

1 Data import and preparation

This part of the document deals with the data preparation of the provided cooper wire data before the data analysis.

1.1 Data import

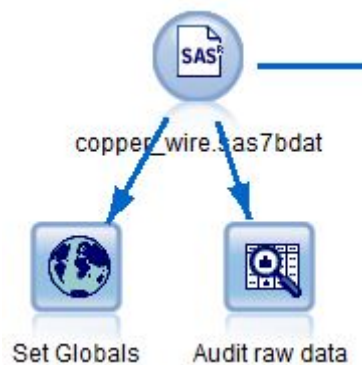


Abbildung 1: Data import in Stream

The data is imported via the node *SAS file*. The node *Set Globals* is used for setting the audited data results of the raw imported data as global values, which get used later on for the data preparation. The node *Data Audit* is used for analyzing the raw data.

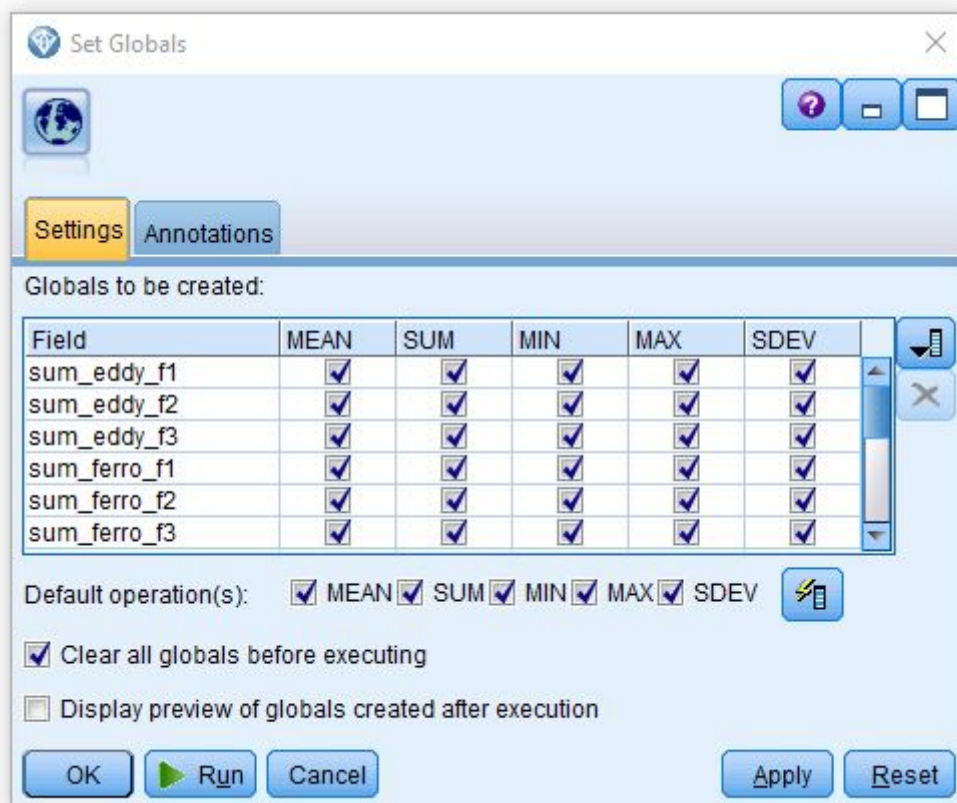
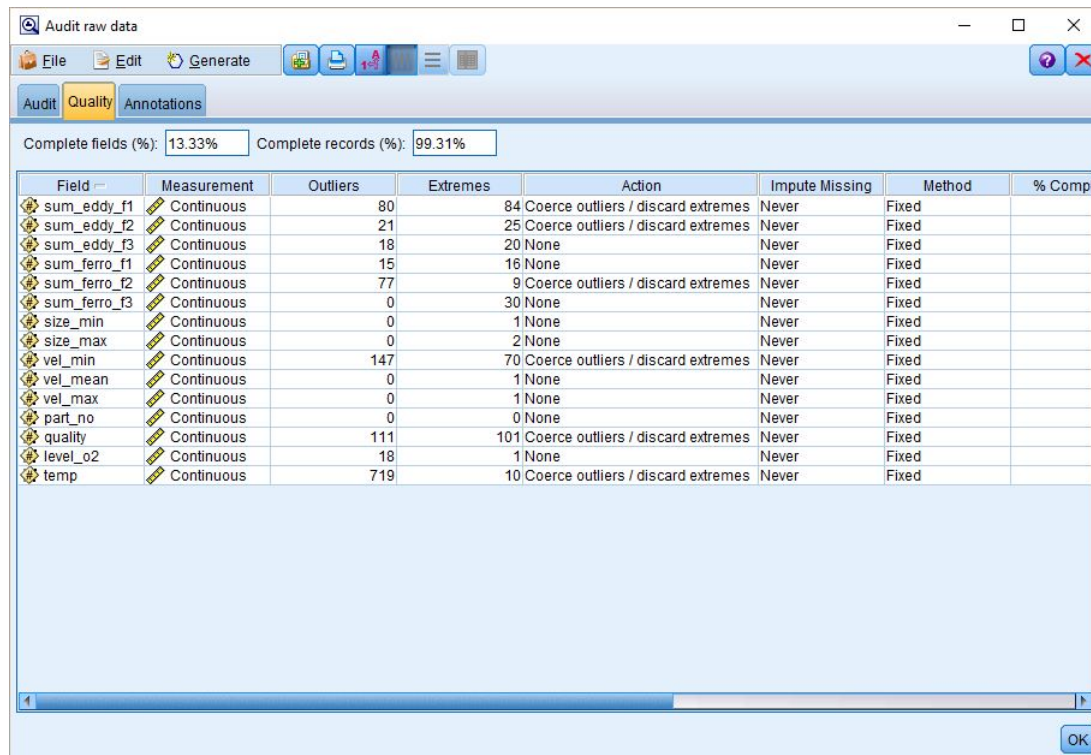


Abbildung 2: Set audit results as global values in the stream

1.2 Data preparation

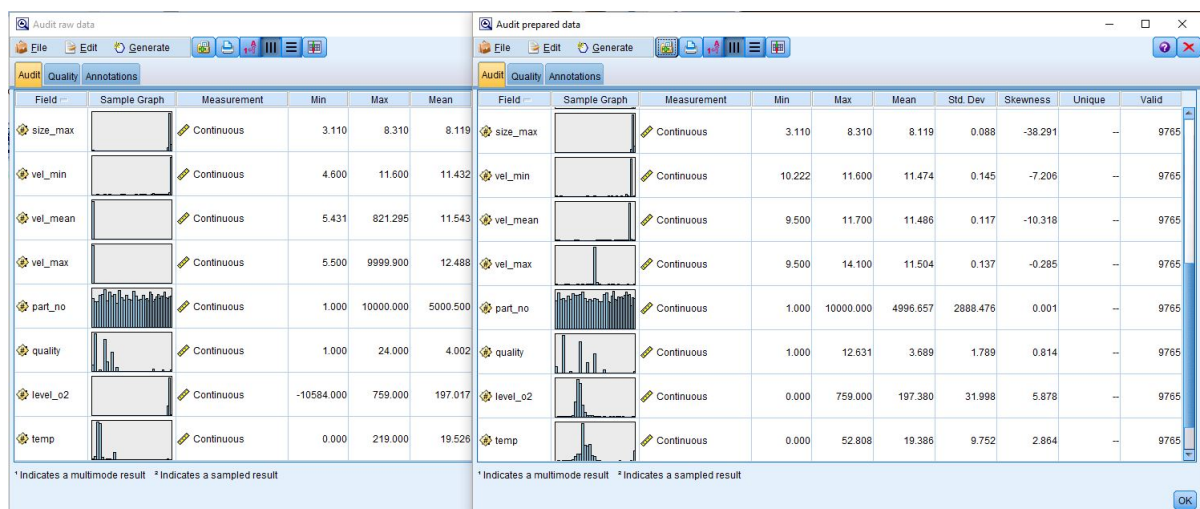
The outliers and extremes were determined during the audit of the raw data.



