Формирование обучающих и контрольных выборок

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Ids_DATA	5960
TRAIN	EMWS1.Part_TRAIN	2382
VALIDATE	EMWS1.Part_VALIDATE	1788
TEST	EMWS1.Part_TEST	1790

- *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?*

Output

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	4771	80.0503	
BAD	1	1	1189	19.9497	

Data=TEST

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	1433	80.0559	
BAD	1	1	357	19.9441	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	1907	80.0588	
BAD	1	1	475	19.9412	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	1431	80.0336	
BAD	1	1	357	19.9664	

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?*

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Ids_DATA	5960
TRAIN	EMWS1.Part_TRAIN	2384
VALIDATE	EMWS1.Part_VALIDATE	1788
TEST	EMWS1.Part_TEST	1788

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?*

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	4771	80.0503	
BAD	1	1	1189	19.9497	

Data=TEST

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	1432	80.0895	
BAD	1	1	356	19.9105	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	1929	80.9144	
BAD	1	1	455	19.0856	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
BAD	0	0	1410	78.8591	
BAD	1	1	378	21.1409	

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.

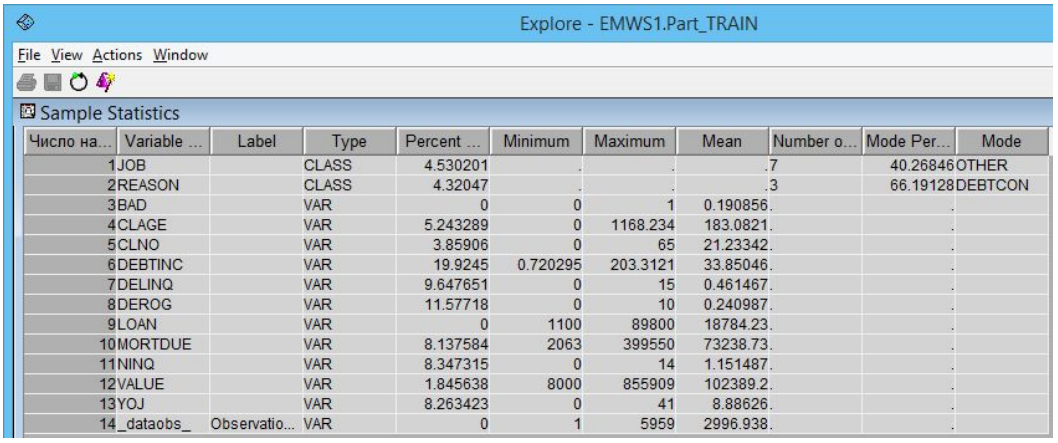
Поиск и удаление выбросов

- *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.

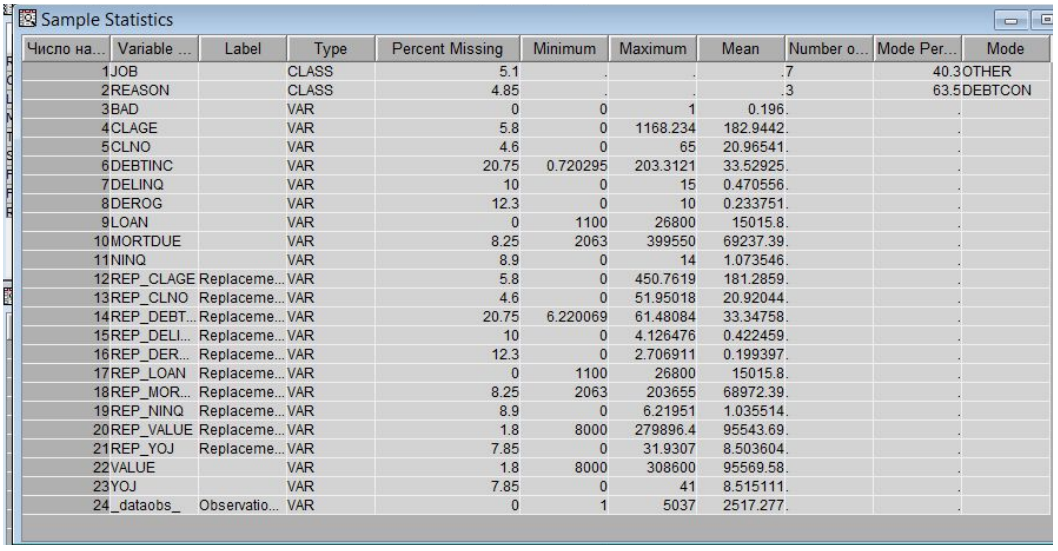
- *В списке “Imported Data” и “Exported data” для роли train посмотрите список переменных. Что изменилось? Как изменилась роль исходных переменных?*

Imported Data



Число на...	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1	JOB		CLASS	4.530201	.	.	.	7	40.26846	OTHER
2	REASON		CLASS	4.32047	.	.	.	3	66.19128	DEBTCON
3	BAD		VAR	0	0	1	0.190856	.	.	
4	CLAGE		VAR	5.243289	0	1168.234	183.0821	.	.	
5	CLNO		VAR	3.85906	0	65	21.23342	.	.	
6	DEBTINC		VAR	19.9245	0.720295	203.3121	33.85046	.	.	
7	DELINQ		VAR	9.647651	0	15	0.461467	.	.	
8	DEROG		VAR	11.57718	0	10	0.240987	.	.	
9	LOAN		VAR	0	1100	89800	18784.23	.	.	
10	MORTDUE		VAR	8.137584	2063	399550	73238.73	.	.	
11	NINQ		VAR	8.347315	0	14	1.151487	.	.	
12	VALUE		VAR	1.845638	8000	855909	102389.2	.	.	
13	YOJ		VAR	8.263423	0	41	8.88626	.	.	
14	_dataobs_	Observatio...	VAR	0	1	5959	2996.938	.	.	

