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

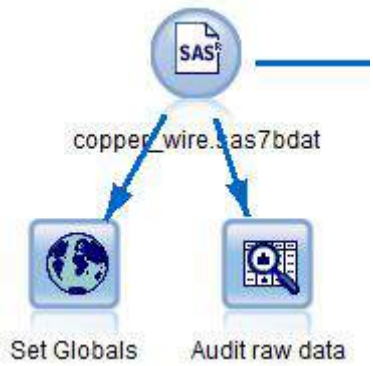


Abbildung 1: Data import in Stream

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.

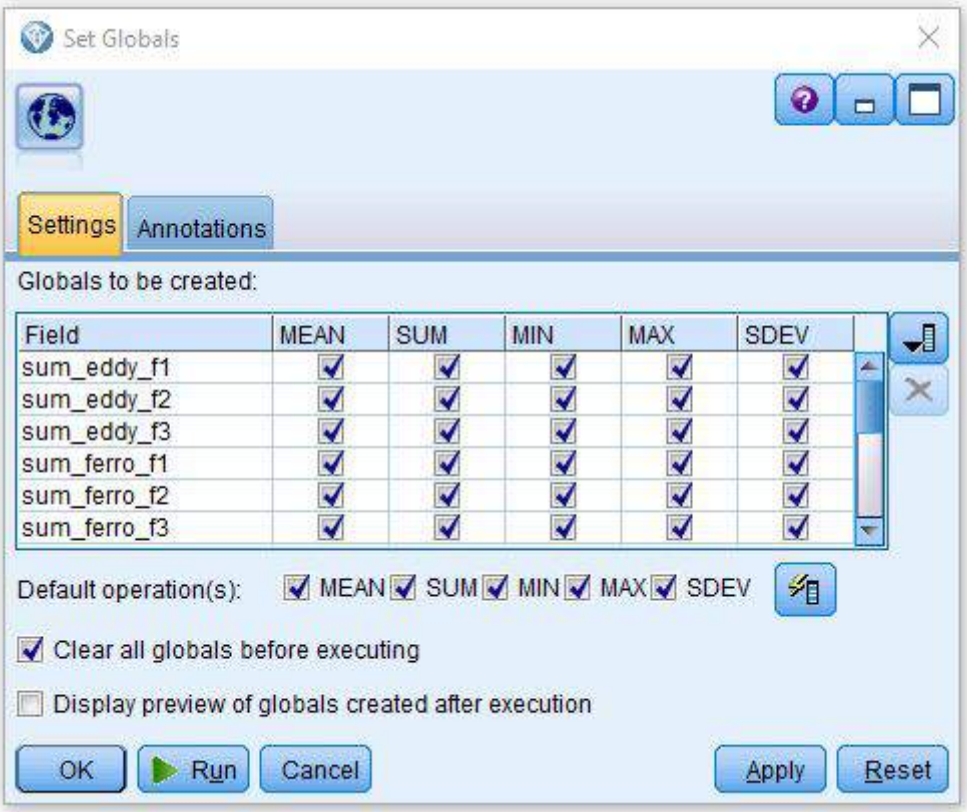
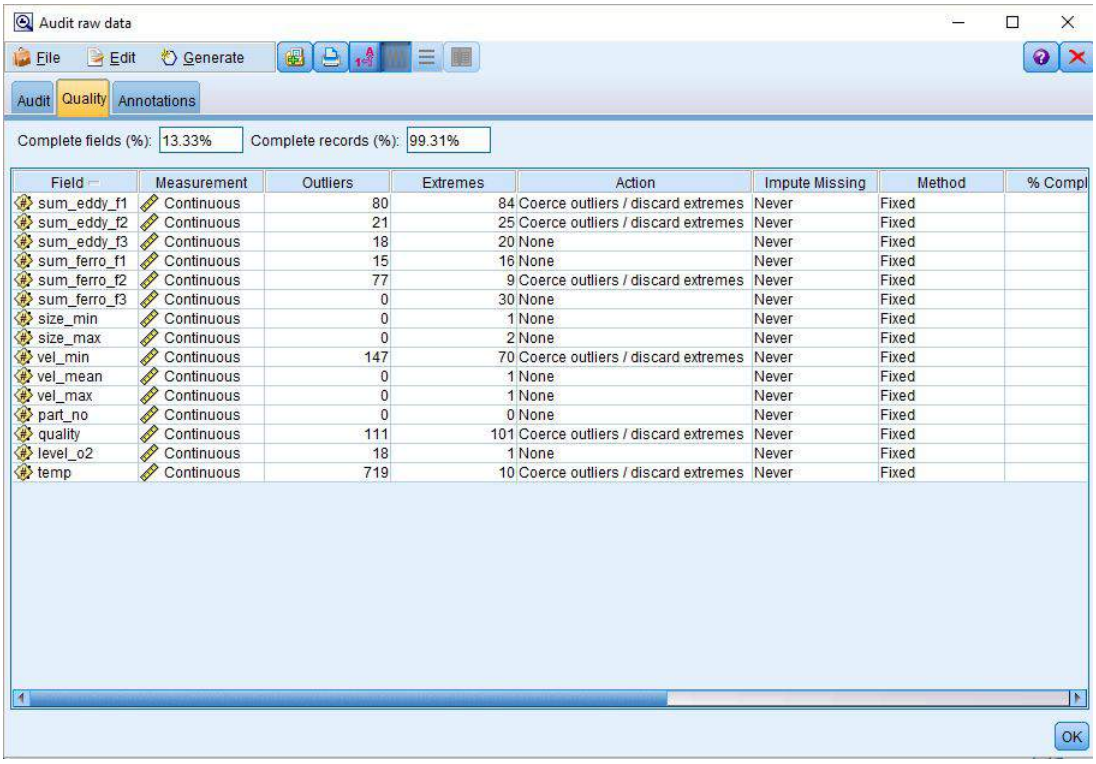


Abbildung 2: Set audit results as global values in the stream

1.2 Data preparation

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



Audit raw data

File Edit Generate

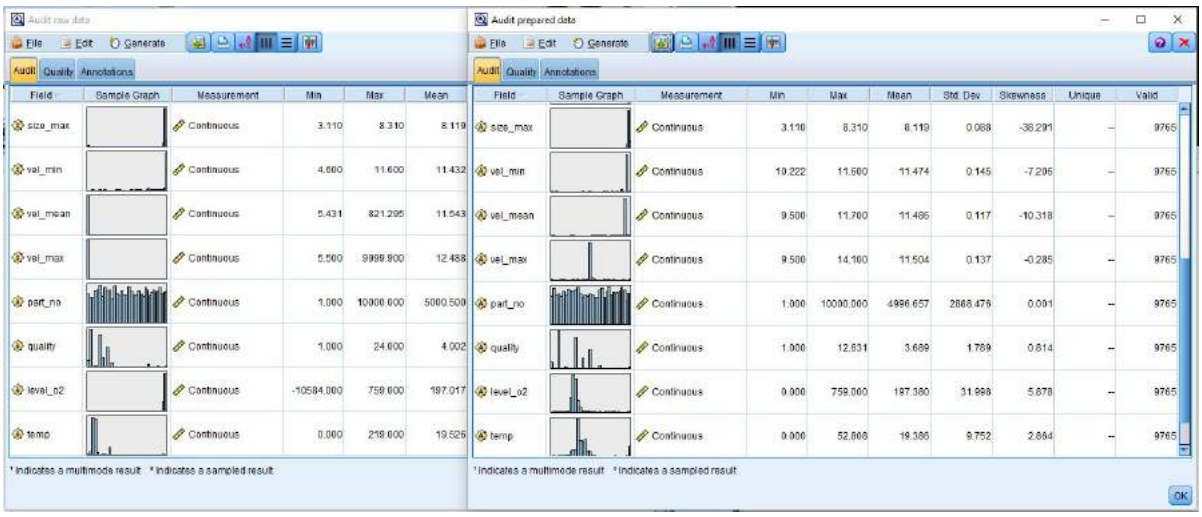
Audit Quality Annotations

Complete fields (%): 13.33% Complete records (%): 99.31%

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	

OK

Abbildung 3: Audit of the raw data



Audit raw data

Field	Sample Graph	Measurement	Min	Max	Mean
size_max		Continuous	3.110	8.310	8.119
vel_min		Continuous	4.600	11.600	11.432
vel_mean		Continuous	5.431	821.295	11.543
vel_max		Continuous	5.500	9999.900	12.488
part_no		Continuous	1.000	10000.000	5000.500
quality		Continuous	1.000	24.000	4.002
level_o2		Continuous	-10584.000	759.000	197.017
temp		Continuous	0.000	219.000	19.526

