Florida Atlantic University
COLLEGE OF BUSINESS

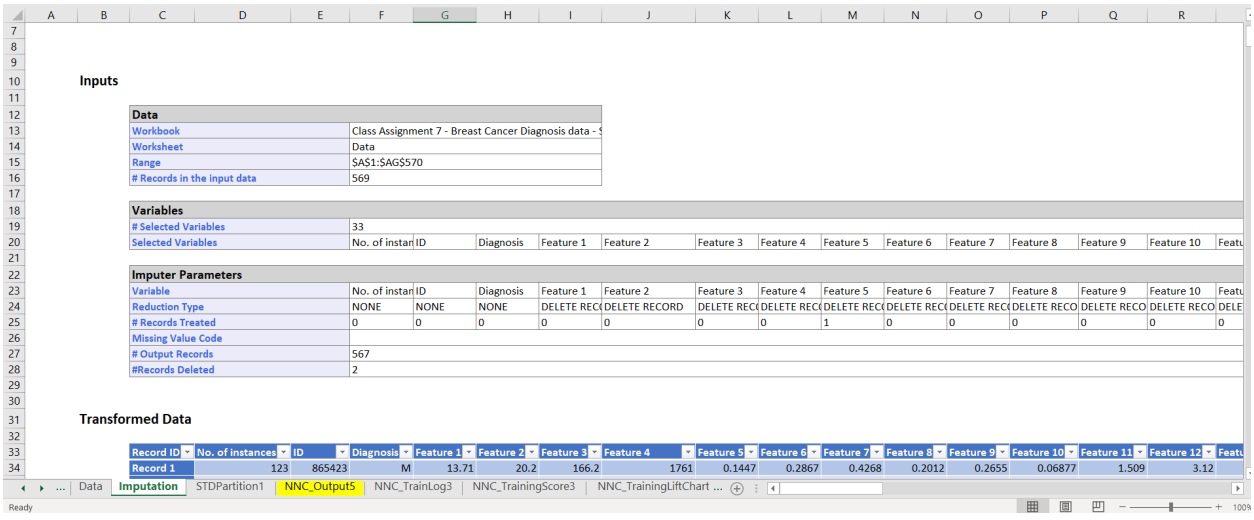
ISM 6136 – Datamining/Predictive Analytics

Jacob Perrone
Class Assignment 7
5 points

TASK: Performing predictive analytics using Neural Nets/ANN in XLMiner OR RapidMiner

Perform the following data mining steps (CLASSIFICATION) in XLMiner

1. Follow the datamining steps below:
   a) Understand the problem and purpose of data mining task
   b) Obtain the dataset for analysis – **Breast Cancer Diagnosis.xls**
   c) Explore, clean and preprocess data
      i. Cleanup any column that is not a predictor
      ii. Perform 'Missing Data Handling'
      iii. Any categorical variables conversion needed – check and remember to perform during modeling



Inputs

Data	
Workbook	Class Assignment 7 - Breast Cancer Diagnosis data - S
Worksheet	Data
Range	$A$1:$AG$570
# Records in the input data	569

Variables	
# Selected Variables	33
Selected Variables	No. of instar, ID, Diagnosis, Feature 1, Feature 2, Feature 3, Feature 4, Feature 5, Feature 6, Feature 7, Feature 8, Feature 9, Feature 10, Featu

Imputer Parameters														
Variable	No. of instar	ID	Diagnosis	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Featu
Reduction Type	NONE	NONE	NONE	DELETE REC	DELETE RECORD	DELETE REC	DELETE REC	DELETE REC	DELETE REC	DELETE REC	DELETE RECO	DELETE RECO	DELETE RECO	DELE
# Records Treated	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Missing Value Code														
# Output Records	567													
#Records Deleted	2													

Transformed Data

Record ID	No. of instances	ID	Diagnosis	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11	Feature 12	Featu
Record 1	123	865423	M	13.71	20.2	166.2	1761	0.1447	0.2867	0.4268	0.2012	0.2655	0.06877	1.509	3.12	