The screenshot shows the 'Audit raw data' window with the 'Quality' tab selected. It displays a table with columns: Field, Measurement, Outliers, Extremes, Action, Impute Missing, Method, and % Compl. The table lists 15 fields with their respective measurement types, outlier counts, extreme counts, and actions taken.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Compl
sum_eddy_f1	Continuous	80	84	Coerce outliers / discard extremes	Never	Fixed	
sum_eddy_f2	Continuous	21	25	Coerce outliers / discard extremes	Never	Fixed	
sum_eddy_f3	Continuous	18	20	None	Never	Fixed	
sum_ferro_f1	Continuous	15	16	None	Never	Fixed	
sum_ferro_f2	Continuous	77	9	Coerce outliers / discard extremes	Never	Fixed	
sum_ferro_f3	Continuous	0	30	None	Never	Fixed	
size_min	Continuous	0	1	None	Never	Fixed	
size_max	Continuous	0	2	None	Never	Fixed	
vel_min	Continuous	147	70	Coerce outliers / discard extremes	Never	Fixed	
vel_mean	Continuous	0	1	None	Never	Fixed	
vel_max	Continuous	0	1	None	Never	Fixed	
part_no	Continuous	0	0	None	Never	Fixed	
quality	Continuous	111	101	Coerce outliers / discard extremes	Never	Fixed	
level_o2	Continuous	18	1	None	Never	Fixed	
temp	Continuous	719	10	Coerce outliers / discard extremes	Never	Fixed	

Abbildung 3: Audit of the raw data



The screenshot shows the 'Audit prepared data' window with the 'Quality' tab selected. It displays a table with columns: Field, Sample Graph, Measurement, Min, Max, Mean, Std. Dev, Skewness, Unique, and Valid. The table lists 15 fields with their respective measurement types, statistical values, and unique counts.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
size_max		Continuous	3.110	8.310	8.119	0.088	-38.291	—	9765
vel_min		Continuous	4.600	11.600	11.432	0.145	-7.206	—	9765
vel_mean		Continuous	5.431	821.295	11.543	0.117	-10.318	—	9765
vel_max		Continuous	5.500	9999.900	12.488	0.137	-0.285	—	9765
part_no		Continuous	1.000	10000.000	5000.500	2888.476	0.001	—	9765
quality		Continuous	1.000	24.000	4.002	1.789	0.814	—	9765
level_o2		Continuous	-10584.000	759.000	197.017	31.998	5.878	—	9765
temp		Continuous	0.000	219.000	19.526	9.752	2.864	—	9765

Abbildung 4: Audit of the raw data

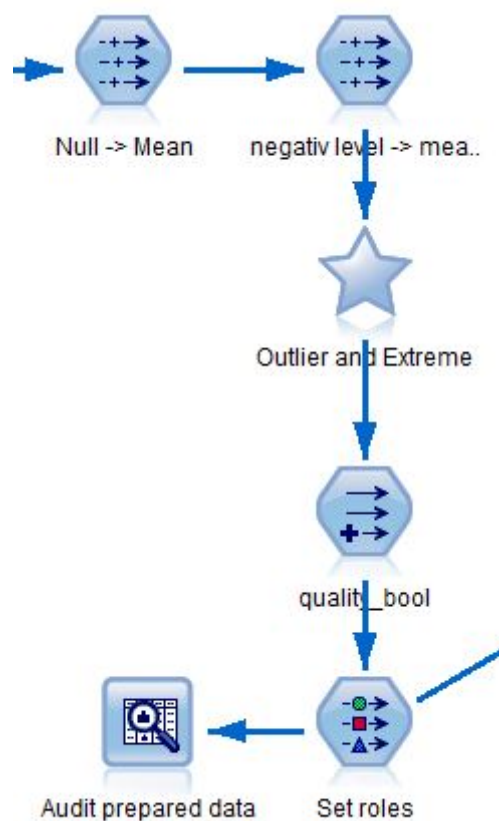


Abbildung 5: Flow of data preparation tasks

This flow prepares the data for the later analysis. The following tasks are performed:

- Null values will be replaced with the mean value set by the *Set Global* node
- The negative value of the field *temp* will be replaced with the global mean of this field
- The outliers and extremes will be handled as you can see in image 3
- A new field will be created *quality_bool* which represents the quality state good or false
- The fields which are not considered to be relevant will be set as ignored and the field *quality_bool* will be set as the target field for the further analysis