Exported data



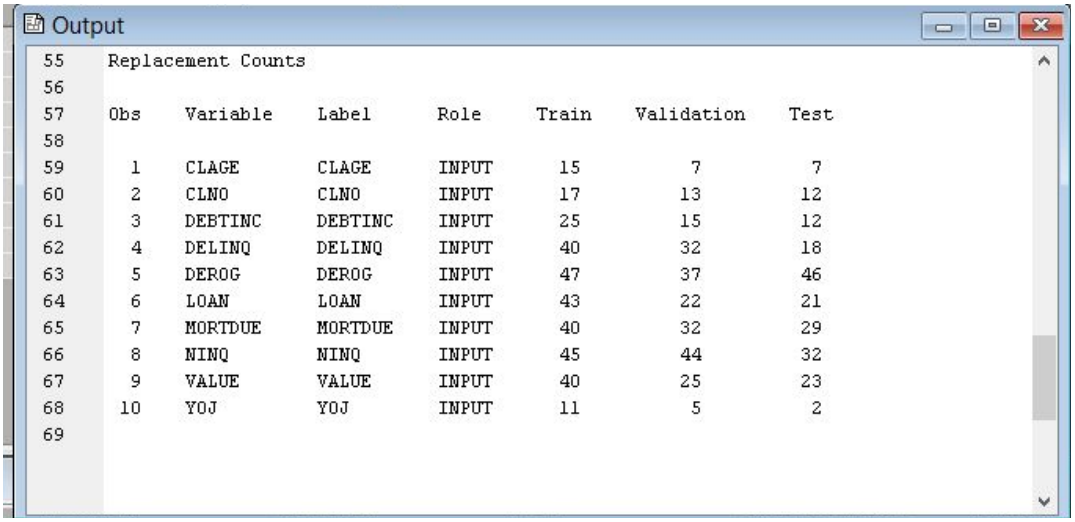
Число на...	Variable ...	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1	JOB		CLASS	5.1	.	.	.	7	40.3	OTHER
2	REASON		CLASS	4.85	.	.	.	3	63.5	DEBTCON
3	BAD		VAR	0	0	1	0.196	.	.	
4	CLAGE		VAR	5.8	0	1168.234	182.9442	.	.	
5	CLNO		VAR	4.6	0	65	20.96541	.	.	
6	DEBTINC		VAR	20.75	0.720295	203.3121	33.52925	.	.	
7	DELINQ		VAR	10	0	15	0.470556	.	.	
8	DEROG		VAR	12.3	0	10	0.233751	.	.	
9	LOAN		VAR	0	1100	26800	15015.8	.	.	
10	MORTDUE		VAR	8.25	2063	399550	69237.39	.	.	
11	NINQ		VAR	8.9	0	14	1.073546	.	.	
12	REP_CLAGE	Replaceme...	VAR	5.8	0	450.7619	181.2859	.	.	
13	REP_CLNO	Replaceme...	VAR	4.6	0	51.95018	20.92044	.	.	
14	REP_DEBT...	Replaceme...	VAR	20.75	6.220069	61.48084	33.34758	.	.	
15	REP_DELI...	Replaceme...	VAR	10	0	4.126476	0.422459	.	.	
16	REP_DER...	Replaceme...	VAR	12.3	0	2.706911	0.199397	.	.	
17	REP_LOAN	Replaceme...	VAR	0	1100	26800	15015.8	.	.	
18	REP_MOR...	Replaceme...	VAR	8.25	2063	203655	68972.39	.	.	
19	REP_NINQ	Replaceme...	VAR	8.9	0	6.21951	1.035514	.	.	
20	REP_VALUE	Replaceme...	VAR	1.8	8000	279896.4	95543.69	.	.	
21	REP_YOJ	Replaceme...	VAR	7.85	0	31.9307	8.503604	.	.	
22	VALUE		VAR	1.8	8000	308600	95569.58	.	.	
23	YOJ		VAR	7.85	0	41	8.515111	.	.	
24	_dataobs_	Observatio...	VAR	0	1	5037	2517.277	.	.	

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?*

Variables

Variable Name	Role	Use	Method	Use Tree	Measurement Level	Order	Label ▲
BAD	Target	Default	DEFAULT	D	Binary		
CLAGE	Rejected	Default	DEFAULT	D	Interval		
CLNO	Rejected	Default	DEFAULT	D	Interval		
DEBTINC	Rejected	Default	DEFAULT	D	Interval		
DELINQ	Rejected	Default	DEFAULT	D	Interval		
DEROG	Rejected	Default	DEFAULT	D	Interval		
JOB	Input	Default	DEFAULT	D	Nominal		
LOAN	Rejected	Default	DEFAULT	D	Interval		
MORTDUE	Rejected	Default	DEFAULT	D	Interval		
NINQ	Rejected	Default	DEFAULT	D	Interval		
REASON	Input	Default	DEFAULT	D	Binary		
VALUE	Rejected	Default	DEFAULT	D	Interval		
YOJ	Rejected	Default	DEFAULT	D	Interval		
REP_CLAGE	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_CLNO	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_DEBT...	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_DELI...	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_DER...	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_LOAN	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_MOR...	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_NINQ	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_VALUE	Input	Default	DEFAULT	D	Interval		Replaceme...
REP_YOJ	Input	Default	DEFAULT	D	Interval		Replaceme...

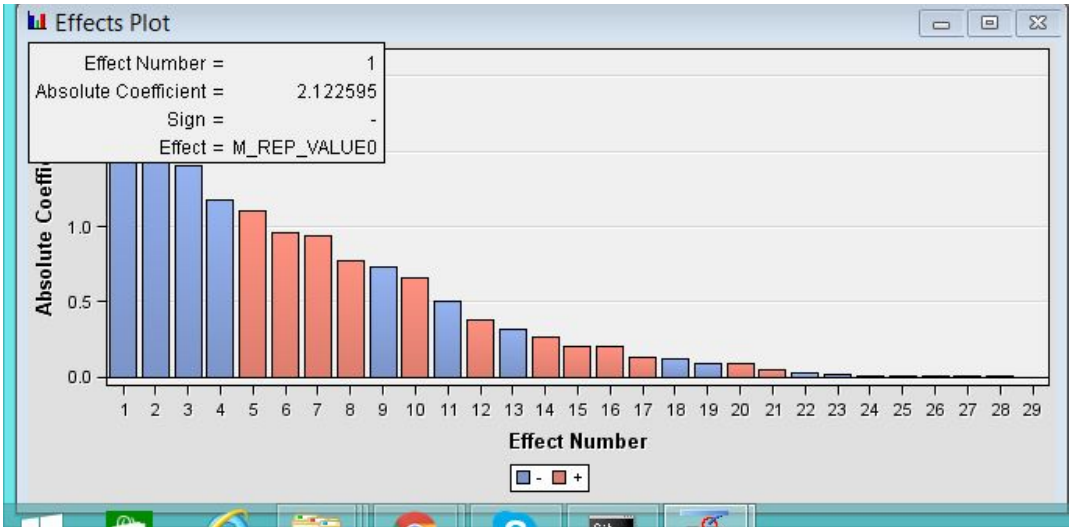
- *Содержат ли пропуски переменные с префиксом IMP_?* Нет
- *Чему равно значение переменной M_IMP_JOB, если значение исходной переменной IMP_JOB было “Office”?* M_IMP_JOB=0