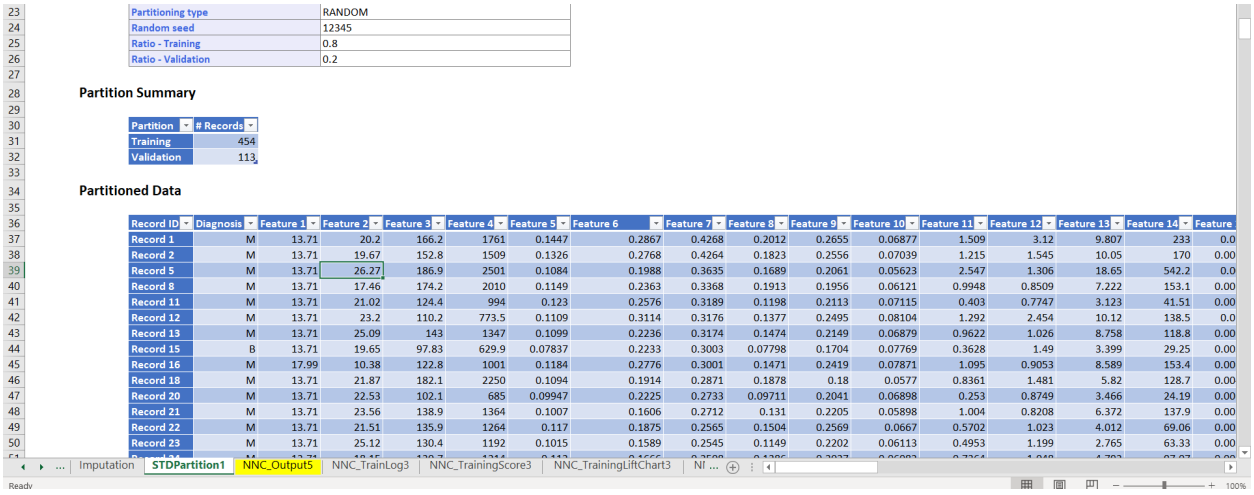
Data | Imputation | STDPartition1 | NNC_Output5 | NNC_TrainLog3 | NNC_TrainingScore3 | NNC_TrainingLiftChart ...

Ready

For this dataset, all the features were numerical, and only 2 records had missing values. The instances with missing values were deleted since there was so few of them.

d) Reduction of data dimension (if needed to get another model)
Not done for this assignment, but could have done PCA to find the best features and only use those for the NN instead of using all 30 features.