* Indicates a multimode result * Indicates a sampled result

Audit prepared data

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	10.222	11.600	11.474	0.145	-7.206	–	9765
vel_mean		Continuous	9.500	11.700	11.486	0.117	-10.318	–	9765
vel_max		Continuous	9.500	14.100	11.504	0.137	-0.285	–	9765
part_no		Continuous	1.000	10000.000	4996.657	2888.476	0.001	–	9765
quality		Continuous	1.000	12.631	3.689	1.789	0.814	–	9765
level_o2		Continuous	0.000	759.000	197.380	31.998	5.878	–	9765
temp		Continuous	0.000	52.808	19.386	9.752	2.864	–	9765

* Indicates a multimode result * Indicates a sampled result

OK

Abbildung 4: Audit of the raw data

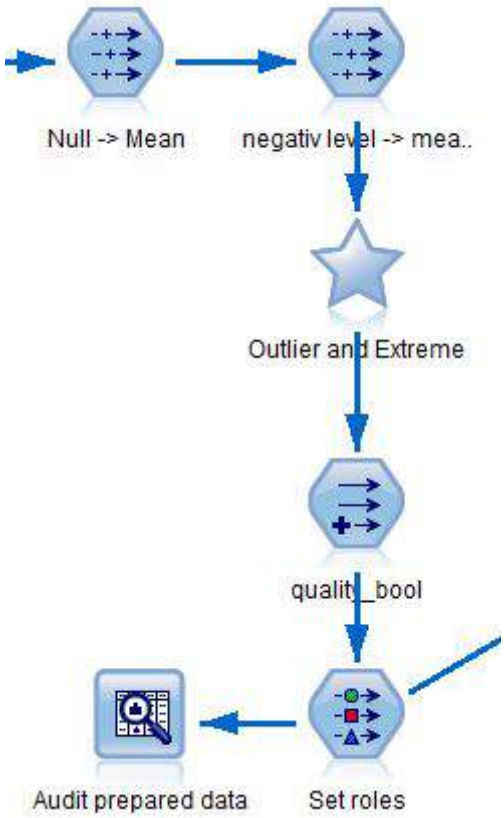


Abbildung 5: Flow of data preparation tasks

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 filed 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

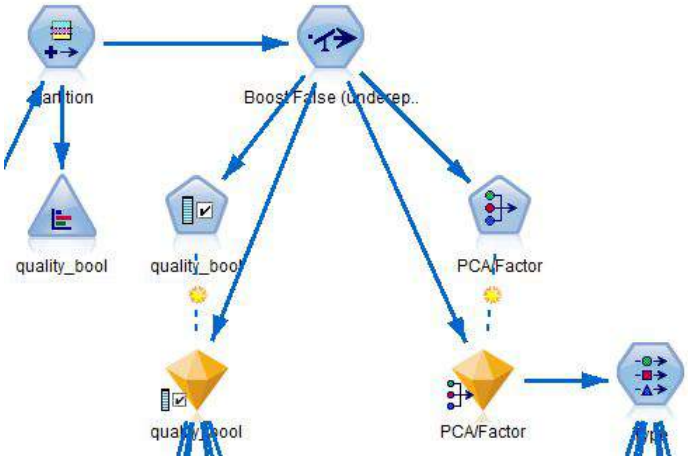


Abbildung 6: Flow of further data preparation

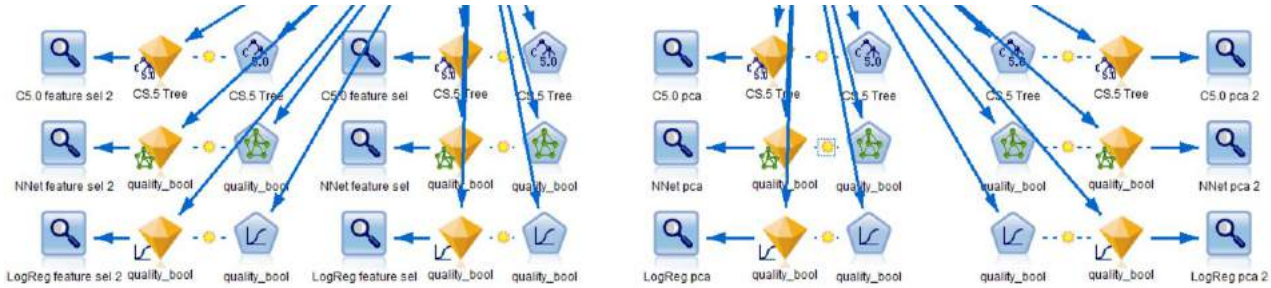


Abbildung 7: Flow of predication models

This part of the stream prepares the data by splitting it into test and training data, additionally the *False qualtity_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*.