1.3 Predictive Model

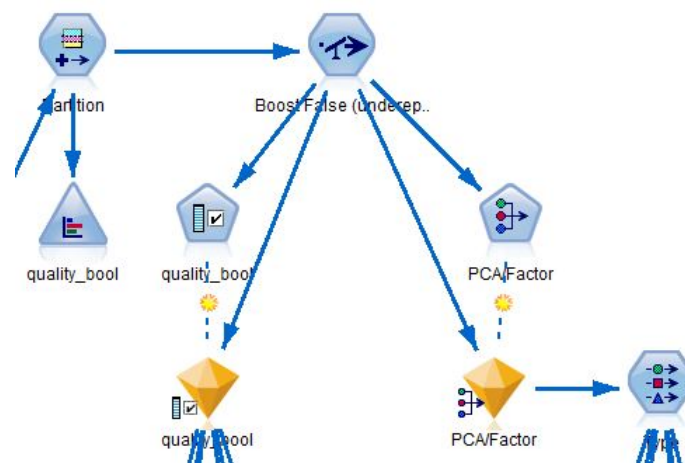


Abbildung 6: Flow of further data preparation

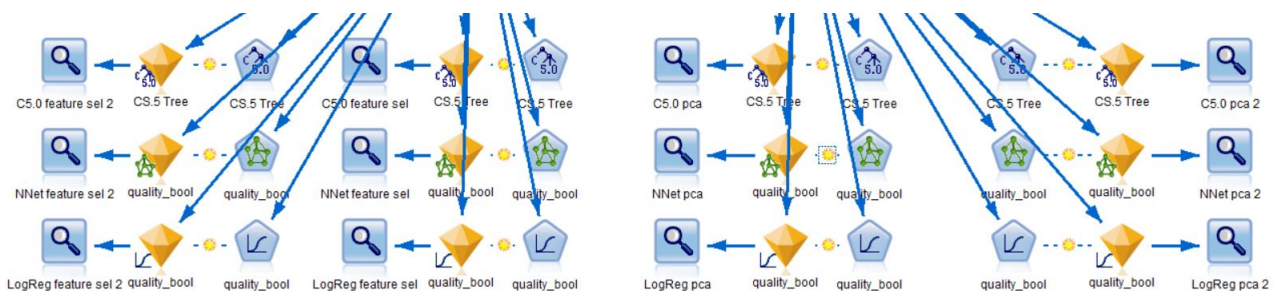
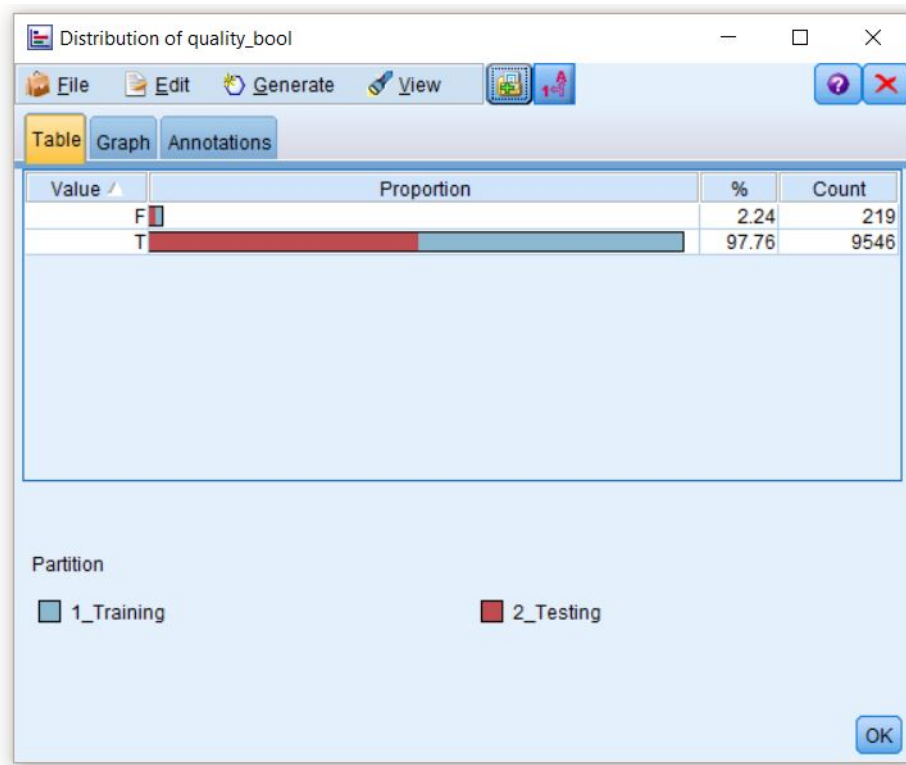
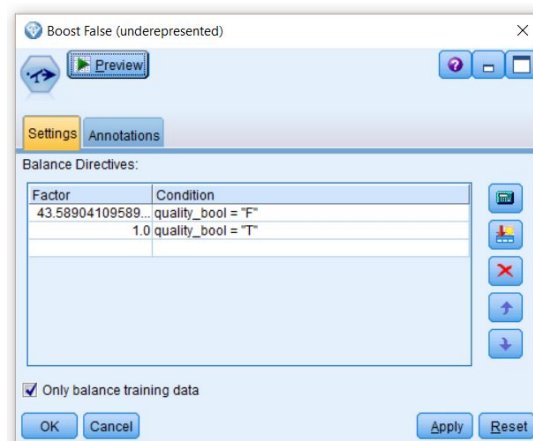


Abbildung 7: Flow of predication models

This part of the stream prepares the data by splitting it into test and training data, additionally the *False quality_bool* will be boost to increase their representation in the data. At last the data gets prepared on the one hand with a *Feature Selection* and on the other hand with a *PCA Factor*.

Abbildung 8: Badly distributed *quality_bool*

As we can see that the *False* quality is underrepresented compared to the *True* quality.

Abbildung 9: Boost of quality *False*

The node *Balance* has been generated by the node *Distribution* and boost the representation of the *False* quality. After this nodes follows the node *Field selection* which removes fields which are not related to the *target*.

The node *Field selection* has reduced the count of fields from 15 down to 10, therefore has removed 5 fields.

