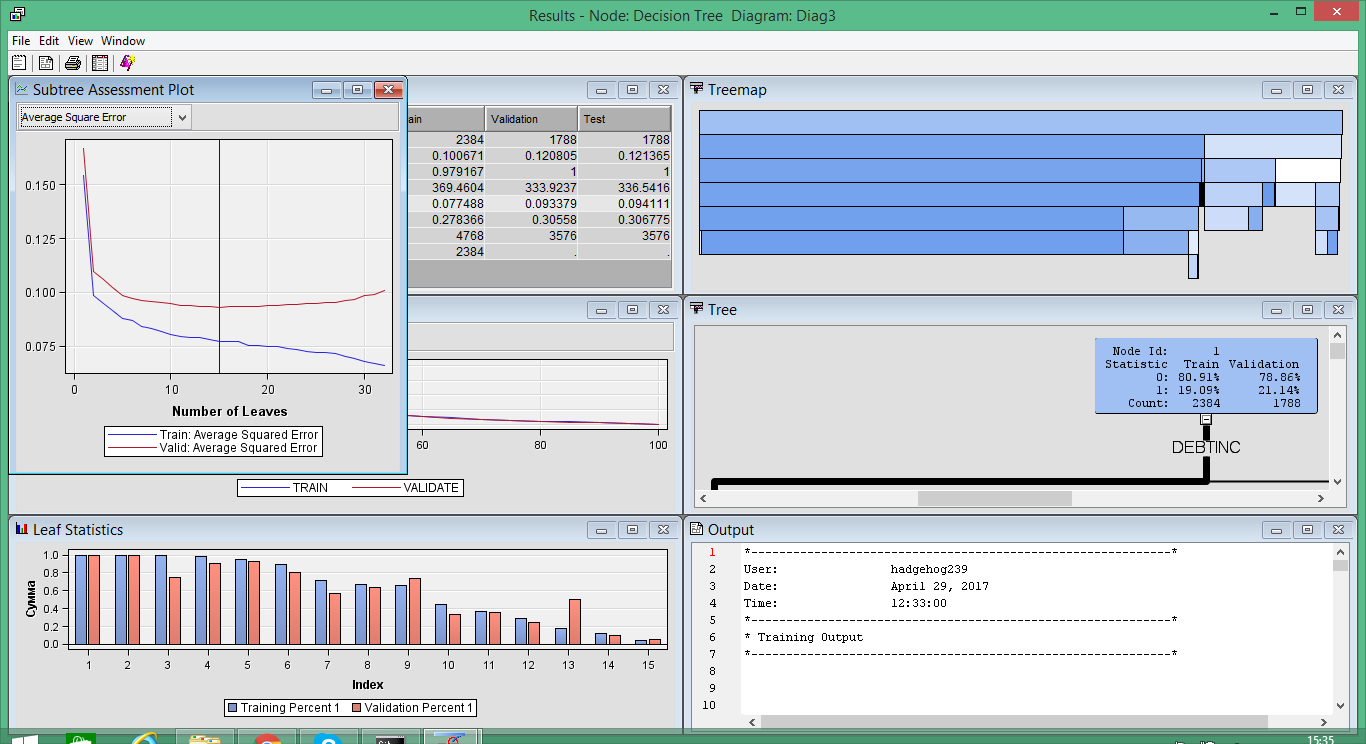
* *Какие переменные вошли в модель?*

DEBTINC DELINQ CLAGE VALUE DEROG CLNO LOAN

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

**

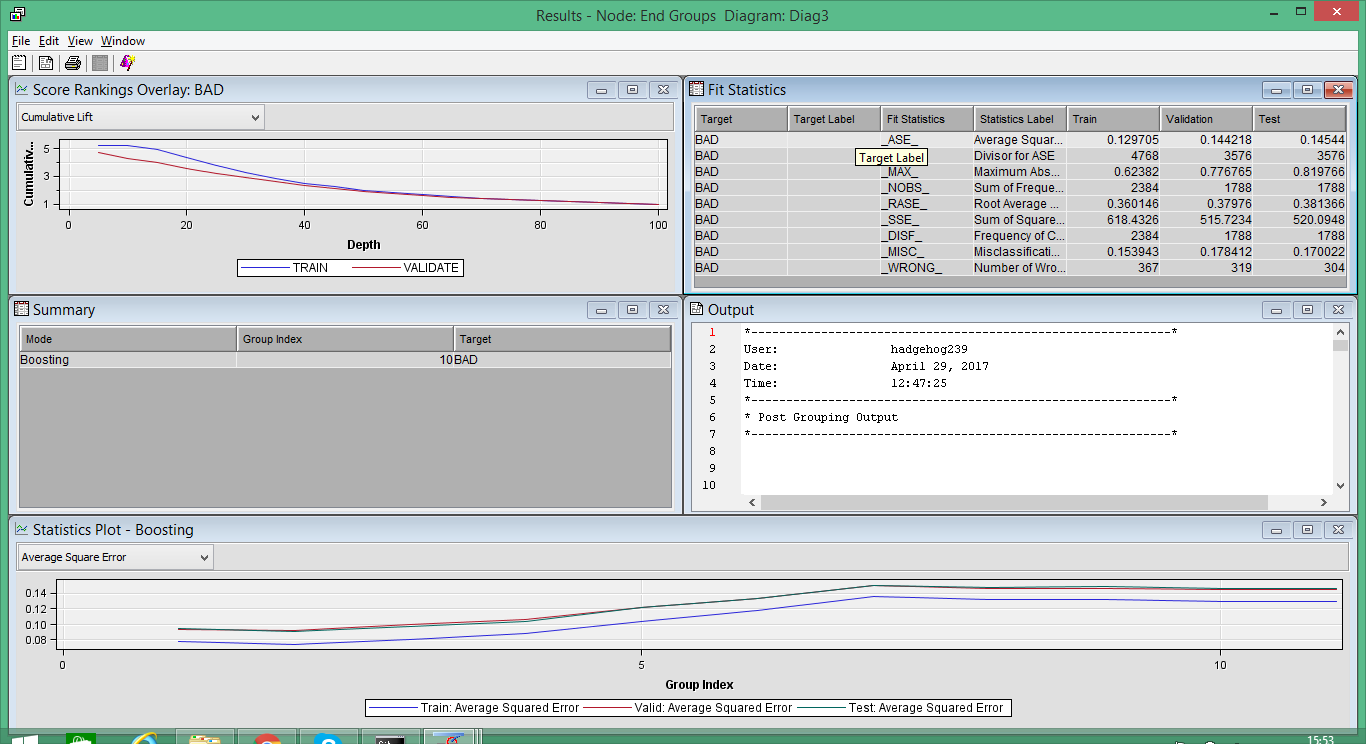
* *Посмотрите на график View->Model->subtree assessment plot. Какой размер дерева (при критерии ASE) был выбран как оптимальный?* 15

**

***Ансамбль деревьев***

*Добавьте узел “start group” перед узлом «Decision tree» и узел “end group” после него. В настройках start group выберете Mode=boosting. Обучите группу.*

* *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

******

***Сравнение моделей***

* *Добавьте узел “Model comparison” и соедините его вход со всеми вашими регрессионными моделями, нейронной сетью и ансамблем деревьев. Выберете в качестве Selection statistics ROC а в качестве Selection table – Test. Обучите Модель.*

*DATA Test*

* *Какие знчения ROC индекса на тестовом наборе получились у всех моделей? Какая модель победила?*

*Reg Tree Neural*

0.87 0.69 0.56

Регрессионная.

* *Подключите после узла “Model comparison” узел Cutoff, в Cutoff Method выберете Maximum Kolmogorov-Smirnov statistics. Запустите узел.*

*Какое значение порога отсечения оптимально для победившей модели?* 0,46

***Подключение метода опорных векторов***

Подключите из раздела HPDM узел «HPSVM» после «Impute» и соедините его выход с