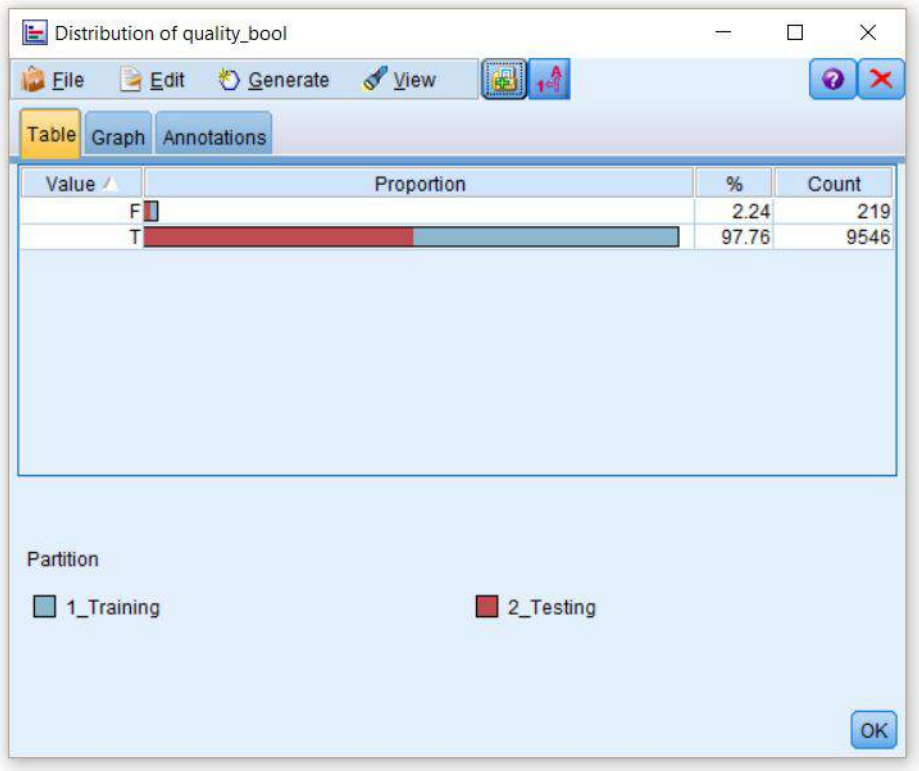


Abbildung 8: Badly distributed *quality_bool*

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

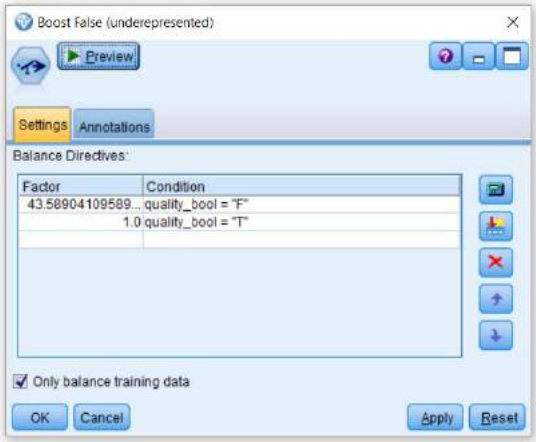
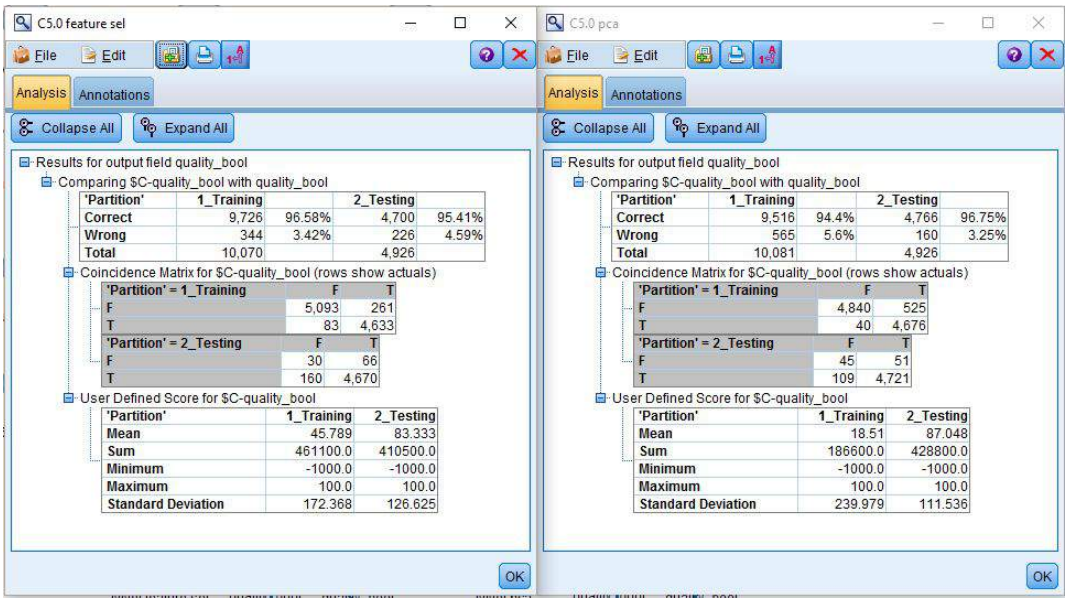


Abbildung 9: Boost of quality *False*

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 ?



C5.0 feature sel

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,726	96.58%	4,700	95.41%
Wrong	344	3.42%	226	4.59%
Total	10,070		4,926	

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

'Partition' = 1_Training	F	T
F	5,093	261
T	83	4,633

'Partition' = 2_Testing	F	T
F	30	66
T	160	4,670

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	45.789	83.333
Sum	461100.0	410500.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	172.368	126.625

C5.0 pca

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,516	94.4%	4,766	96.75%
Wrong	565	5.6%	160	3.25%
Total	10,081		4,926	

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

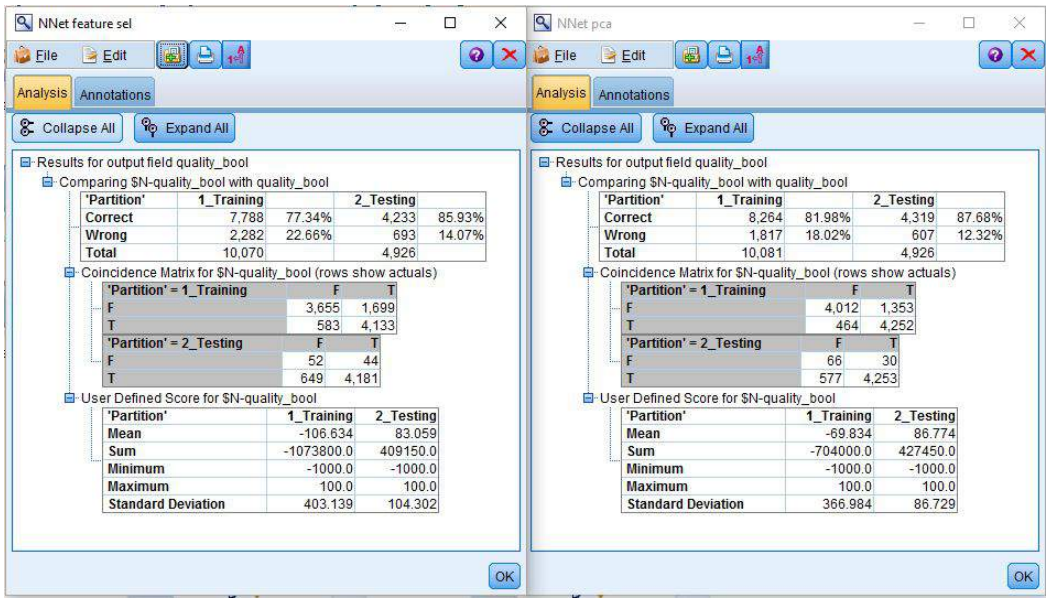
'Partition' = 1_Training	F	T
F	4,840	525
T	40	4,676

'Partition' = 2_Testing	F	T
F	45	51
T	109	4,721

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	18.51	87.048
Sum	186600.0	428800.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	239.979	111.536

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



NNet feature sel

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,788	77.34%	4,233	85.93%
Wrong	2,282	22.66%	693	14.07%
Total	10,070		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,655	1,699
T	583	4,133

'Partition' = 2_Testing	F	T
F	52	44
T	649	4,181

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-106.634	83.059
Sum	-1073800.0	409150.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	403.139	104.302

NNet pca

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,264	81.98%	4,319	87.68%
Wrong	1,817	18.02%	607	12.32%
Total	10,081		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