1.3.1 What are results of the use of *Feature Selection* and *PCA Factor* with defaults ?

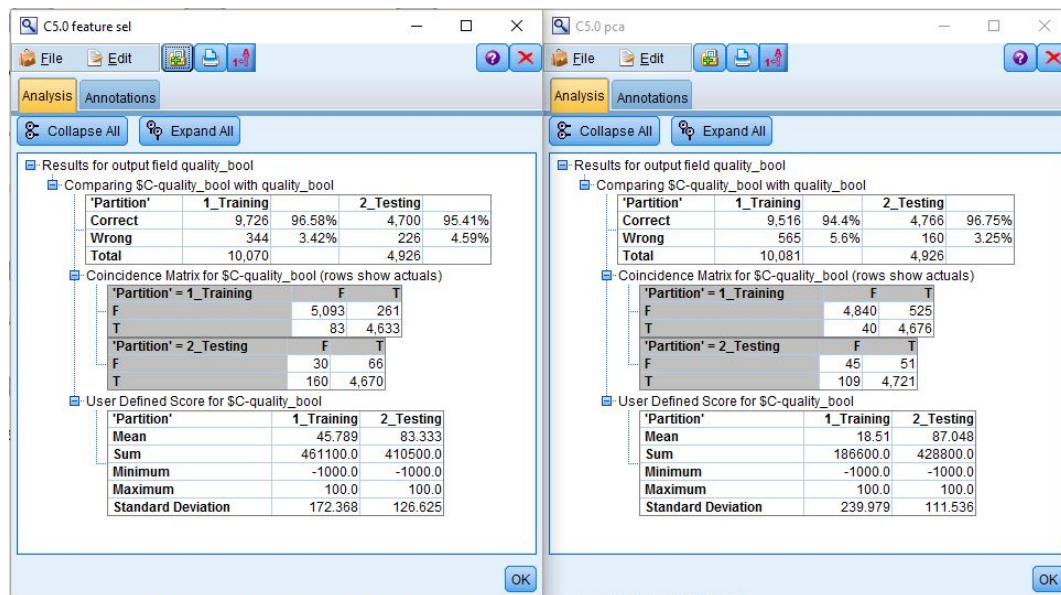


Abbildung 10: C5.0 with *Feature Selection* compared to *PCA*

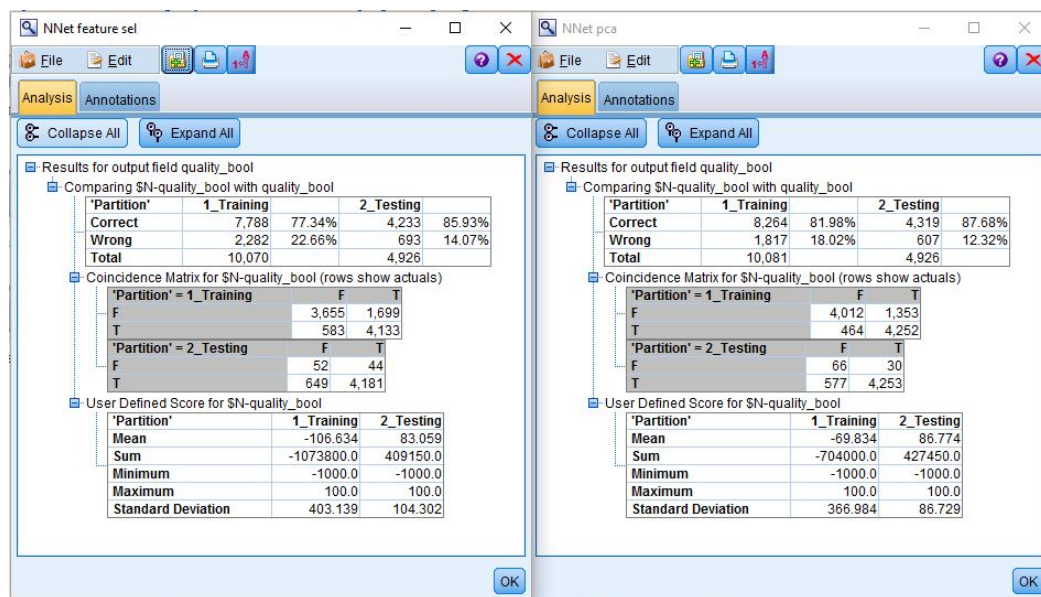
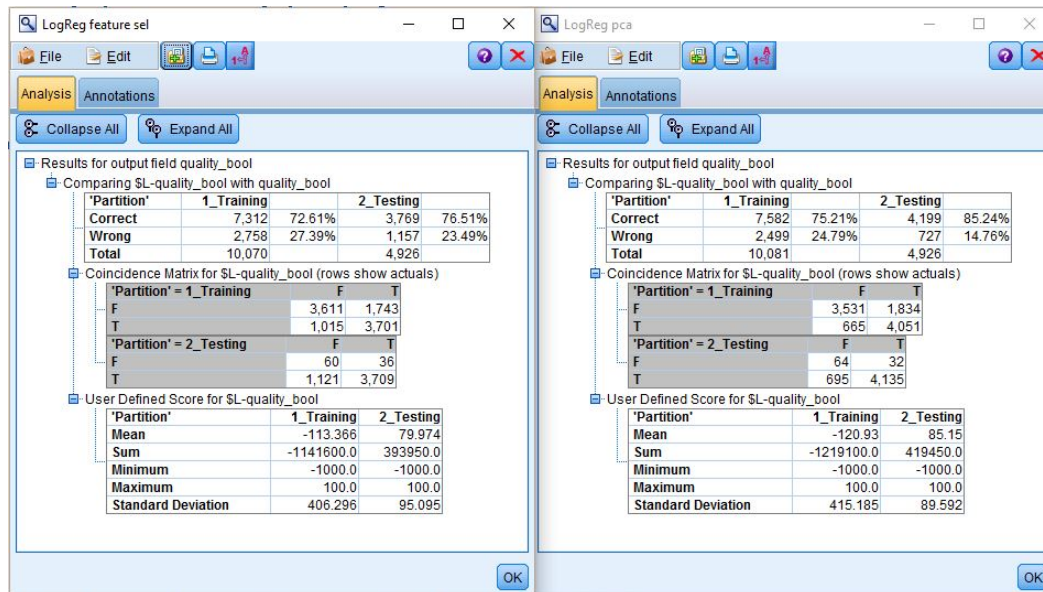
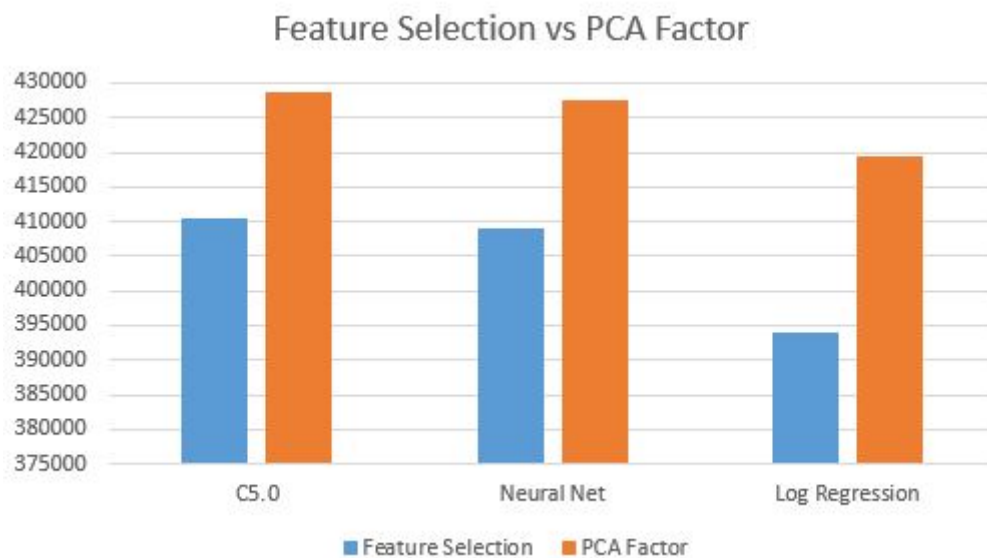
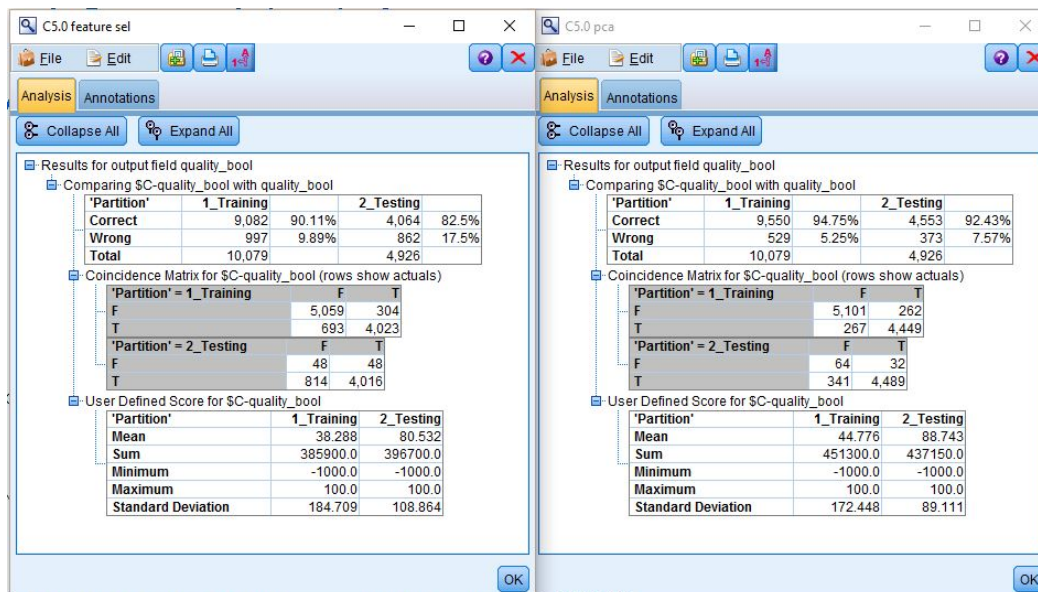


Abbildung 11: Neuronal Net with *Feature Selection* compared to *PCA*


Abbildung 12: Logistic Regression with *Feature Selection* compared to *PCA*

Abbildung 13: The chart shows the results of the comparison between *Feature Selection* and *PCA*

1.3.2 What are results when C5.0 is modified ?



C5.0 feature sel

Results for output field quality_bool

Comparing \$C-quality_bool with quality_bool

'Partition'	1_Training	2_Testing	
Correct	9,082	4,064	82.5%
Wrong	997	862	17.5%
Total	10,079	4,926	

Coincidence Matrix for \$C-quality_bool (rows show actuals)

'Partition' = 1_Training

	F	T
F	5,059	304
T	693	4,023

'Partition' = 2_Testing

	F	T
F	48	48
T	814	4,016

User Defined Score for \$C-quality_bool

'Partition'	1_Training	2_Testing
Mean	38.288	80.532
Sum	385900.0	396700.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	184.709	108.864

C5.0 pca

Results for output field quality_bool

Comparing \$C-quality_bool with quality_bool

'Partition'	1_Training	2_Testing	
Correct	9,550	4,553	92.43%
Wrong	529	373	7.57%
Total	10,079	4,926	

Coincidence Matrix for \$C-quality_bool (rows show actuals)