узлом Model Comparison, выберете Optimization method:

**Вариант II: “Active set“,** обучите модель:

* *Сколько опорных векторов получилось в модели (см. Fit Statistics)? Сколько из них лежит на «зазоре»?*

Number of Support Vectors 568

Number of Support Vectors on Margin 112

* Каково значение среднеквадратичной ошибки ASE и площади под ROC кривой на тренировочном, тестовом и валидационном наборах?

Train Validation Test

ASE 0.16 0.175 0.176

Reg 0,88 0,88 0,87

HPSVM 0,98 0,85 0,84

Tree 0,78 0,69 0,69

Neural 0,52 0,52 0,56

*Поменяйте в разделе “Interior point Options” (или Active set Options”) тип модели с линейной на полиномиальную второго порядка:*

* Сколько опорных векторов получилось в модели (см. Fit Statistics)? Сколько из них лежит на «зазоре»? Почему число векторов изменилось по сравнению с линейно моделью?

Number of Support Vectors 689

Number of Support Vectors on Margin 105

По мере увеличения степени полинома растет и количество решений. Так как в нашем случае степень двойки, очевидно, что векторов будет больше.

* Каково значение среднеквадратичной ошибки ASE и площади под ROC кривой на тренировочном, тестовом и валидационном наборах?

Train Validation Test

ASE 0.16 0.175 0.176

Reg 0,09 0,88 0,87

HPSVM 0,16 0,85 0,84

Tree 0,16 0,69 0,69

Neural 0,15 0,52 0,56