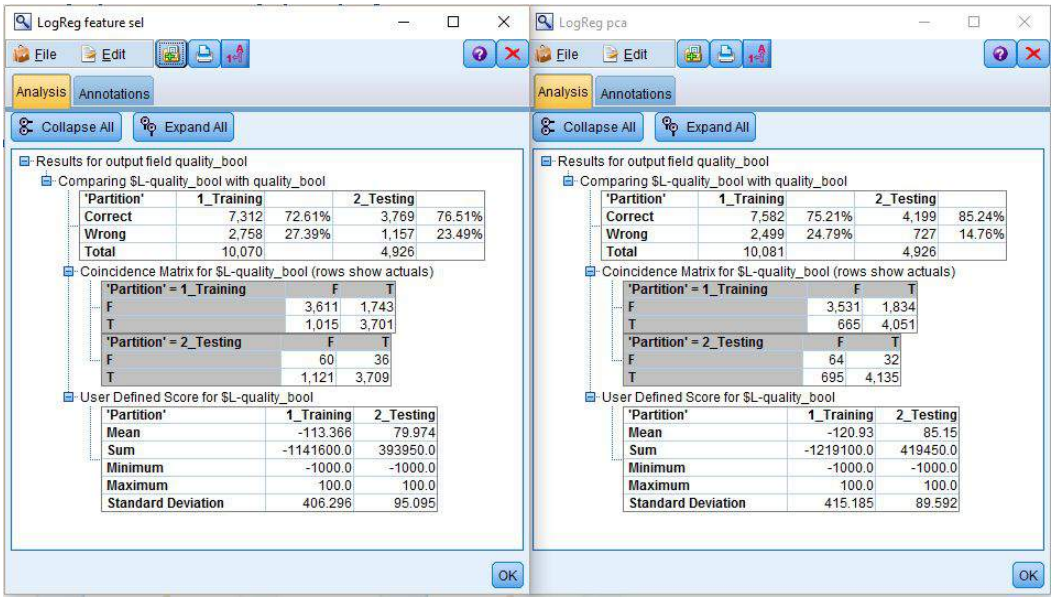
'Partition' = 1_Training	F	T
F	4,012	1,353
T	464	4,252

'Partition' = 2_Testing	F	T
F	66	30
T	577	4,253

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-69.834	86.774
Sum	-704000.0	427450.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	366.984	86.729

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



LogReg feature sel

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,312	72.61%	3,769	76.51%
Wrong	2,758	27.39%	1,157	23.49%
Total	10,070		4,926	

Coincidence Matrix for $L-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,611	1,743
T	1,015	3,701

'Partition' = 2_Testing	F	T
F	60	36
T	1,121	3,709

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-113.366	79.974
Sum	-1141600.0	393950.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	406.296	95.095

LogReg pca

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,582	75.21%	4,199	85.24%
Wrong	2,499	24.79%	727	14.76%
Total	10,081		4,926	

Coincidence Matrix for $L-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,531	1,834
T	665	4,051

'Partition' = 2_Testing	F	T
F	64	32
T	695	4,135

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-120.93	85.15
Sum	-1219100.0	419450.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	415.185	89.592

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

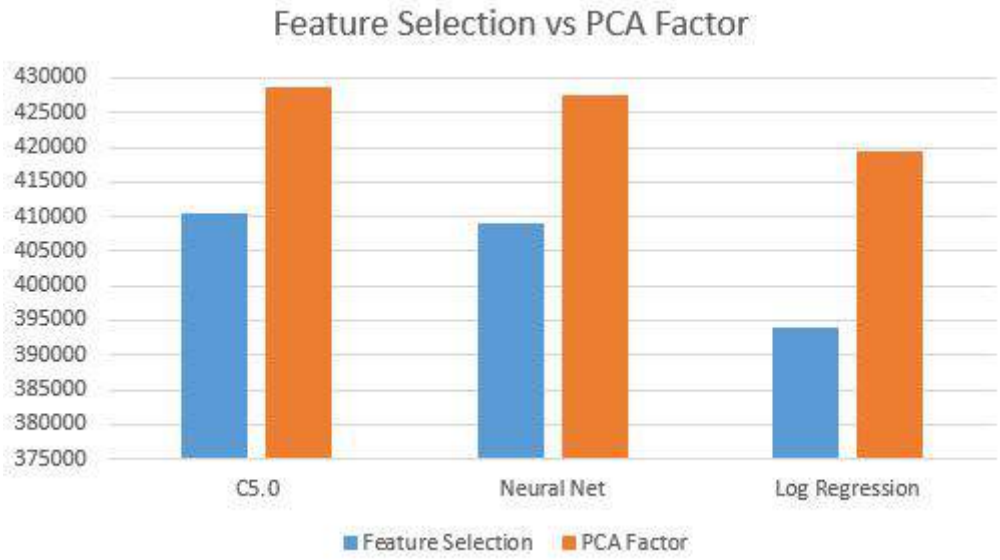
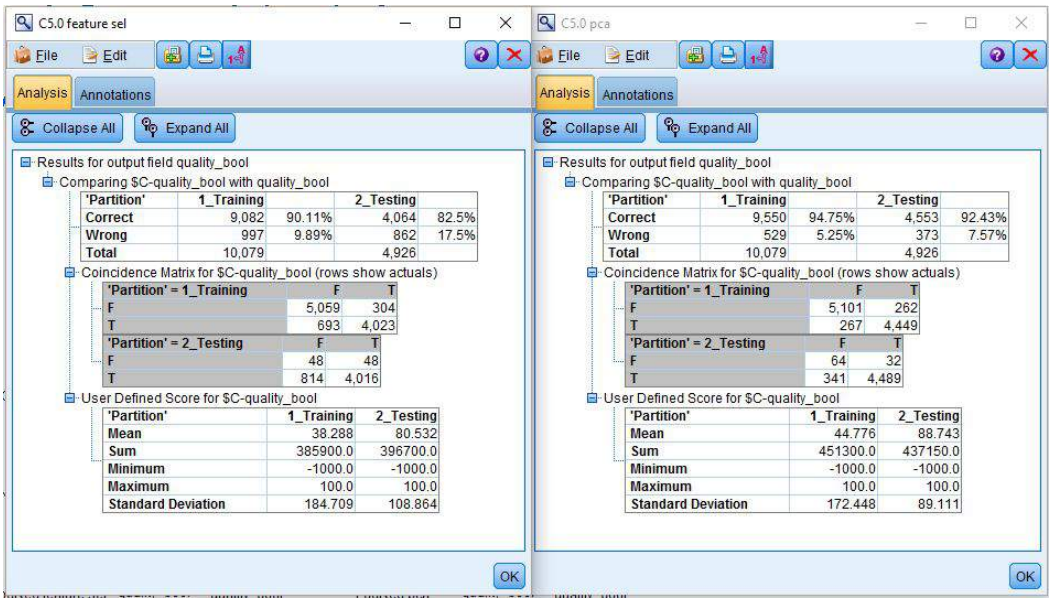


Abbildung 13: The chart shows the results of the comparison between *Feature Selection* and *PCA*

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	90.11%	4,064	82.5%
Wrong	997	9.89%	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	94.75%	4,553	92.43%
Wrong	529	5.25%	373	7.57%
Total	10,079		4,926	

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

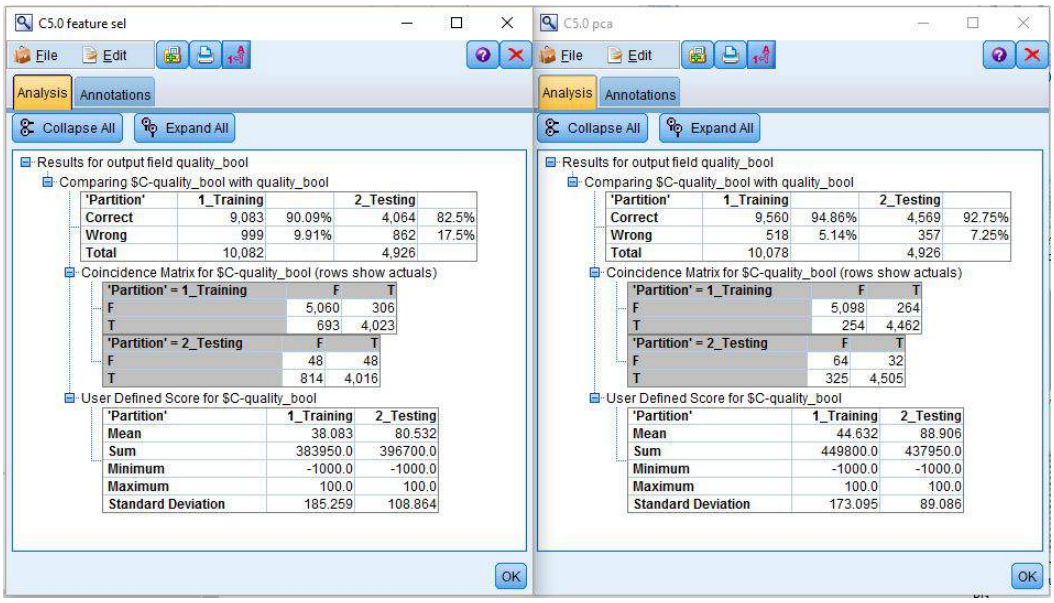
'Partition' = 1_Training	F	T
F	5,101	262
T	267	4,449