'Partition' = 1_Training

	F	T
F	5,101	262
T	267	4,449

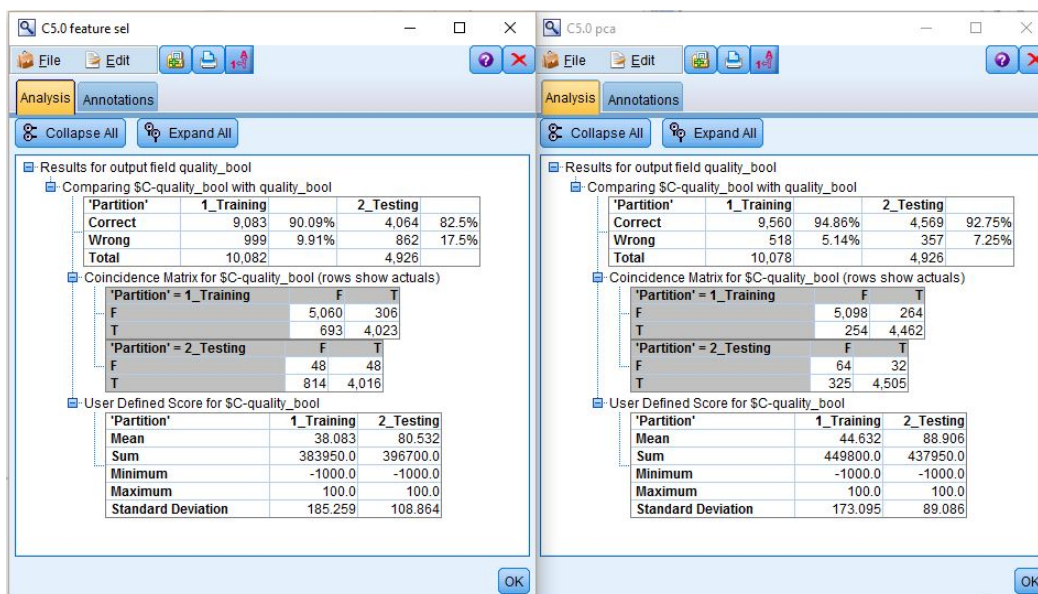
'Partition' = 2_Testing

	F	T
F	64	32
T	341	4,489

User Defined Score for \$C-quality_bool

'Partition'	1_Training	2_Testing
Mean	44.776	88.743
Sum	451300.0	437150.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	172.448	89.111

Abbildung 14: C5.0 (costs for TF=10) with *Feature Selection* compared to *PCA*



C5.0 feature sel

Results for output field quality_bool

Comparing \$C-quality_bool with quality_bool

'Partition'	1_Training	2_Testing	
Correct	9,083	4,064	82.5%
Wrong	999	862	17.5%
Total	10,082	4,926	

Coincidence Matrix for \$C-quality_bool (rows show actuals)

'Partition' = 1_Training

	F	T
F	5,060	306
T	693	4,023

'Partition' = 2_Testing

	F	T
F	48	48
T	814	4,016

User Defined Score for \$C-quality_bool

'Partition'	1_Training	2_Testing
Mean	38.083	80.532
Sum	383950.0	396700.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	185.259	108.864

C5.0 pca

Results for output field quality_bool

Comparing \$C-quality_bool with quality_bool

'Partition'	1_Training	2_Testing	
Correct	9,560	4,569	92.75%
Wrong	518	357	7.25%
Total	10,078	4,926	

Coincidence Matrix for \$C-quality_bool (rows show actuals)

'Partition' = 1_Training

	F	T
F	5,098	264
T	254	4,462

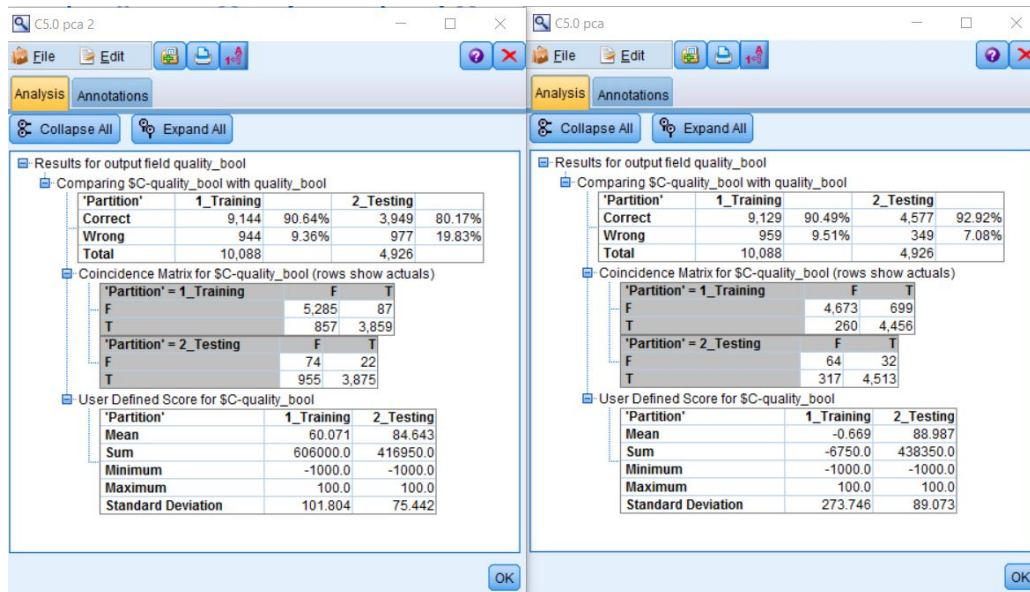
'Partition' = 2_Testing

	F	T
F	64	32
T	325	4,505

User Defined Score for \$C-quality_bool

'Partition'	1_Training	2_Testing
Mean	44.632	88.906
Sum	449800.0	437950.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	173.095	89.086

Abbildung 15: C5.0 (costs for TF=10, prun serv=10) with *Feature Selection* compared to *PCA*



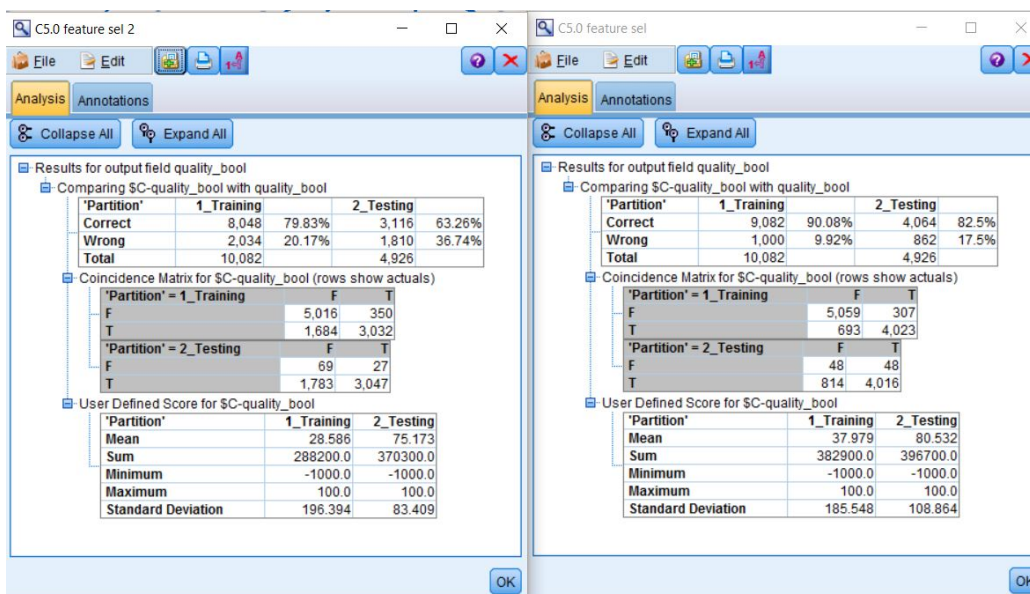
Left Window: C5.0 pca 2

'Partition'	1_Training	2_Testing
Correct	9,144	3,949
Wrong	944	977
Total	10,088	4,926