e) Partition data



Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.8
Ratio - Validation	0.2

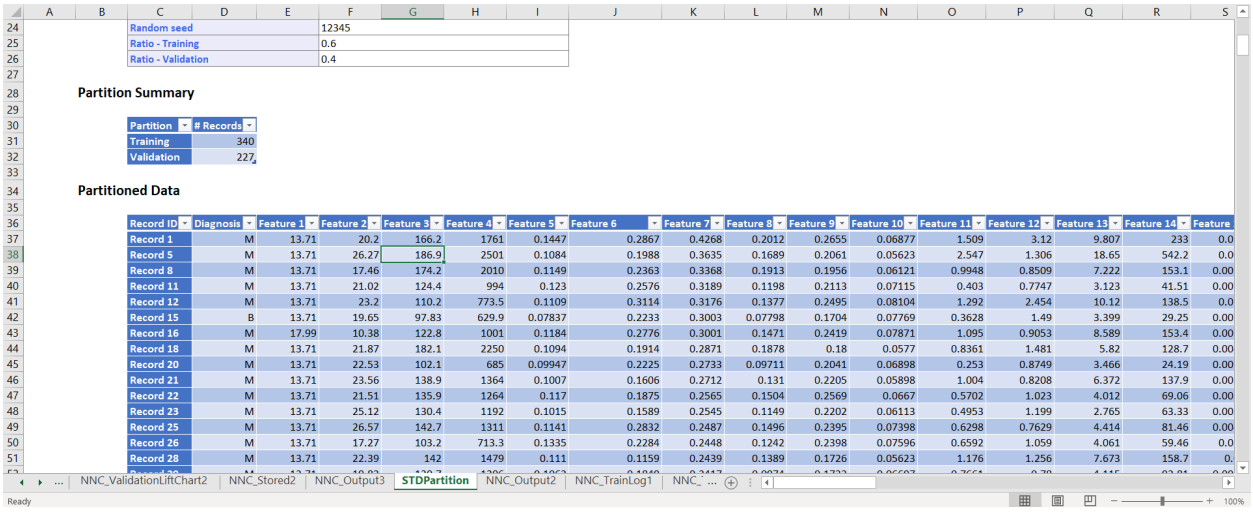
Partition Summary

Partition	# Records
Training	454
Validation	113

Partitioned Data

Record ID	Diagnosis	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11	Feature 12	Feature 13	Feature 14	Feature
Record 1	M	13.71	20.2	166.2	1761	0.1447	0.2867	0.4268	0.2012	0.2655	0.06877	1.509	3.12	9.807	233	0.0
Record 2	M	13.71	19.67	152.8	1509	0.1326	0.2768	0.4264	0.1823	0.2556	0.07039	1.215	1.545	10.05	170	0.00
Record 5	M	13.71	26.27	186.9	2501	0.1084	0.1988	0.3635	0.1689	0.2061	0.05623	2.547	1.306	18.65	542.2	0.0
Record 8	M	13.71	17.46	174.2	2010	0.1149	0.2363	0.3368	0.1913	0.1956	0.06121	0.9948	0.8509	7.222	153.1	0.00
Record 11	M	13.71	21.02	124.4	994	0.123	0.2576	0.3189	0.1198	0.2113	0.07115	0.403	0.7747	3.123	41.51	0.00
Record 12	M	13.71	23.2	110.2	773.5	0.1109	0.3114	0.3176	0.1377	0.2495	0.08104	1.292	2.454	10.12	138.5	0.0
Record 13	M	13.71	25.09	143	1347	0.1099	0.2236	0.3174	0.1474	0.2149	0.06879	0.9622	1.026	8.758	118.8	0.00
Record 15	B	13.71	19.65	97.83	629.9	0.07837	0.2233	0.3003	0.07798	0.1704	0.07769	0.3628	1.49	3.399	29.25	0.00
Record 16	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.00
Record 18	M	13.71	21.87	182.1	2250	0.1094	0.1914	0.2871	0.1878	0.18	0.0577	0.8361	1.481	5.82	128.7	0.00
Record 20	M	13.71	22.53	102.1	685	0.09947	0.2225	0.2733	0.09711	0.2041	0.06898	0.253	0.8749	3.466	24.19	0.00
Record 21	M	13.71	23.56	138.9	1364	0.1007	0.1606	0.2712	0.131	0.2205	0.05898	1.004	0.8208	6.372	137.9	0.00
Record 22	M	13.71	21.51	135.9	1264	0.117	0.1875	0.2565	0.1504	0.2569	0.0667	0.5702	1.023	4.012	69.06	0.00
Record 23	M	13.71	25.12	130.4	1192	0.1015	0.1589	0.2545	0.1149	0.2202	0.06113	0.4953	1.199	2.765	63.33	0.00

Imputation | STDPartition1 | NNC_Output5 | NNC_TrainLog3 | NNC_TrainingScore3 | NNC_TrainingLiftChart3 | NI ...