Построение регрессионной модели

- *Какие переменные вошли в модель?*

Variable Name	Use	Report	Role	Measurement Level	Order	Label
BAD	Yes	No	Target	Binary		
CLAGE	Default	No	Rejected	Interval		
CLNO	Default	No	Rejected	Interval		
DEBTINC	Default	No	Rejected	Interval		
DELINQ	Default	No	Rejected	Interval		
DEROG	Default	No	Rejected	Interval		
IMP_JOB	Default	No	Input	Nominal		Imputed JOB
IMP_REAS...	Default	No	Input	Binary		Imputed RE...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
LOAN	Default	No	Rejected	Interval		
MORTDUE	Default	No	Rejected	Interval		
M_JOB	Default	No	Input	Binary		Imputation I...
M_REASON	Default	No	Input	Binary		Imputation I...
M_REP_CL...	Default	No	Input	Binary		Imputation I...
M_REP_CL...	Default	No	Input	Binary		Imputation I...
M_REP_DE...	Default	No	Input	Binary		Imputation I...
M_REP_DE...	Default	No	Input	Binary		Imputation I...
M_REP_DE...	Default	No	Input	Binary		Imputation I...
M_REP_M...	Default	No	Input	Binary		Imputation I...
M_REP_NI...	Default	No	Input	Binary		Imputation I...
M_REP_VA...	Default	No	Input	Binary		Imputation I...
M_REP_YOJ	Default	No	Input	Binary		Imputation I...
NINQ	Default	No	Rejected	Interval		
REP_LOAN	Default	No	Input	Interval		Replaceme...
VALUE	Default	No	Rejected	Interval		
YOJ	Default	No	Rejected	Interval		

- *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:

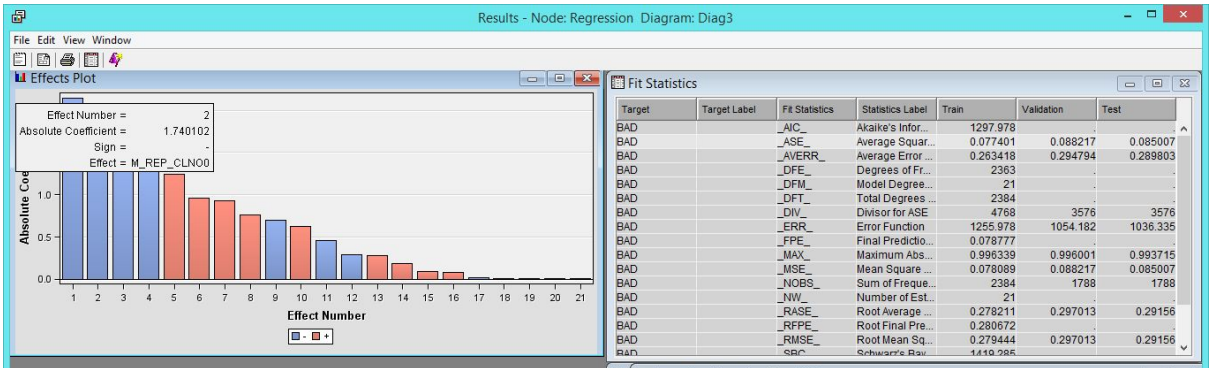
Вариант I: Forward

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



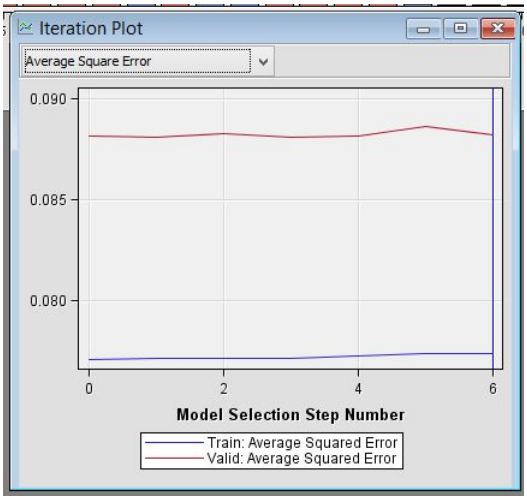
- *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

- *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)*

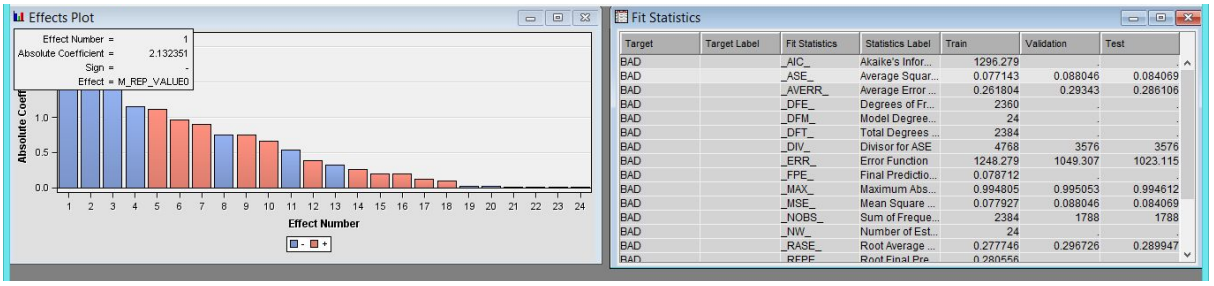
Variables

Variable Name	Use	Report	Role	Measurement Level	Order	Label
BAD	Yes	No	Target	Binary		
CLAGE	Default	No	Rejected	Interval		
CLNO	Default	No	Rejected	Interval		
DEBTINC	Default	No	Rejected	Interval		
DELINQ	Default	No	Rejected	Interval		
DEROG	Default	No	Rejected	Interval		
IMP_JOB	Default	No	Input	Nominal		Imputed JOB
IMP_REAS...	Default	No	Input	Binary		Imputed RE...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
IMP_REP_...	Default	No	Input	Interval		Imputed: R...
LOAN	Default	No	Rejected	Interval		
MORTDUE	Default	No	Rejected	Interval		
M_JOB	Default	No	Input	Binary		Imputation I...
M_REASON	Default	No	Input	Binary		Imputation I...
M_REP_CL...	Default	No	Input	Binary		Imputation I...
M_REP_CL...	Default	No	Input	Binary		Imputation I...
M_REP_DE...	Default	No	Input	Binary		Imputation I...
M_REP_DE...	Default	No	Input	Binary		Imputation I...
M_REP_DE...	Default	No	Input	Binary		Imputation I...
M_REP_M...	Default	No	Input	Binary		Imputation I...
M_REP_NI...	Default	No	Input	Binary		Imputation I...
M_REP_VA...	Default	No	Input	Binary		Imputation I...
M_REP_YOJ	Default	No	Input	Binary		Imputation I...
NINQ	Default	No	Rejected	Interval		
REP_LOAN	Default	No	Input	Interval		Replaceme...
VALUE	Default	No	Rejected	Interval		
YOJ	Default	No	Rejected	Interval		