Right Window: C5.0 pca

'Partition'	1_Training	2_Testing
Correct	9,129	4,577
Wrong	959	349
Total	10,088	4,926

Abbildung 16: C5.0 (costs for TF=10, prun serv=1, rec=100) with PCA



Left Window: C5.0 feature sel 2

'Partition'	1_Training	2_Testing
Correct	8,048	3,116
Wrong	2,034	1,810
Total	10,082	4,926

Right Window: C5.0 feature sel

'Partition'	1_Training	2_Testing
Correct	9,082	4,064
Wrong	1,000	862
Total	10,082	4,926

Abbildung 17: C5.0 (costs for TF=10, prun serv=1, rec=100) with PCA

1.3.3 What are results when the Neural Net is modified ?

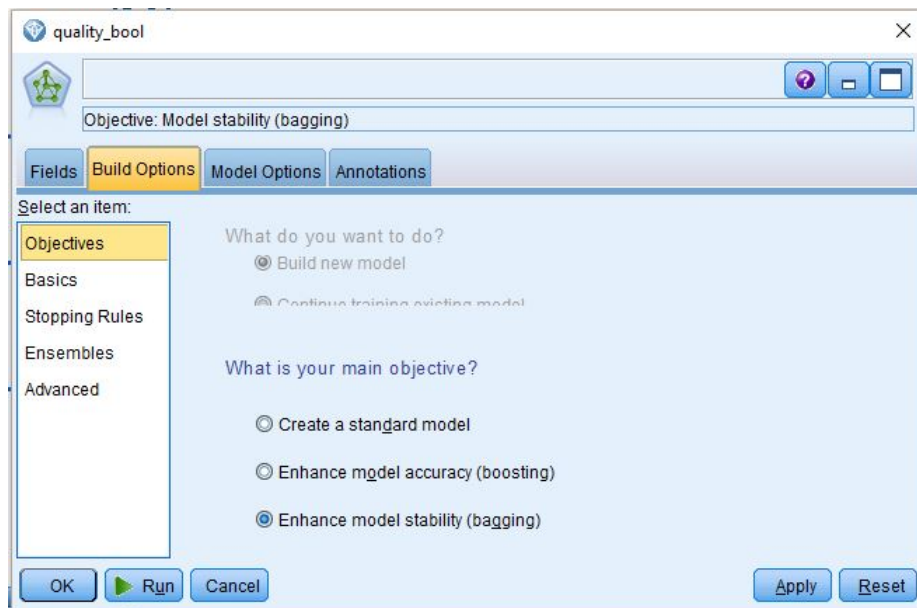


Abbildung 18: Part one of Neural Net settings

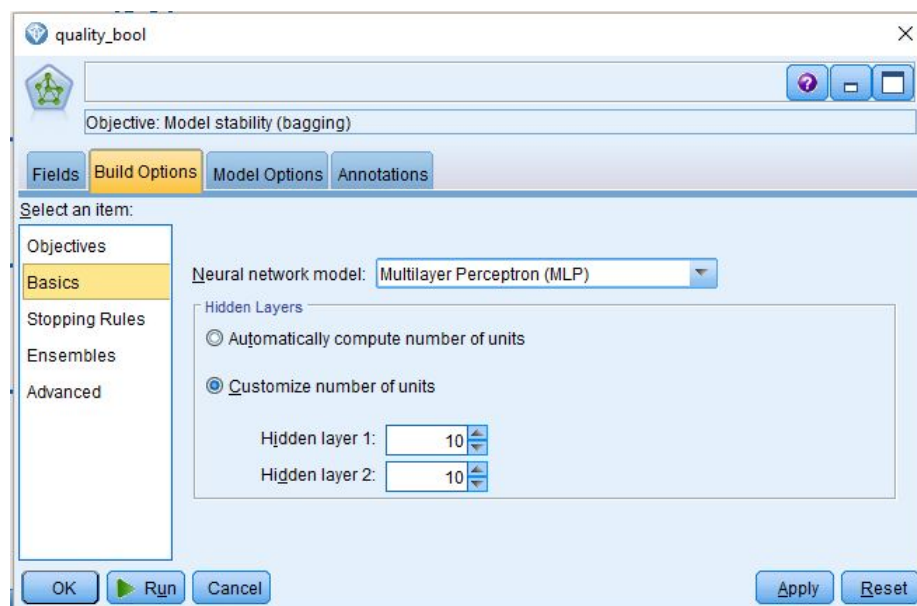
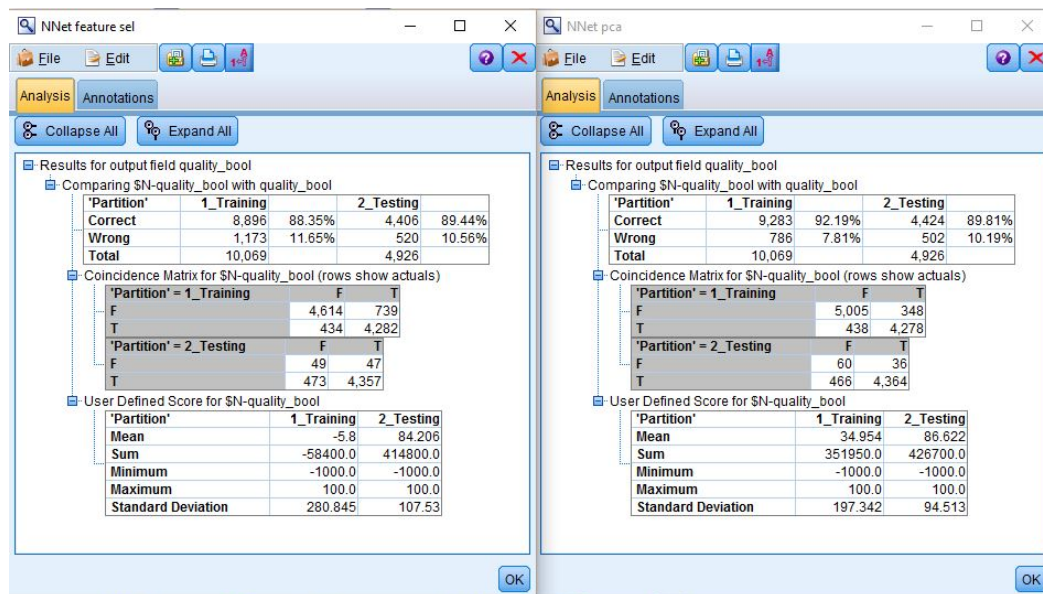
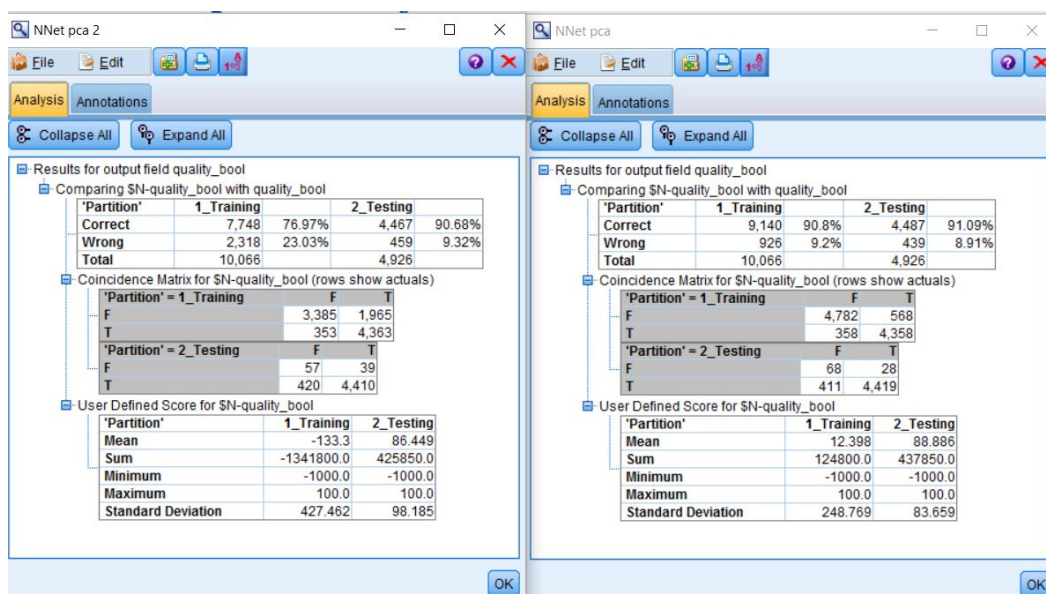
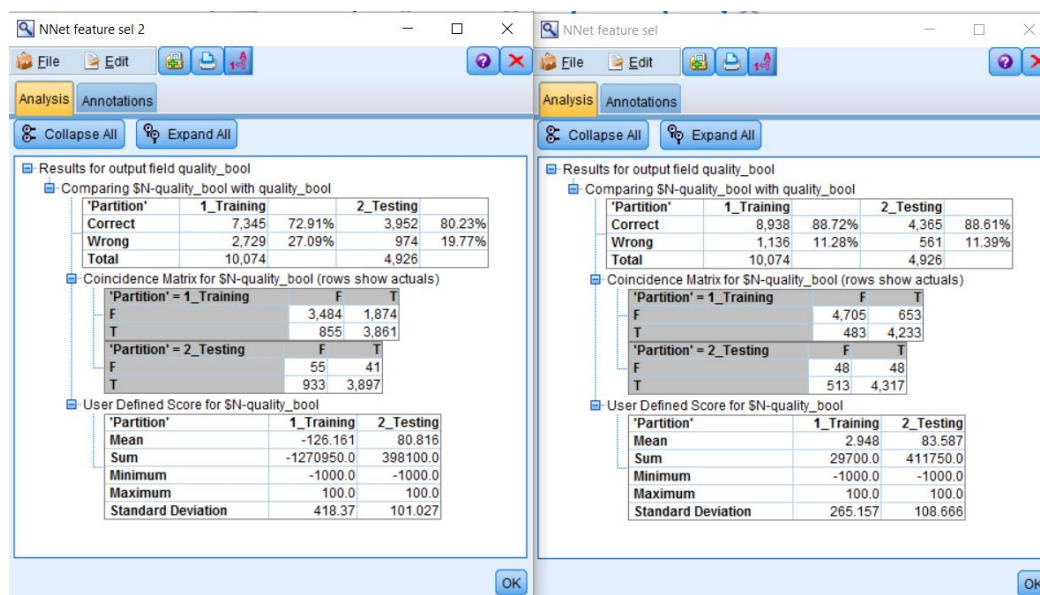