Random seed	12345
Ratio - Training	0.6
Ratio - Validation	0.4

Partition Summary

Partition	# Records
Training	340
Validation	227

Partitioned Data

Record ID	Diagnosis	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11	Feature 12	Feature 13	Feature 14	Feature
Record 1	M	13.71	20.2	166.2	1761	0.1447	0.2867	0.4268	0.2012	0.2655	0.06877	1.509	3.12	9.807	233	0.0
Record 5	M	13.71	26.27	186.9	2501	0.1084	0.1988	0.3635	0.1689	0.2061	0.05623	2.547	1.306	18.65	542.2	0.0
Record 8	M	13.71	17.46	174.2	2010	0.1149	0.2363	0.3368	0.1913	0.1956	0.06121	0.9948	0.8509	7.222	153.1	0.00
Record 11	M	13.71	21.02	124.4	994	0.123	0.2576	0.3189	0.1198	0.2113	0.07115	0.403	0.7747	3.123	41.51	0.00
Record 12	M	13.71	23.2	110.2	773.5	0.1109	0.3114	0.3176	0.1377	0.2495	0.08104	1.292	2.454	10.12	138.5	0.0
Record 15	B	13.71	19.65	97.83	629.9	0.07837	0.2233	0.3003	0.07798	0.1704	0.07769	0.3628	1.49	3.399	29.25	0.00
Record 16	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.00
Record 18	M	13.71	21.87	182.1	2250	0.1094	0.1914	0.2871	0.1878	0.18	0.0577	0.8361	1.481	5.82	128.7	0.00
Record 20	M	13.71	22.53	102.1	685	0.09947	0.2225	0.2733	0.09711	0.2041	0.06898	0.253	0.8749	3.466	24.19	0.00
Record 21	M	13.71	23.56	138.9	1364	0.1007	0.1606	0.2712	0.131	0.2205	0.05898	1.004	0.8208	6.372	137.9	0.00
Record 22	M	13.71	21.51	135.9	1264	0.117	0.1875	0.2565	0.1504	0.2569	0.0667	0.5702	1.023	4.012	69.06	0.00
Record 23	M	13.71	25.12	130.4	1192	0.1015	0.1589	0.2545	0.1149	0.2202	0.06113	0.4953	1.199	2.765	63.33	0.00
Record 25	M	13.71	26.57	142.7	1311	0.1141	0.2832	0.2487	0.1496	0.2395	0.07398	0.6298	0.7629	4.414	81.46	0.00
Record 26	M	13.71	17.27	103.2	713.3	0.1335	0.2284	0.2448	0.1242	0.2398	0.07596	0.6592	1.059	4.061	59.46	0.0
Record 28	M	13.71	22.39	142	1479	0.111	0.1159	0.2439	0.1389	0.1726	0.05623	1.176	1.256	7.673	158.7	0.

NNC_ValidationLiftChart2 | NNC_Stored2 | NNC_Output3 | STDPartition | NNC_Output2 | NNC_TrainLog1 | NNC_ ...

I chose two partitioning levels for the NN's, a 60/40 split and a 80/20 split. For other models (in other assignments) an 80/20 split normally produced better results than 60/40 splits and I was curious if the same would be for NN.

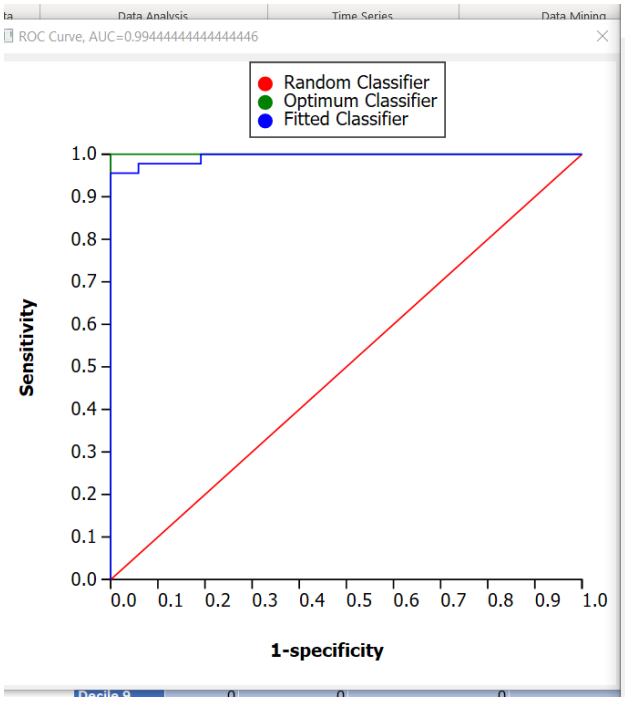
f) Choose the data mining techniques/algorithms – **Classify > Neural Networks** > **Automatic or Manual** and **build 4 models - (Remember to Rescale** (use standardization) **the dataset)**