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

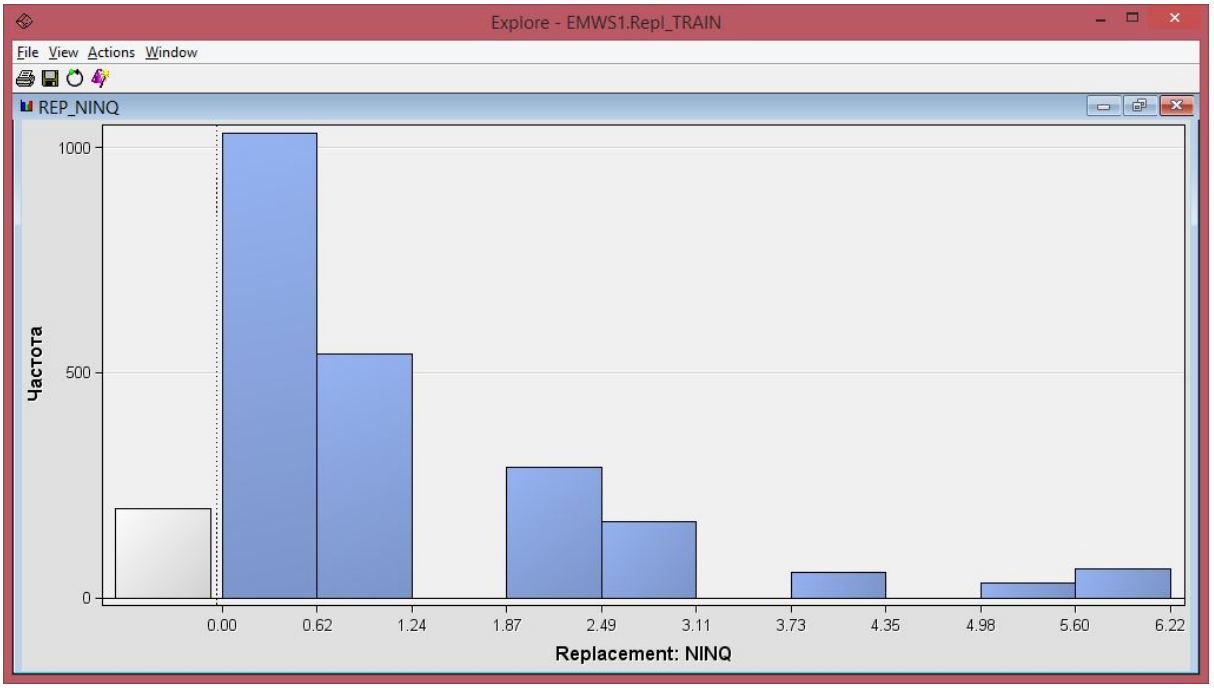
- *Посмотрите на график view->model->iteration plot, выбрав в критериях “Error function”. Обратите внимание вертикальную синюю черту. На каком шаге была выбрана финальная модель?* На 3-ем шаге

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?*



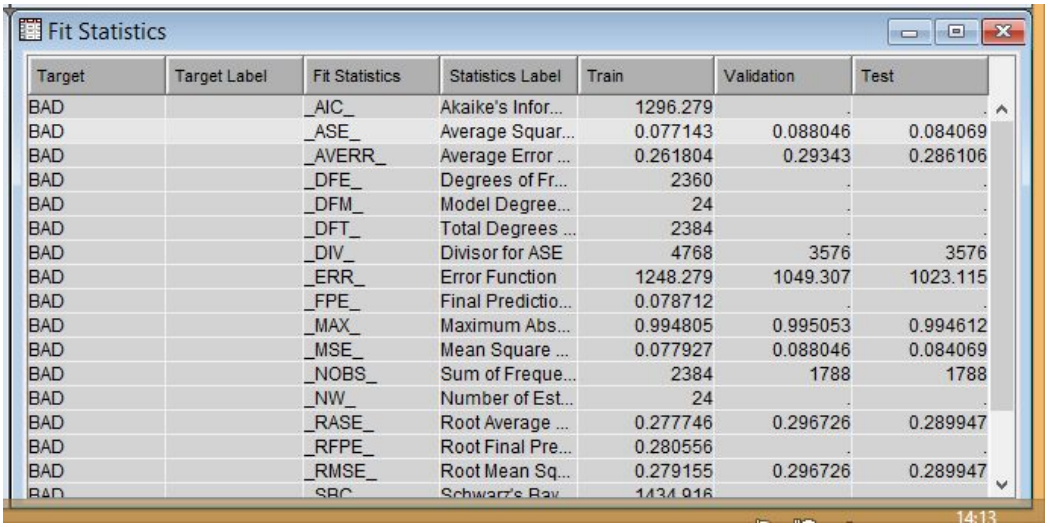
Присутствует левосторонняя асимметрия. Положительное значение коэффициента асимметрии указывает, что размер правого «хвоста» распределения больше, чем левого (относительно среднего).

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?*


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_AIC_	Akaike's Infor...	1296.279	.	.
BAD		_ASE_	Average Squar...	0.077143	0.088046	0.084069
BAD		_AVERR_	Average Error ...	0.261804	0.29343	0.286106
BAD		_DFE_	Degrees of Fr...	2360	.	.
BAD		_DFM_	Model Degree...	24	.	.
BAD		_DFT_	Total Degrees ...	2384	.	.
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_ERR_	Error Function	1248.279	1049.307	1023.115
BAD		_FPE_	Final Predictio...	0.078712	.	.
BAD		_MAX_	Maximum Abs...	0.994805	0.995053	0.994612
BAD		_MSE_	Mean Square ...	0.077927	0.088046	0.084069
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_NW_	Number of Est...	24	.	.
BAD		_RASE_	Root Average ...	0.277746	0.296726	0.289947
BAD		_RFPE_	Root Final Pre...	0.280556	.	.
BAD		_RMSE_	Root Mean Sq...	0.279155	0.296726	0.289947
BAD		_SBC_	Schwarz's Bay...	1434.916		