Abbildung 19: Part two of Neural Net settings

Abbildung 20: Neural Net with *Feature Selection* compared to *PCA*Abbildung 21: Logistic regression (unit 1.1 to 10.10) with *PCA*


Abbildung 22: Logistic regression (unit 1.1 to 10.10) with *Feature Selection*

1.4 What are the results if the Logistic regression is modified ?

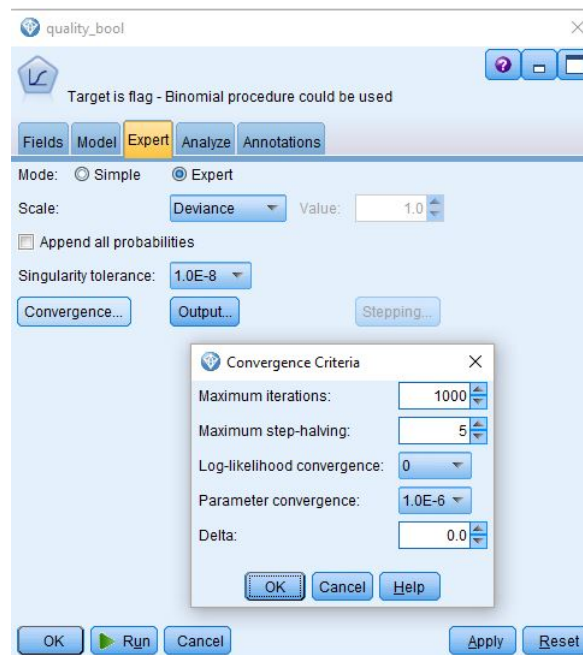
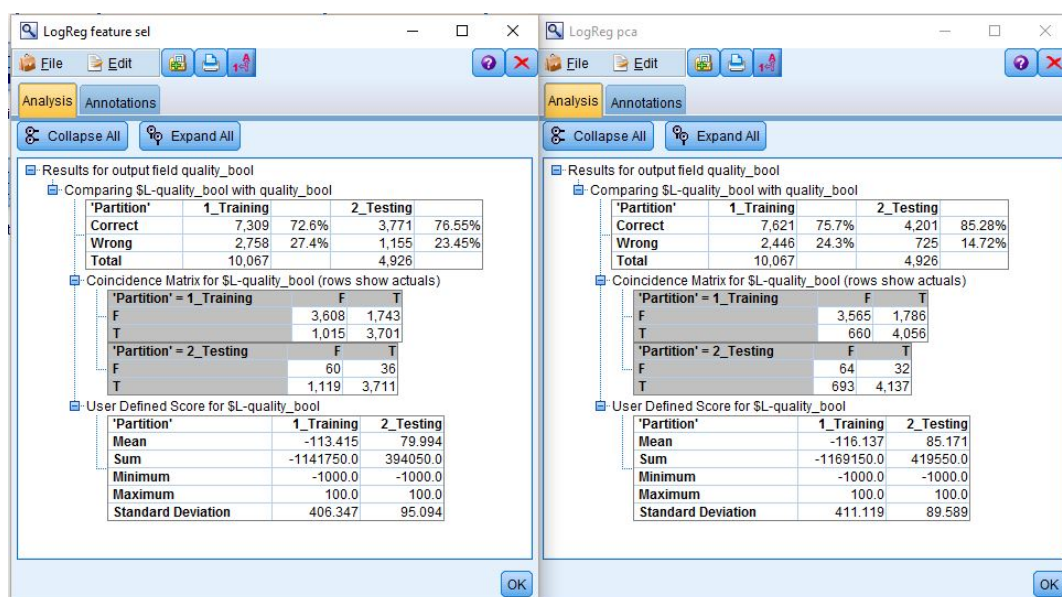


Abbildung 23: Logistic regression (unit 1.1 to 10.10) with *Feature Selection* compared to *PCA*