'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	90.09%	4,064	82.5%
Wrong	999	9.91%	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	94.86%	4,569	92.75%
Wrong	518	5.14%	357	7.25%
Total	10,078		4,926	

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

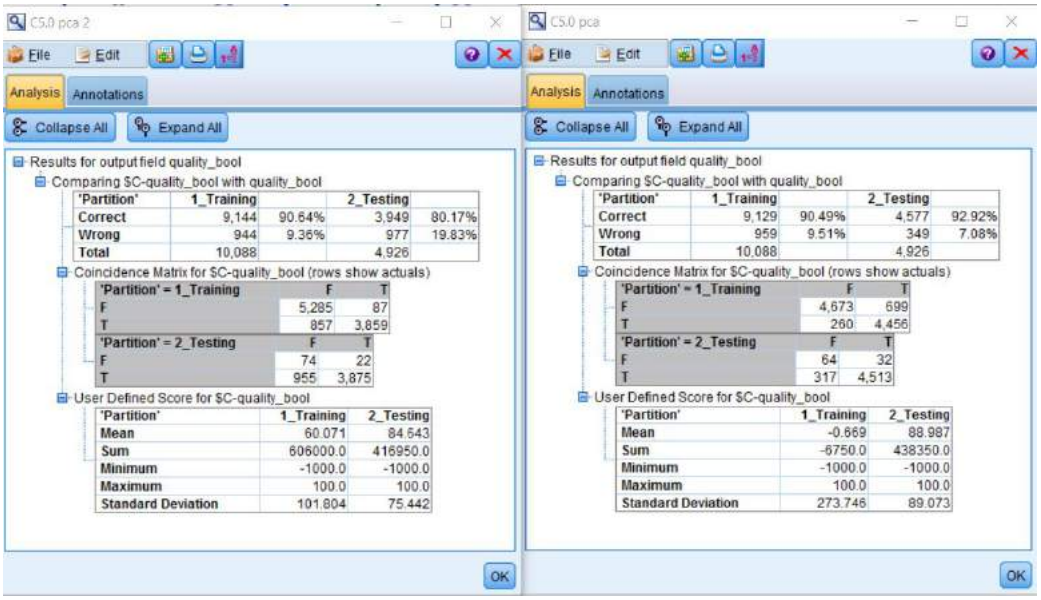
'Partition' = 1_Training	F	T
F	5,098	264
T	254	4,462

'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*



C5.0 pca 2

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,144	90.64%	3,949	80.17%
Wrong	944	9.36%	977	19.83%
Total	10,088		4,926	

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

'Partition' = 1_Training	F	T
F	5,285	87
T	857	3,859

'Partition' = 2_Testing	F	T
F	74	22
T	955	3,875

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	60.071	84.543
Sum	606000.0	416950.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	101.804	75.442

C5.0 pca

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,129	90.49%	4,577	92.92%
Wrong	959	9.51%	349	7.08%
Total	10,088		4,926	

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

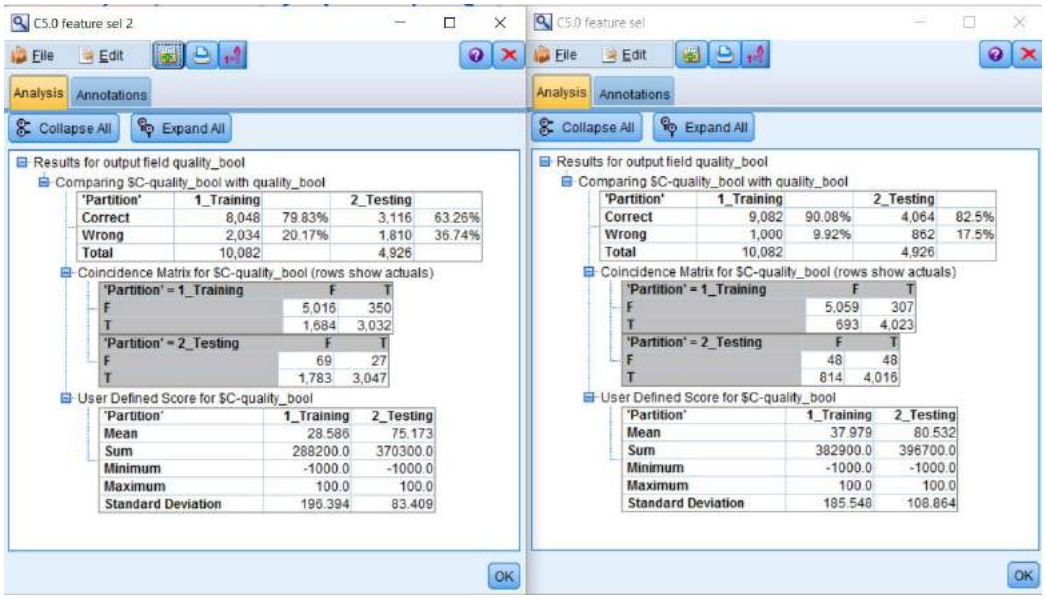
'Partition' = 1_Training	F	T
F	4,673	699
T	260	4,456

'Partition' = 2_Testing	F	T
F	64	32
T	317	4,513

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	-0.669	88.987
Sum	-6750.0	438350.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	273.746	89.073

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



C5.0 feature sel 2

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,048	79.83%	3,116	63.26%
Wrong	2,034	20.17%	1,810	36.74%
Total	10,082		4,926	

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

'Partition' = 1_Training	F	T
F	5,016	350
T	1,684	3,032

'Partition' = 2_Testing	F	T
F	69	27
T	1,783	3,047

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	28.586	75.173
Sum	288200.0	370300.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	196.394	83.409

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	90.08%	4,064	82.5%
Wrong	1,000	9.92%	862	17.5%
Total	10,082		4,926	

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

'Partition' = 1_Training	F	T
F	5,059	307
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	37.979	80.532
Sum	382900.0	396700.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	185.548	108.864

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 ?