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?*

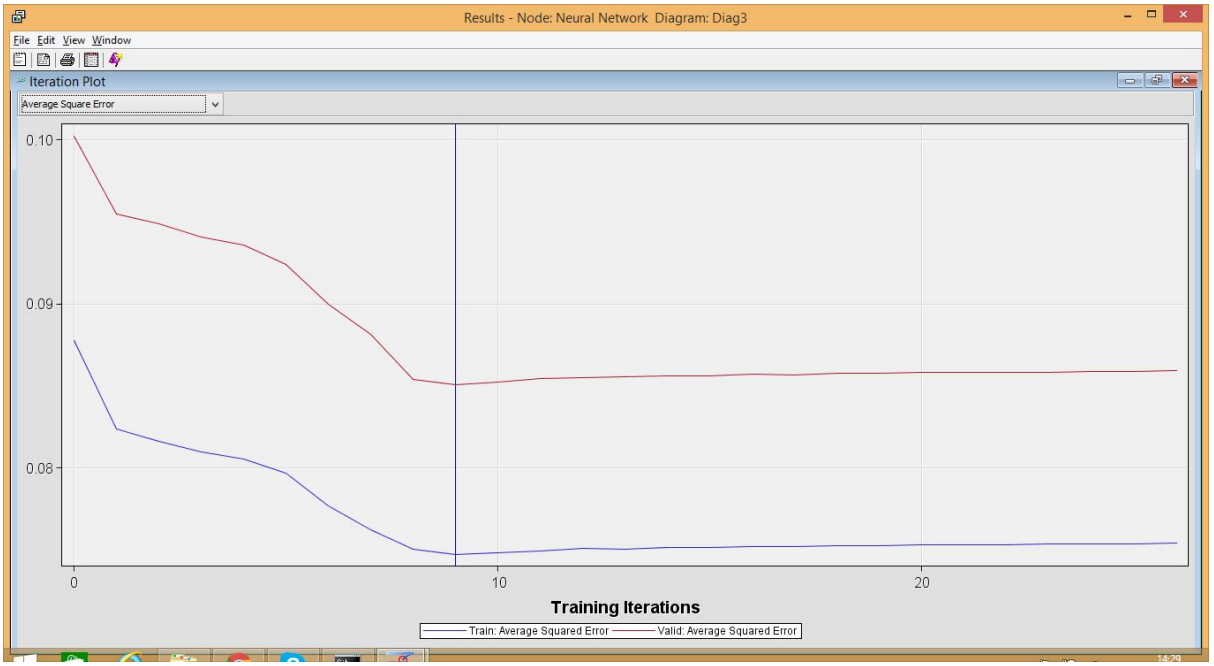
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_DFT_	Total Degrees of Freedom	2384	.	.
BAD		_DFE_	Degrees of Freedom for Error	2291	.	.
BAD		_DFM_	Model Degrees of Freedom	93	.	.
BAD		_NW_	Number of Estimated Weights	93	.	.
BAD		_AIC_	Akaike's Information Criterion	1397.53	.	.
BAD		_SBC_	Schwarz's Bayesian Criterion	1934.748	.	.
BAD		_ASE_	Average Squared Error	0.074726	0.085046	0.081716
BAD		_MAX_	Maximum Absolute Error	0.989486	0.986088	0.990279
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_NOBS_	Sum of Frequencies	2384	1788	1788
BAD		_RASE_	Root Average Squared Error	0.27336	0.291626	0.28586
BAD		_SSE_	Sum of Squared Errors	356.2923	304.1233	292.2159
BAD		_SUMW_	Sum of Case Weights Times Freq	4768	3576	3576
BAD		_FPE_	Final Prediction Error	0.080793	.	.
BAD		_MSE_	Mean Squared Error	0.077759	0.085046	0.081716
BAD		_RFPE_	Root Final Prediction Error	0.28424		

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?*

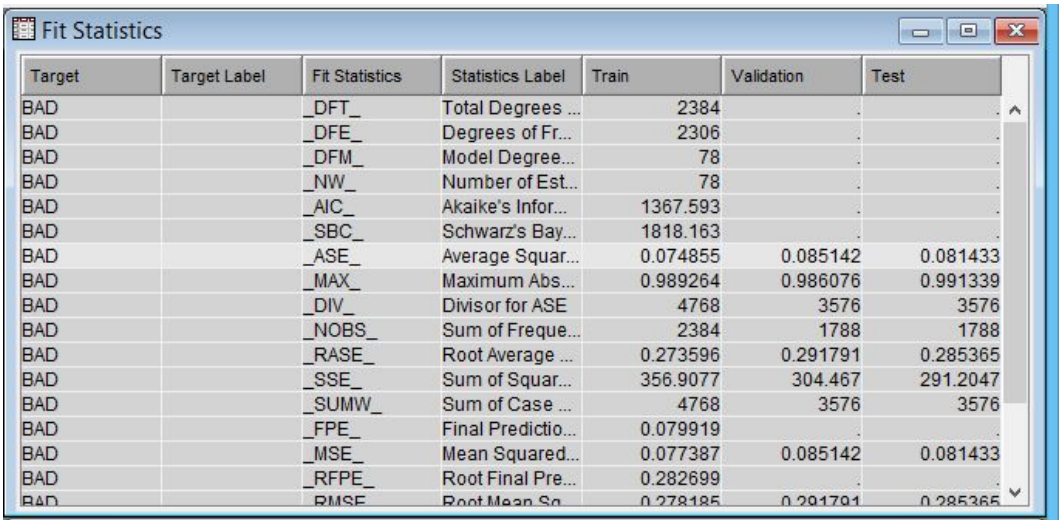
- *Обратите внимание на график Iteration plot: на каком шаге обучения выбрана оптимальная модель?*



Выбор значимых переменных

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

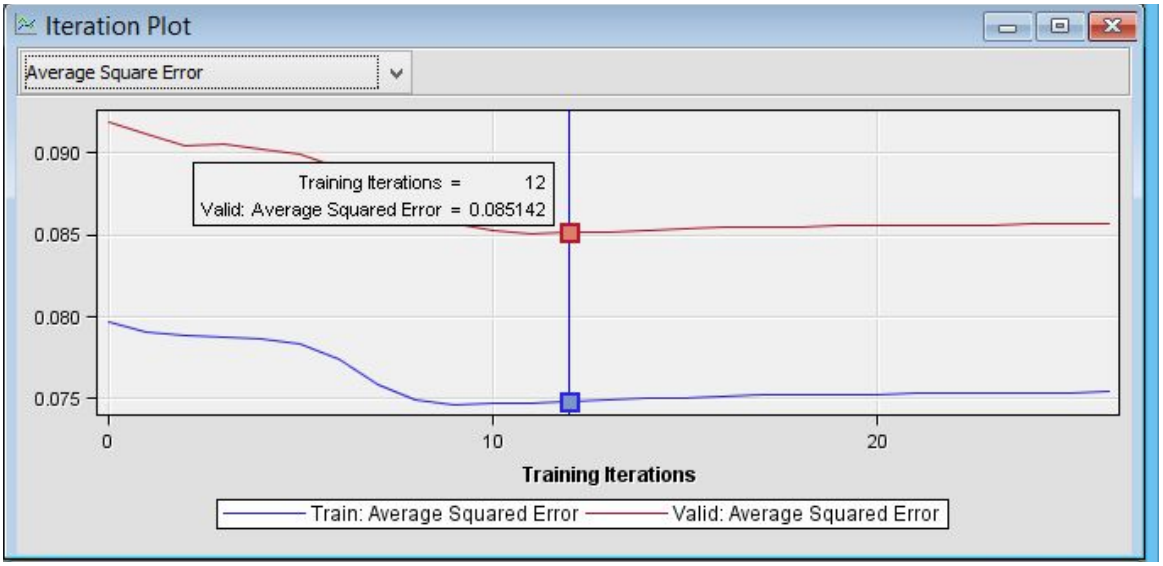
Число степеней свободы - это минимально необходимое число значений зависимой переменной, которых достаточно для получения искомой характеристики выборки и которые могут свободно варьироваться с учетом того, что для этой выборки известны все другие величины, используемые для расчета искомой характеристики.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