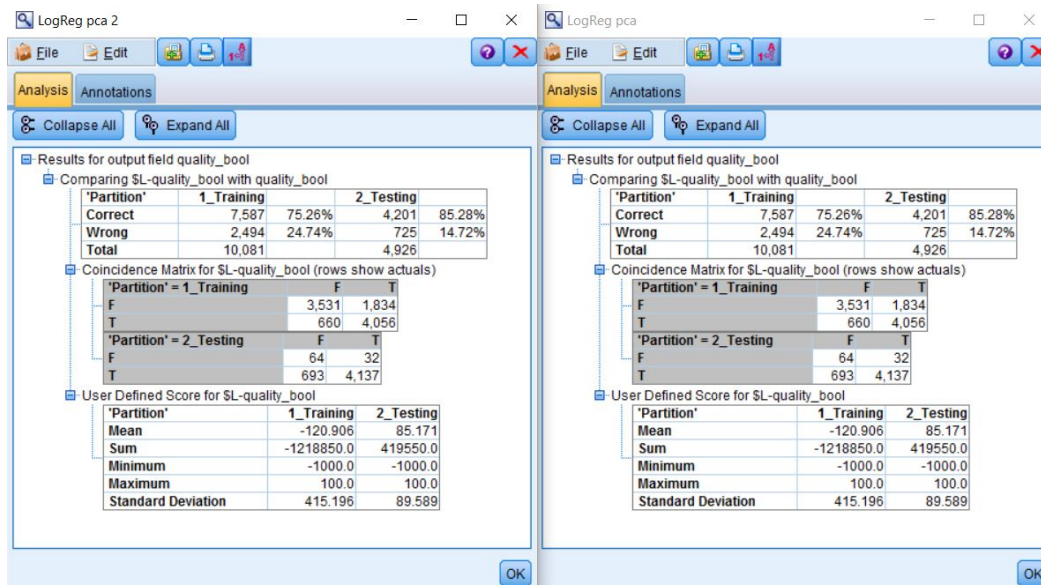
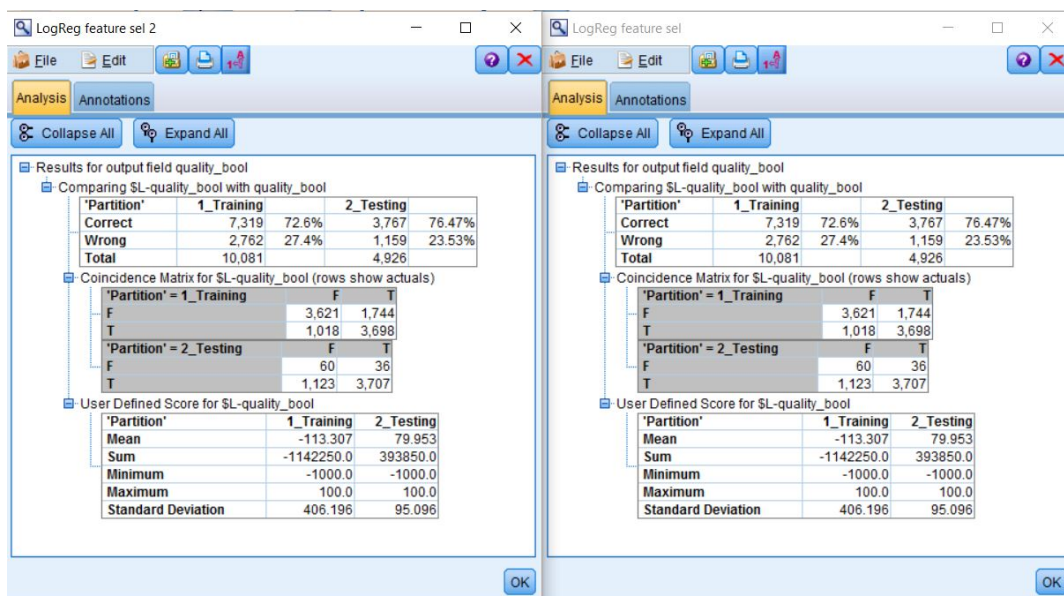
LogReg feature sel Results:

'Partition'	1_Training	2_Testing
Correct	7,309 72.6%	3,771 76.55%
Wrong	2,758 27.4%	1,155 23.45%
Total	10,067	4,926

LogReg pca Results:

'Partition'	1_Training	2_Testing
Correct	7,621 75.7%	4,201 85.28%
Wrong	2,446 24.3%	725 14.72%
Total	10,067	4,926

Abbildung 24: Logistic Regression settings


Abbildung 25: Logistic regression (it=200, steps=200) with *PCA*

Abbildung 26: Logistic regression (it=200, steps=200) with *Feature Selection*

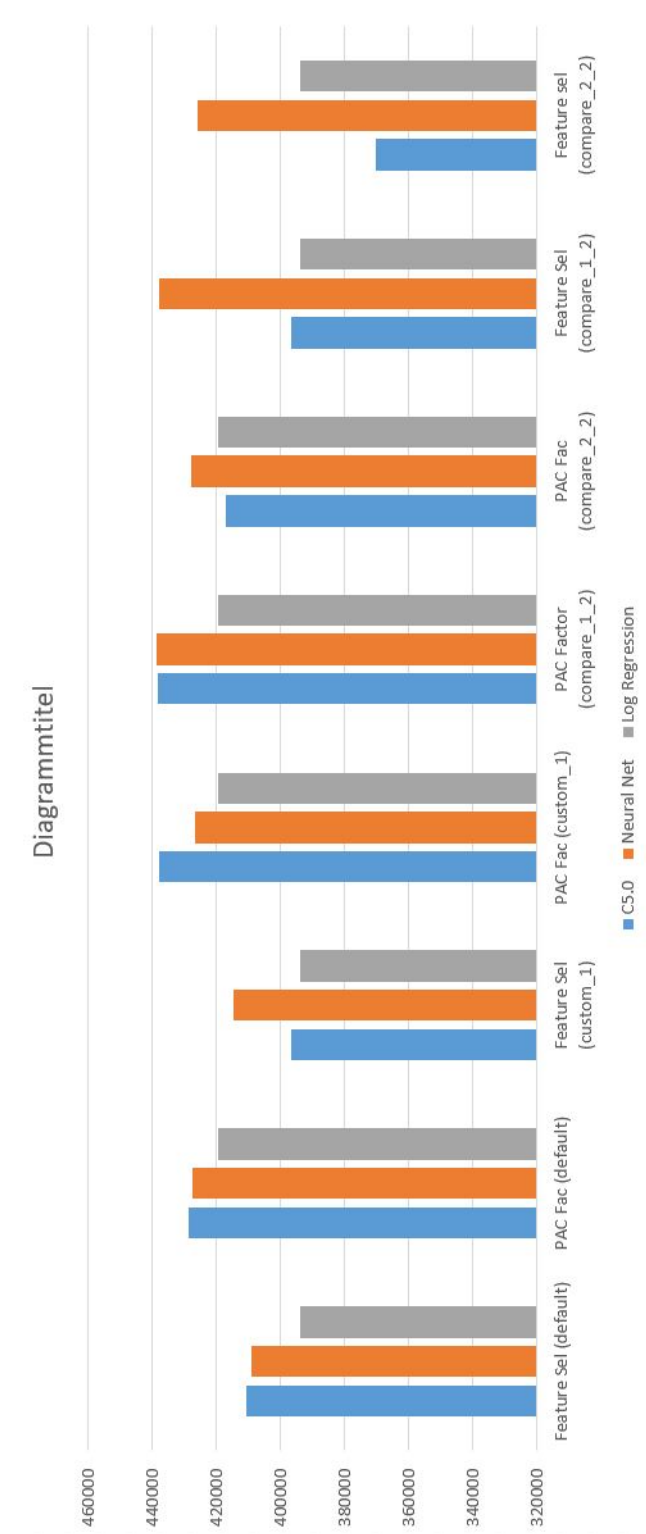


Abbildung 27: All results of the experiments in a table