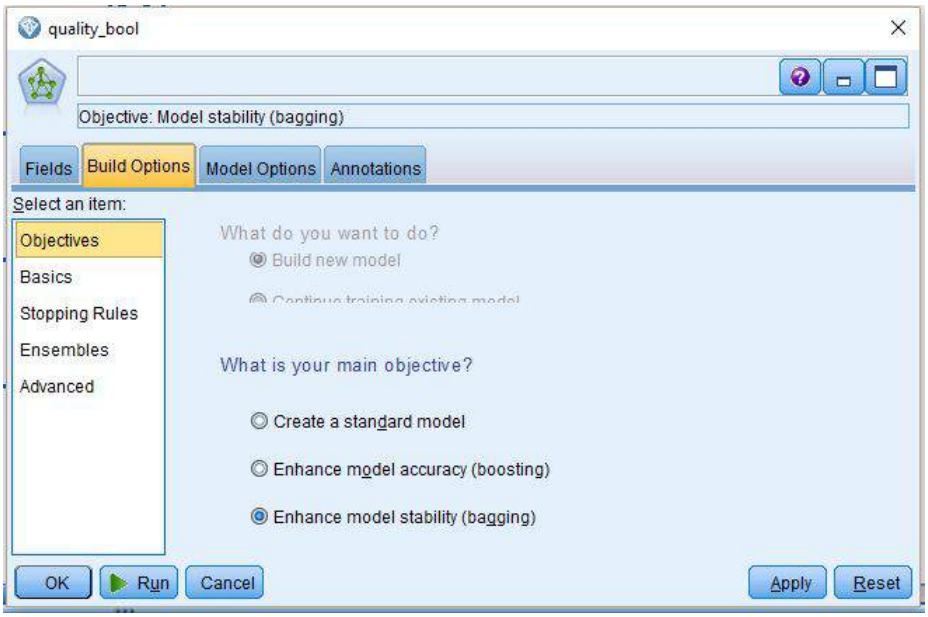


Abbildung 18: Part one of Neural Net settings

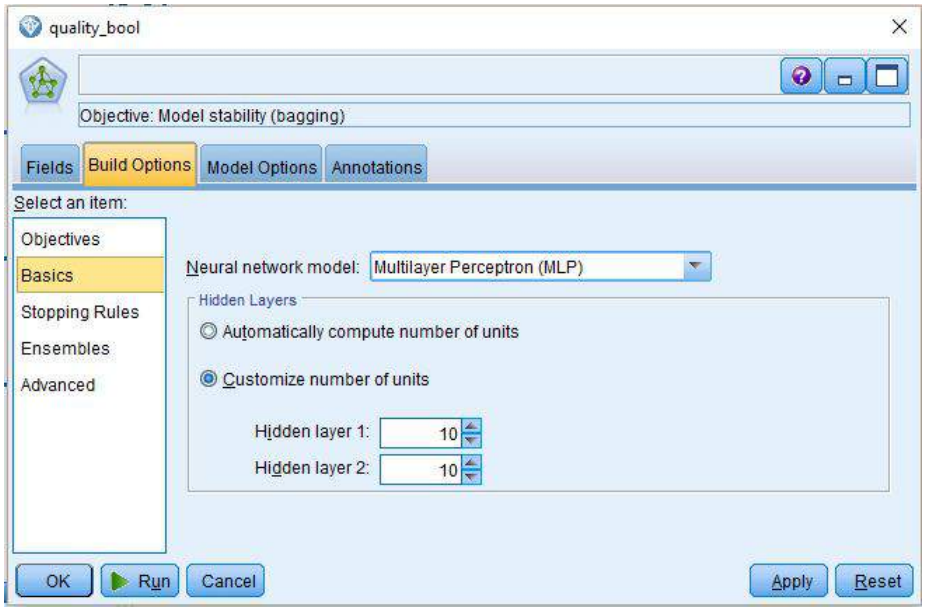
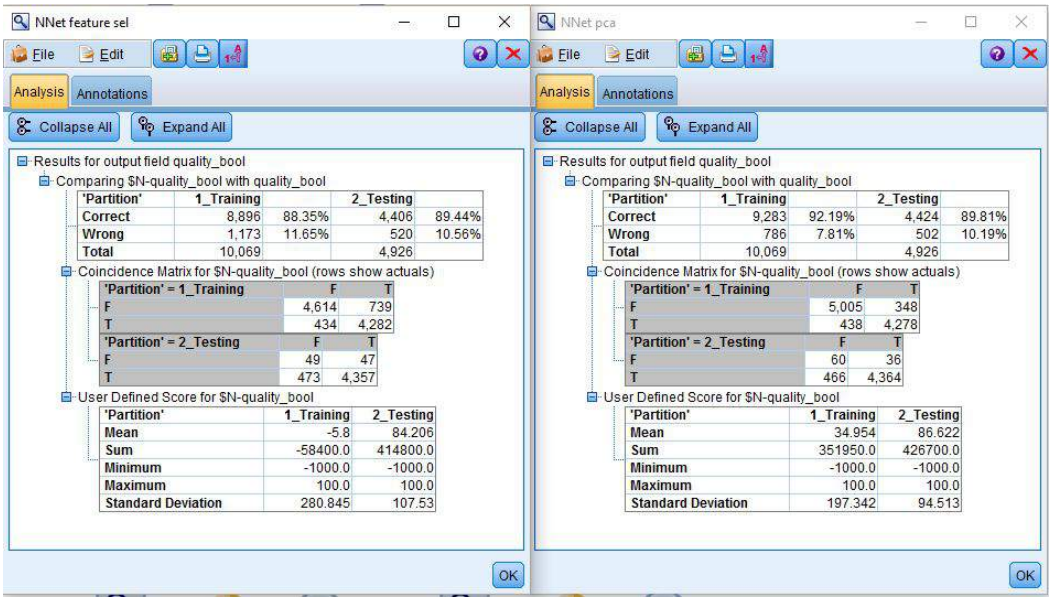


Abbildung 19: Part two of Neural Net settings



NNet feature sel

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,896	88.35%	4,406	89.44%
Wrong	1,173	11.65%	520	10.56%
Total	10,069		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	4,614	739
T	434	4,282

'Partition' = 2_Testing	F	T
F	49	47
T	473	4,357

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-5.8	84.206
Sum	-58400.0	414800.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	280.845	107.53

NNet pca

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,283	92.19%	4,424	89.81%
Wrong	786	7.81%	502	10.19%
Total	10,069		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

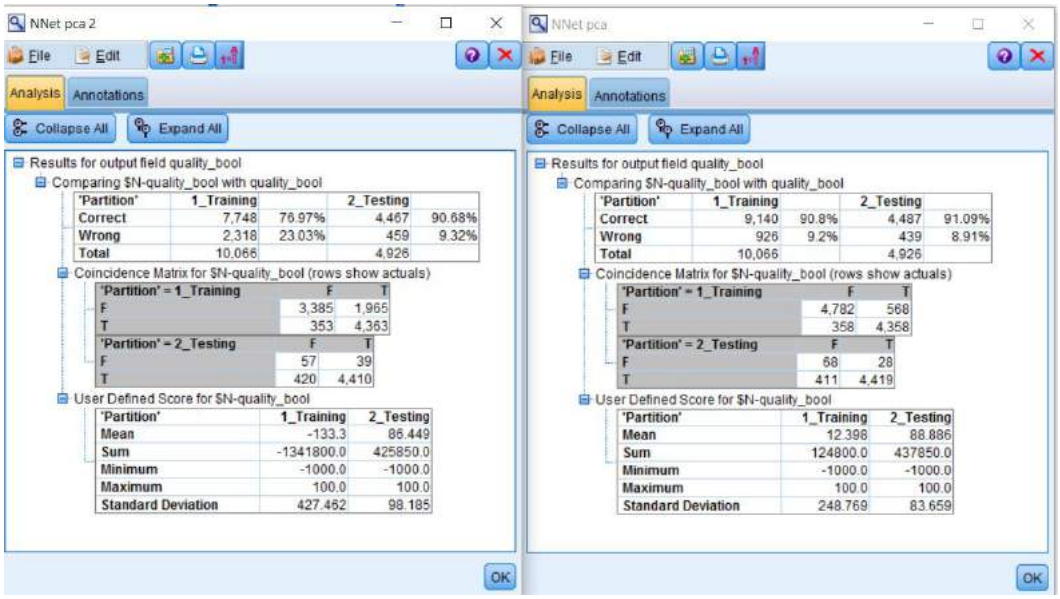
'Partition' = 1_Training	F	T
F	5,005	348
T	438	4,278