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_DFT_	Total Degrees ...	2384	.	.
BAD		_DFE_	Degrees of Fr...	2306	.	.
BAD		_DFM_	Model Degree...	78	.	.
BAD		_NW_	Number of Est...	78	.	.
BAD		_AIC_	Akaike's Infor...	1367.593	.	.
BAD		_SBC_	Schwarz's Bay...	1818.163	.	.
BAD		_ASE_	Average Squar...	0.074855	0.085142	0.081433
BAD		_MAX_	Maximum Abs...	0.989264	0.986076	0.991339
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_RASE_	Root Average ...	0.273596	0.291791	0.285365
BAD		_SSE_	Sum of Squar...	356.9077	304.467	291.2047
BAD		_SUMW_	Sum of Case ...	4768	3576	3576
BAD		_FPE_	Final Predictio...	0.079919	.	.
BAD		_MSE_	Mean Squared...	0.077387	0.085142	0.081433
BAD		_RFPE_	Root Final Pre...	0.282699	.	.
BAD		_RMSE_	Root Mean Sq...	0.278185	0.291791	0.285365

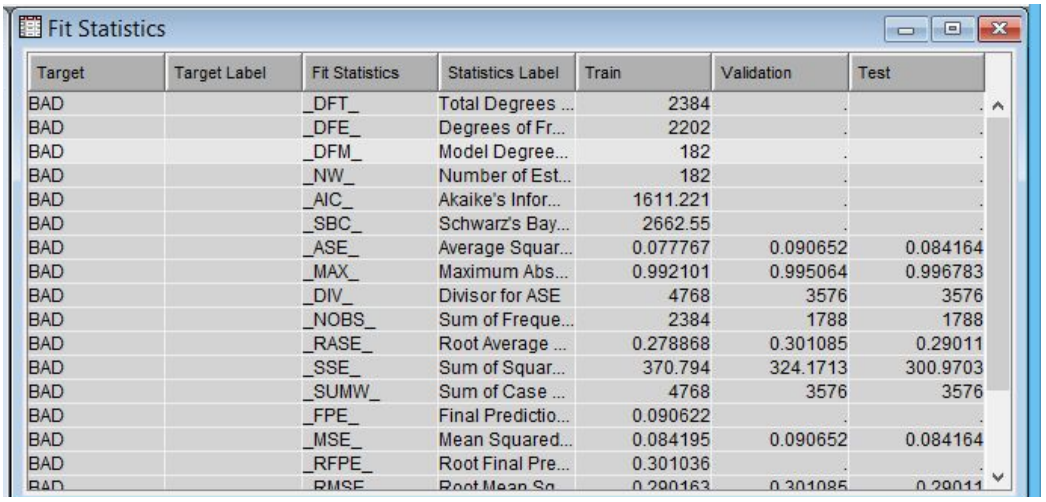
- *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*

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



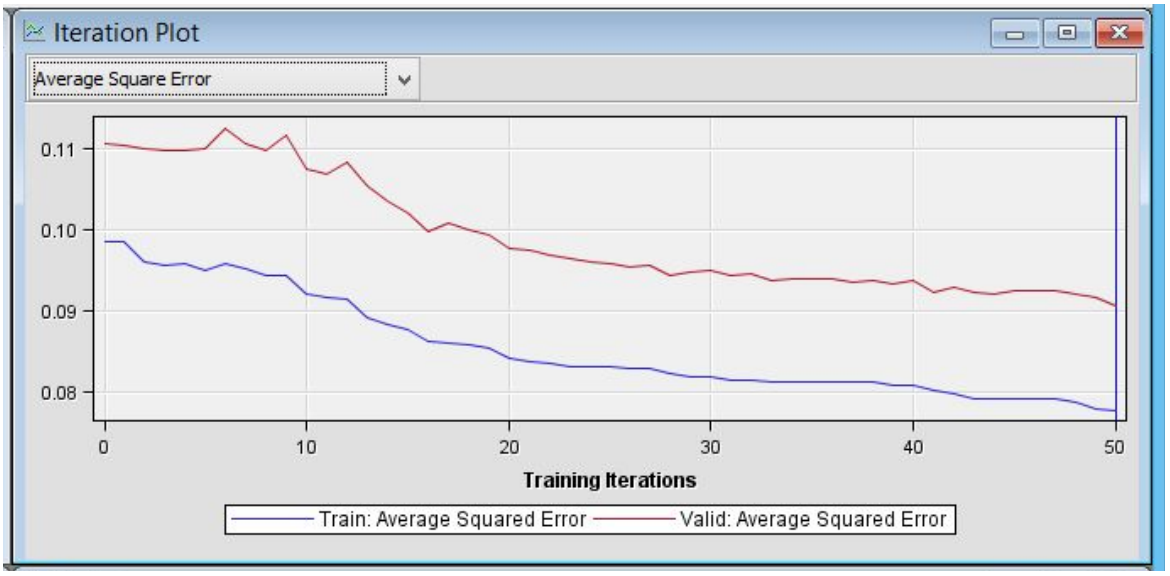
Усложнение архитектуры

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


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_DFT_	Total Degrees ...	2384	.	.
BAD		_DFE_	Degrees of Fr...	2202	.	.
BAD		_DFM_	Model Degree...	182	.	.
BAD		_NW_	Number of Est...	182	.	.
BAD		_AIC_	Akaike's Infor...	1611.221	.	.
BAD		_SBC_	Schwarz's Bay...	2662.55	.	.
BAD		_ASE_	Average Squar...	0.077767	0.090652	0.084164
BAD		_MAX_	Maximum Abs...	0.992101	0.995064	0.996783
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_RASE_	Root Average ...	0.278868	0.301085	0.29011
BAD		_SSE_	Sum of Squar...	370.794	324.1713	300.9703
BAD		_SUMW_	Sum of Case ...	4768	3576	3576
BAD		_FPE_	Final Predictio...	0.090622	.	.
BAD		_MSE_	Mean Squared...	0.084195	0.090652	0.084164
BAD		_RFPE_	Root Final Pre...	0.301036	.	.
BAD		_RMSE_	Root Mean Sq...	0.290163	0.301085	0.29011

- *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

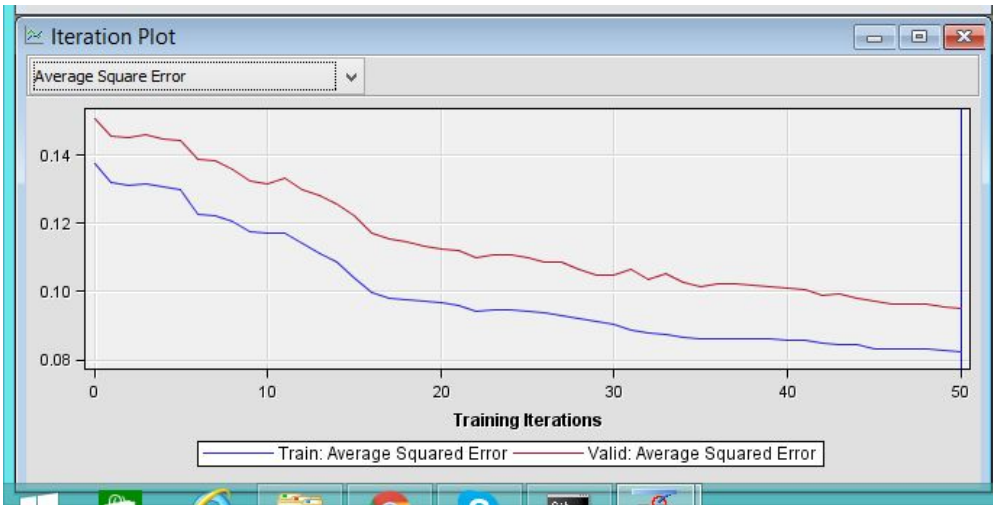
Подключите узел «SOM» после «Impute», и подключите вашу нейронную сеть после узла SOM. Обучите модель.