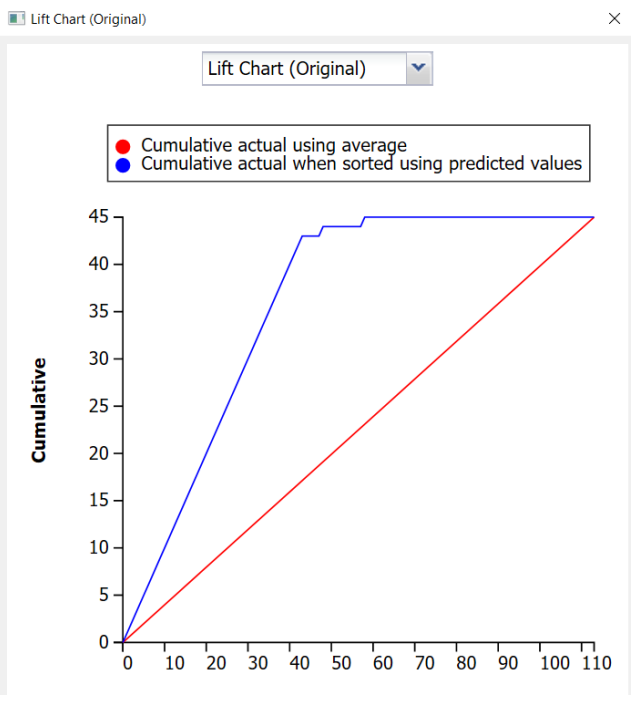
For getting the four models, I chose to do an automatic NN which took longer to build but created vastly more NN options of layers and nodes in each layer and there results. By doing this, I sorted those models by lowest error % to choose the best models. I chose 2 from each partitioning (shown in table).

g) Follow the steps of creating a model as shown in Lecture 9 slides
After obtaining NNC_Output and NNC_Output3, which are the results of running the automatic version of NN in the classify section. After sorting by lowest error %, I chose the two best models from each (60/40 and 80/20 partitioning for the respective output tabs). By clicking on the net name, it brough up the selection criteria similar to how a manual NN would be made, I made sure that the data was rescaled using standardization (also done in previous step), also made sure neural net weights was checked and that the detailed reports had the lift charts selected as well. That was done for each model chosen on each output tab.

h) Interpret the results and depending on the model selection criteria choose the **best model**

		Partitioning	Training accuracy %	Training error %	Validation accuracy %	Validation error %	Validation F1 score
Model 1		80/20	98.0170%	1.9824%	98.2300%	1.7699%	0.97727
Model 2		80/20	98.2379%	1.7621%	95.5752%	4.4247%	0.94252
Model 3		60/40	98.5294%	1.4706%	92.0704%	7.9295%	0.88607
Model 4		60/40	99.1176%	0.8824%	96.0352%	3.9647%	0.94478





First, for selecting the best model I looked at the accuracy of training and validation, which model 1 had the highest of. The validation error % was also the lowest as a result of the high accuracy, and model 1 also had a high F1 score. After that I looked at the lift chart and the AUC for model 1 validation (shown above). Model 1 had a very high AUC of 0.99444 which is very good. The lift chart also shows a good lift from the actual line (red).

i) Deploy **best model** on the new data and explain your prediction results (how many records/instances are Malignant/Benign. Paste a screenshot of your prediction for new data.

Scoring

Record ID	Prediction: Diagnosis	PostProb: B	PostProb: M
Record 1	M	0	1
Record 2	B	0.973067347	0.026932653
Record 3	B	0.758983695	0.241016305
Record 4	B	0.764889439	0.235110561
Record 5	B	0.780730933	0.219269067
Record 6	B	0.563100975	0.436899025
Record 7	B	0.924888077	0.075111923
Record 8	B	1	0
Record 9	B	0.882267969	0.117732031
Record 10	B	0.79475974	0.20524026
Record 11	B	0.783120536	0.216879464
Record 12	B	0.913090709	0.086909291
Record 13	M	0	1
Record 14	B	0.690918622	0.309081378
Record 15	M	0	1
Record 16	M	0.418919674	0.581080326
Record 17	B	0.700133645	0.299866355
Record 18	B	0.782586808	0.217413192

j) Submit the **Excel workbook and this word document with explanation/screenshots for steps c) through i).**