'Partition' = 2_Testing	F	T
F	60	36
T	466	4,364

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	34.954	86.622
Sum	351950.0	426700.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	197.342	94.513

Abbildung 20: Neural Net with *Feature Selection* compared to *PCA*



NNet pca 2

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,748	76.97%	4,467	90.68%
Wrong	2,318	23.03%	459	9.32%
Total	10,066		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,385	1,965
T	353	4,363

'Partition' = 2_Testing	F	T
F	57	39
T	420	4,410

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-133.3	86.449
Sum	-1341800.0	425850.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	427.462	98.185

NNet pca

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,140	90.8%	4,487	91.09%
Wrong	926	9.2%	439	8.91%
Total	10,066		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

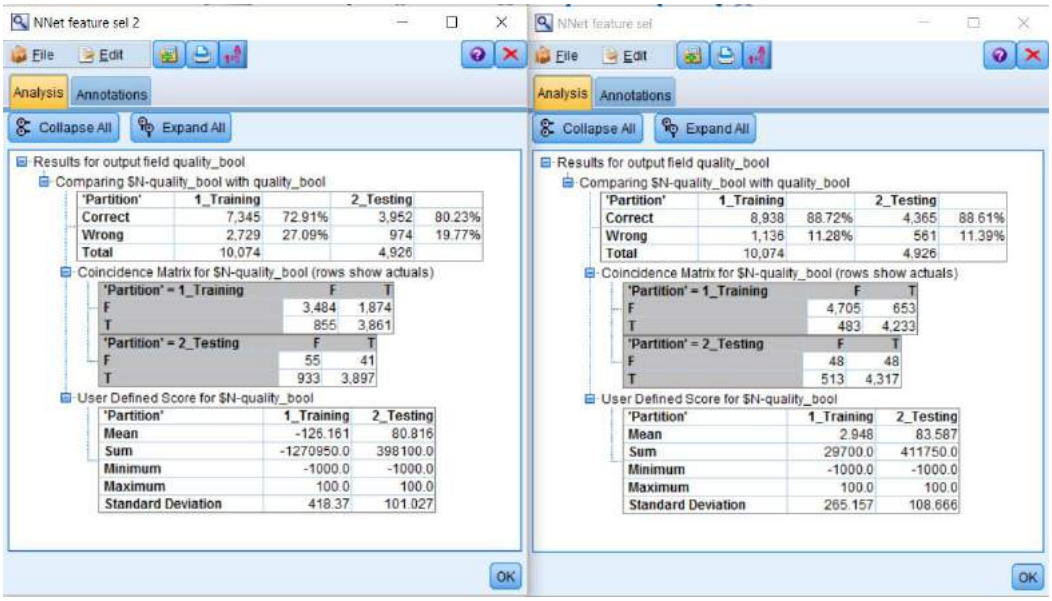
'Partition' = 1_Training	F	T
F	4,782	568
T	358	4,358

'Partition' = 2_Testing	F	T
F	68	28
T	411	4,419

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	12.398	88.886
Sum	124800.0	437850.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	248.769	83.659

Abbildung 21: Logistic regression (unit 1.1 to 10.10) with *PCA*



NNet feature sel 2

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,345	72.91%	3,952	80.23%
Wrong	2,729	27.09%	974	19.77%
Total	10,074		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,484	1,874
T	855	3,861

'Partition' = 2_Testing	F	T
F	55	41
T	933	3,897

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-126.161	80.816
Sum	-1270950.0	398100.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	418.37	101.027

OK

NNet feature sel

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,938	88.72%	4,365	88.61%
Wrong	1,136	11.28%	561	11.39%
Total	10,074		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	4,705	653
T	483	4,233

'Partition' = 2_Testing	F	T
F	48	48
T	513	4,317

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	2.948	83.587
Sum	29700.0	411750.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	265.157	108.666

OK

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*

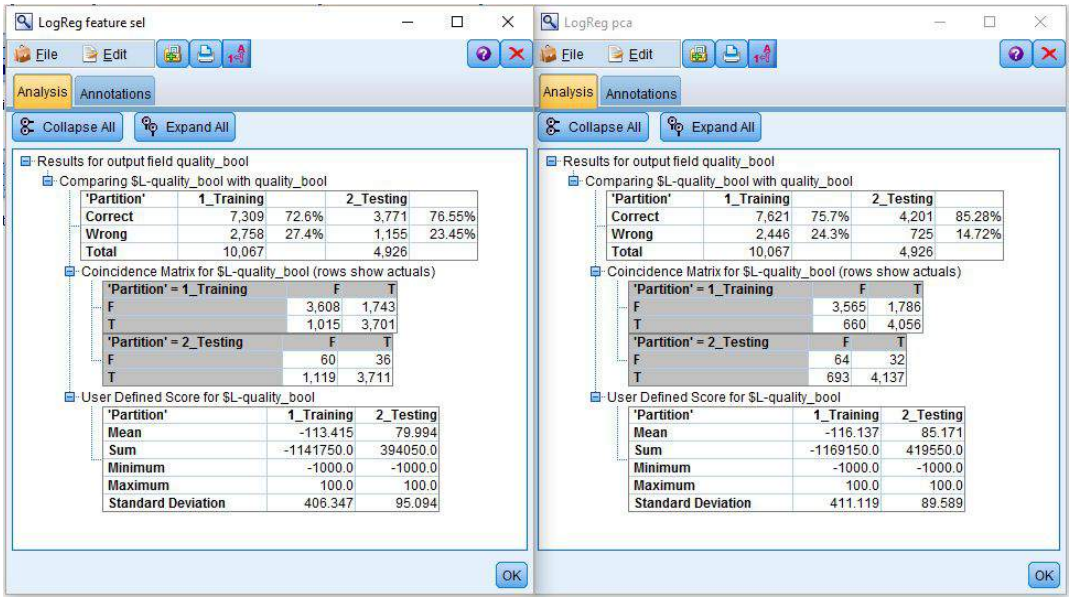
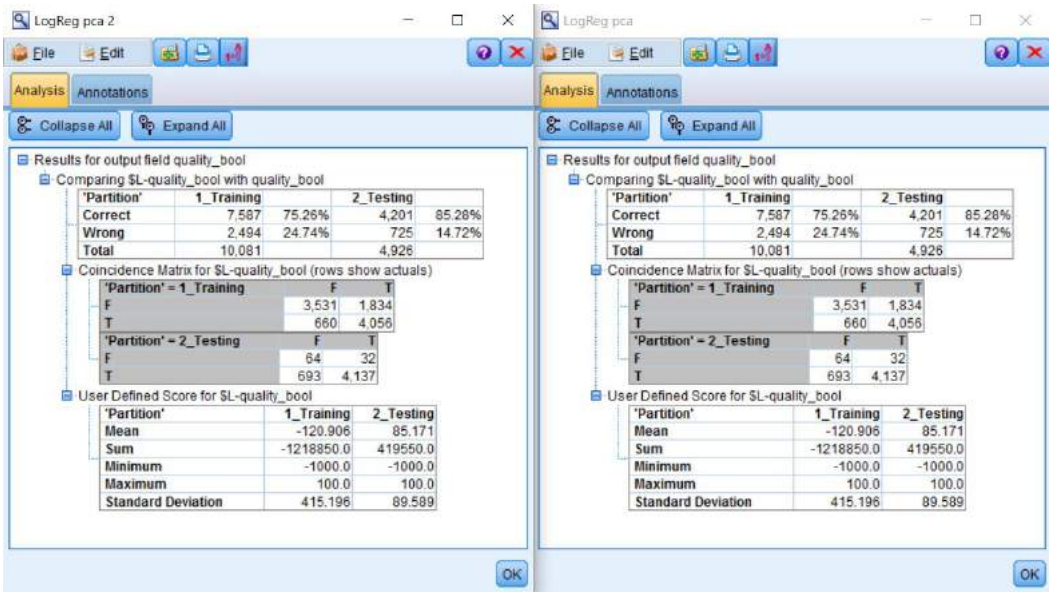