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

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_DFT_	Total Degrees ...	2384	.	.
BAD		_DFE_	Degrees of Fr...	2139	.	.
BAD		_DFM_	Model Degree...	245	.	.
BAD		_NW_	Number of Est...	245	.	.
BAD		_AIC_	Akaike's Infor...	1811.405	.	.
BAD		_SBC_	Schwarz's Bay...	3226.656	.	.
BAD		_ASE_	Average Squar...	0.082431	0.095033	0.087246
BAD		_MAX_	Maximum Abs...	0.985801	0.986796	0.993121
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_RASE_	Root Average ...	0.287108	0.308274	0.295375
BAD		_SSE_	Sum of Squar...	393.0307	339.8373	311.9931
BAD		_SUMW_	Sum of Case ...	4768	3576	3576
BAD		_FPE_	Final Predictio...	0.101314	.	.
BAD		_MSE_	Mean Squared...	0.091873	0.095033	0.087246
BAD		_RFPE_	Root Final Pre...	0.318299	.	.
BAD		_RMSE_	Root Mean Sq...	0.303105	0.308274	0.295375

- *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»?*

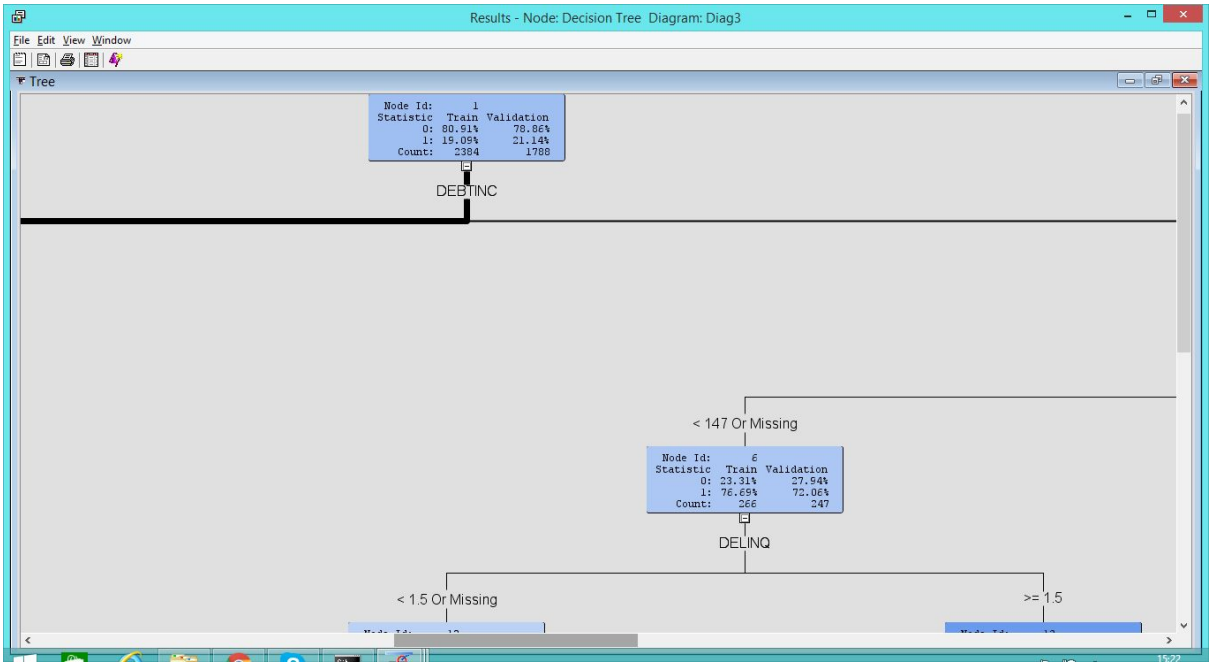
Потому что дерево решений используется как вспомогательный инструмент например для замены пропущенных значений, что нельзя сказать о регрессии или о нейронной сети.

- *Сколько листьев получилось в дереве?*
  32

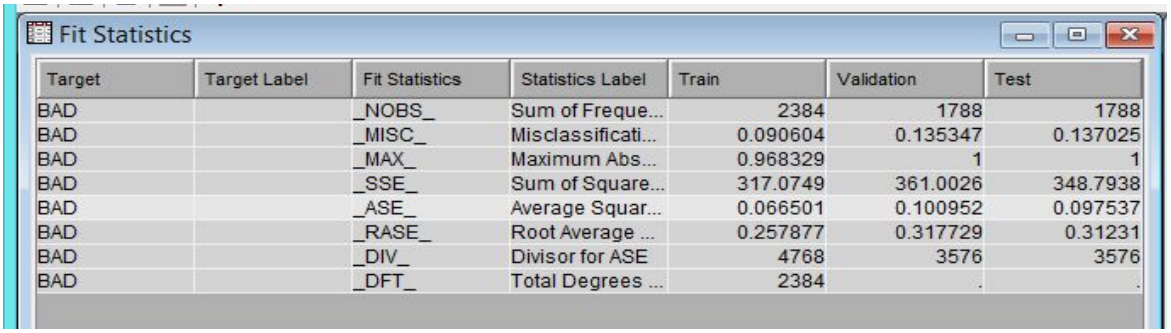
- *Какие переменные вошли в модель?*

DEBTINC DELINQ VALUE CLAGE DEROG MORTDUE LOAN JOB CLNO YOJ NINQ

- *По какой переменной произошло первое разбиение?* DEBTINC

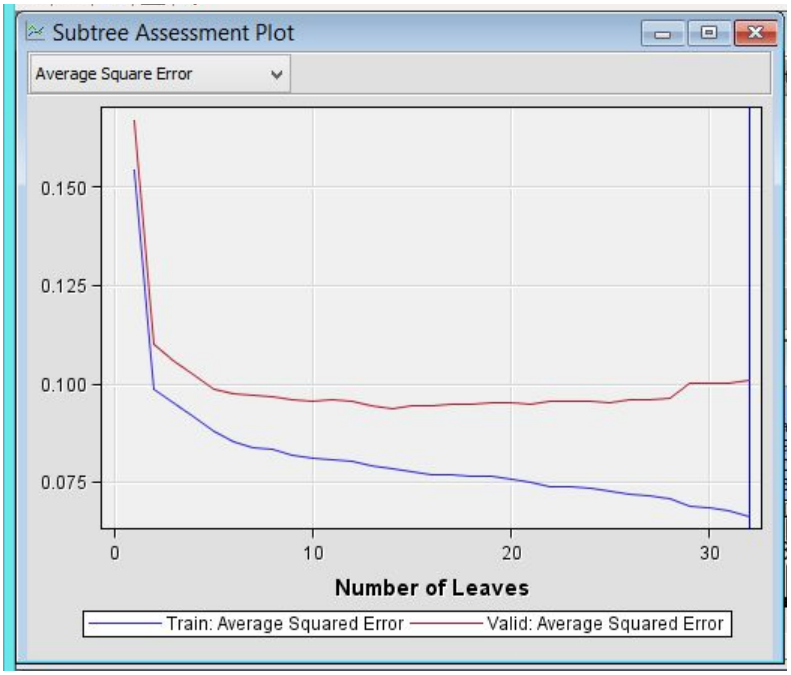


- *Каково значение среднеквадратичной ошибки ASE на тренировочном, тестовом и валидационном наборах?*



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_MISC_	Misclassificati...	0.090604	0.135347	0.137025
BAD		_MAX_	Maximum Abs...	0.968329	1	1
BAD		_SSE_	Sum of Square...	317.0749	361.0026	348.7938
BAD		_ASE_	Average Squar...	0.066501	0.100952	0.097537
BAD		_RASE_	Root Average ...	0.257877	0.317729	0.31231
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_DFT_	Total Degrees ...	2384	.	.

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



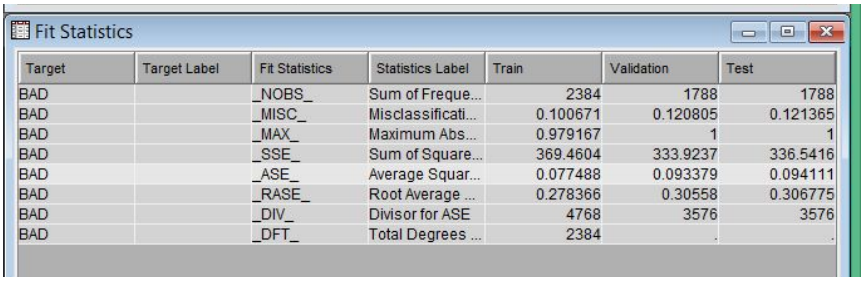
Обрубание дерева решений

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

- *Какие переменные вошли в модель?*

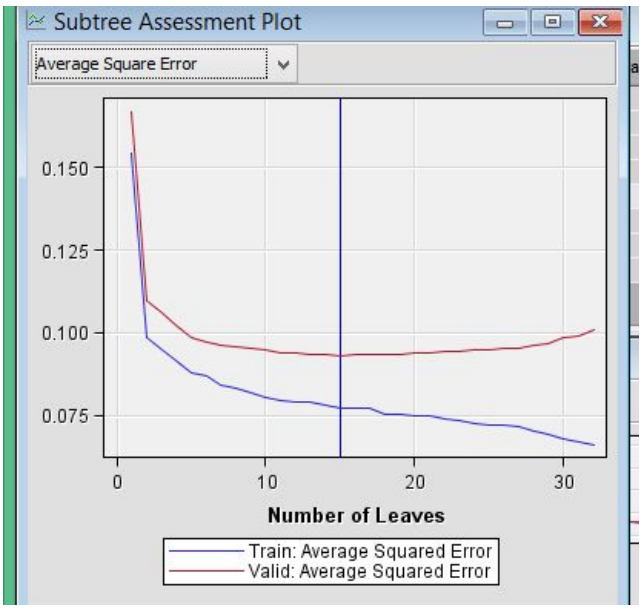
DEBTINC DELINQ CLAGE VALUE DEROG CLNO LOAN

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



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_MISC_	Misclassificati...	0.100671	0.120805	0.121365
BAD		_MAX_	Maximum Abs...	0.979167	1	1
BAD		_SSE_	Sum of Square...	369.4604	333.9237	336.5416
BAD		_ASE_	Average Squar...	0.077488	0.093379	0.094111
BAD		_RASE_	Root Average ...	0.278366	0.30558	0.306775
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_DFT_	Total Degrees ...	2384	.	.

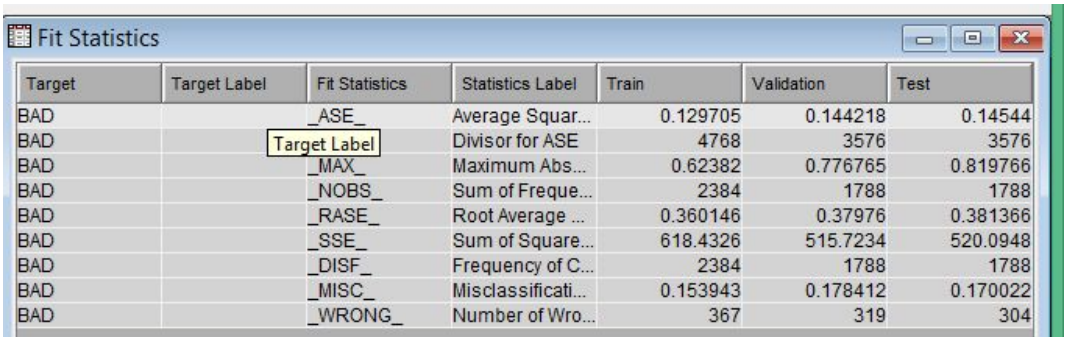
- *Посмотрите на график View->Model->subtree assessment plot. Какой размер дерева (при критерии ASE) был выбран как оптимальный?* 15



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?*


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
BAD		_ASE_	Average Squar...	0.129705	0.144218	0.14544
BAD		_DIV_	Divisor for ASE	4768	3576	3576
BAD		_MAX_	Maximum Abs...	0.62382	0.776765	0.819766
BAD		_NOBS_	Sum of Freque...	2384	1788	1788
BAD		_RASE_	Root Average ...	0.360146	0.37976	0.381366
BAD		_SSE_	Sum of Square...	618.4326	515.7234	520.0948
BAD		_DISF_	Frequency of C...	2384	1788	1788
BAD		_MISC_	Misclassificati...	0.153943	0.178412	0.170022
BAD		_WRONG_	Number of Wro...	367	319	304

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
HPSVM0,98	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
HPSVM0,16	0,85	0,84	
Tree	0,16	0,69	0,69
Neural	0,15	0,52	0,56