Abbildung 24: Logistic Regression settings



LogReg pca 2

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,587	75.26%	4,201	85.28%
Wrong	2,494	24.74%	725	14.72%
Total	10,081		4,926	

Coincidence Matrix for $L-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,531	1,834
T	660	4,056

'Partition' = 2_Testing	F	T
F	64	32
T	693	4,137

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-120.906	85.171
Sum	-1218850.0	419550.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	415.196	89.589

LogReg pca

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,587	75.26%	4,201	85.28%
Wrong	2,494	24.74%	725	14.72%
Total	10,081		4,926	

Coincidence Matrix for $L-quality_bool (rows show actuals)

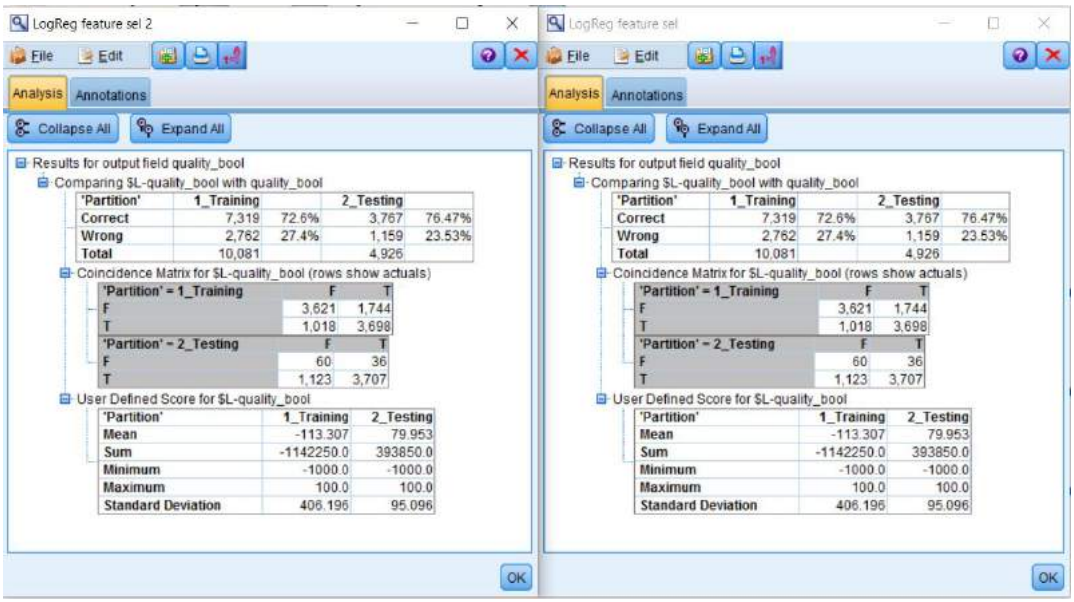
'Partition' = 1_Training	F	T
F	3,531	1,834
T	660	4,056

'Partition' = 2_Testing	F	T
F	64	32
T	693	4,137

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-120.906	85.171
Sum	-1218850.0	419550.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	415.196	89.589

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



LogReg feature sel 2

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,319	72.6%	3,767	76.47%
Wrong	2,762	27.4%	1,159	23.53%
Total	10,081		4,926	

Coincidence Matrix for $L-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,621	1,744
T	1,018	3,698

'Partition' = 2_Testing	F	T
F	60	36
T	1,123	3,707

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-113.307	79.953
Sum	-1142250.0	393850.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	406.196	95.096

LogReg feature sel

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,319	72.6%	3,767	76.47%
Wrong	2,762	27.4%	1,159	23.53%
Total	10,081		4,926	

Coincidence Matrix for $L-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	3,621	1,744
T	1,018	3,698

'Partition' = 2_Testing	F	T
F	60	36
T	1,123	3,707

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-113.307	79.953
Sum	-1142250.0	393850.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	406.196	95.096

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

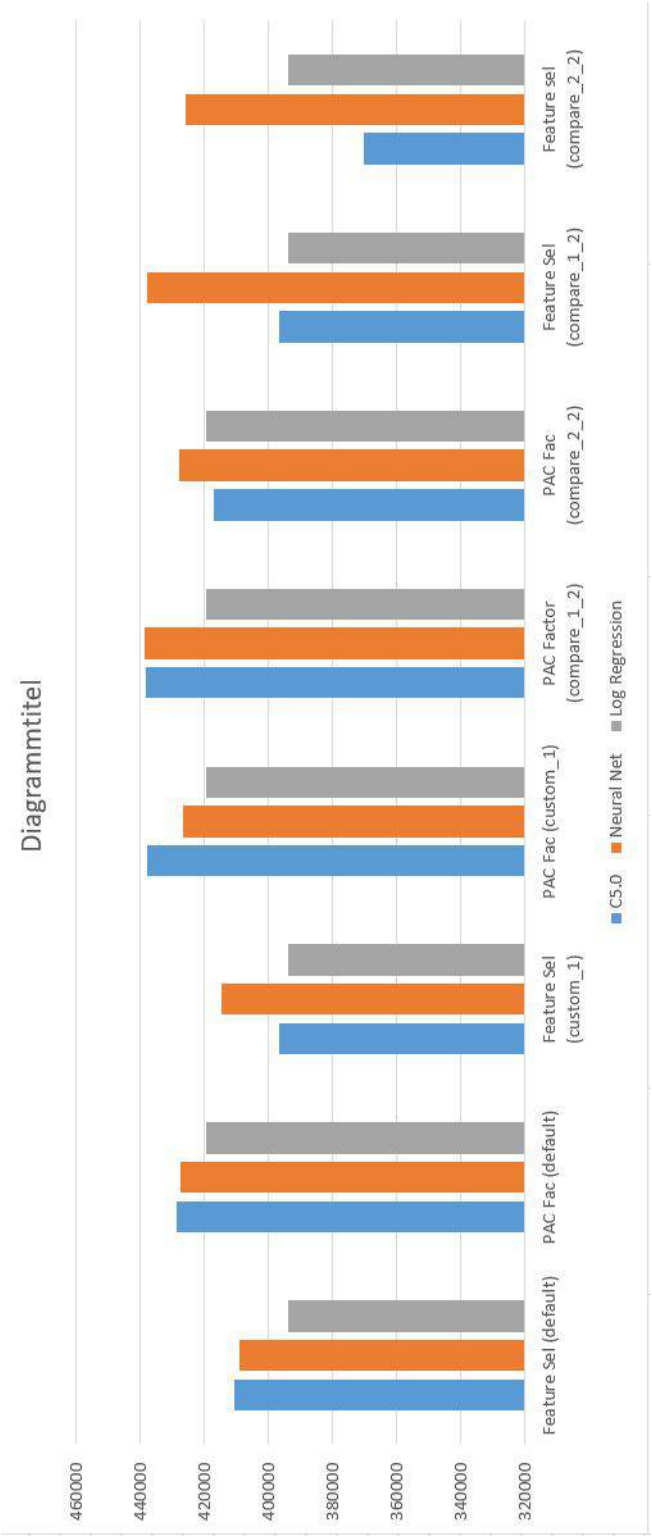


Abbildung 27: All results of the experiments in a table