Jacob Perrone

Assignment 4 Report

Part 1

Part C-D)

Data								
Workbook	Class Assignment 4 - Flying_Fitness - Student Use Dataset-1.							
Worksheet	Data							
Range	$A$1:$H$139							
# Records in the input data	138							
Variables								
# Selected Variables	8							
Selected Variables	No.	Obs	Outcome c	Var2	Var3	Var4	Var5	Var6
Imputer Parameters								
Variable	No.	Obs	Outcome c	Var2	Var3	Var4	Var5	Var6
Reduction Type	NONE	NONE	NONE	DELETE REC	DELETE RECORD	DELETE REC	DELETE REC	DELETE RECORD
# Records Treated	0	0	0	1	2	1	1	1
Missing Value Code								
# Output Records	136							
#Records Deleted	2							

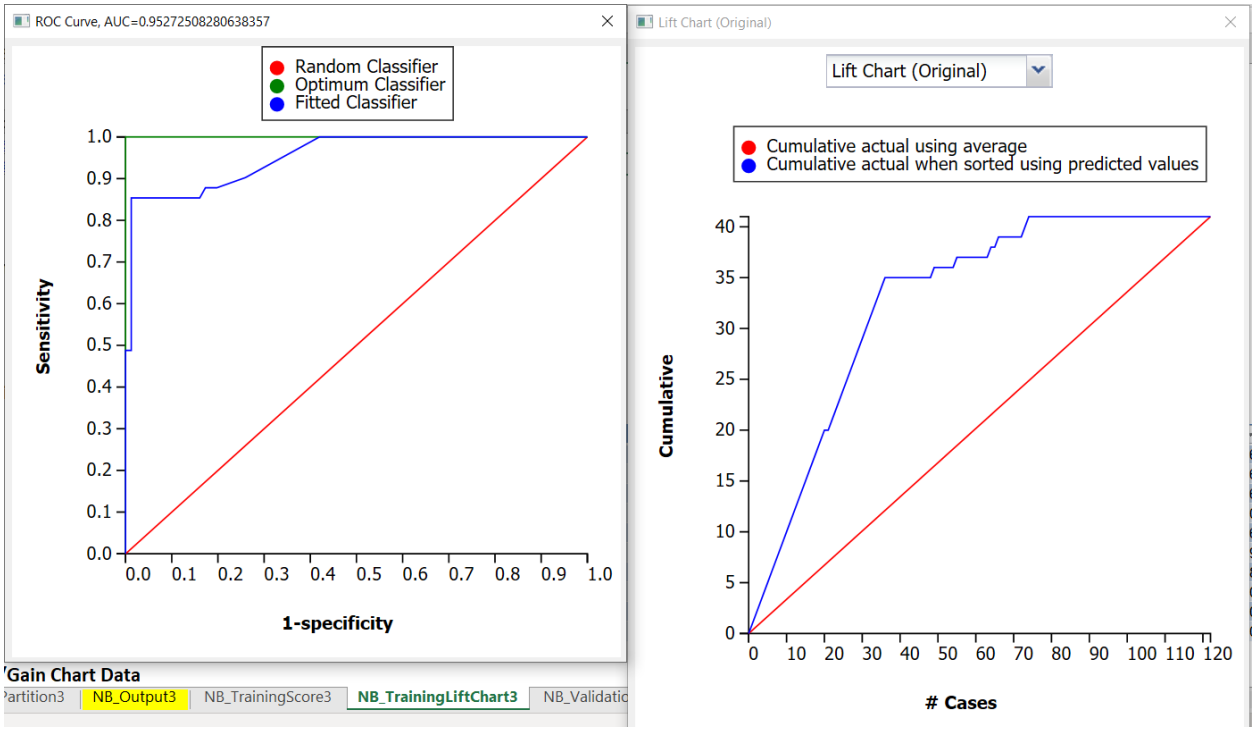
From the original dataset I went and sorted each column by largest value to see if there was any incorrect or missing data. I deleted the records that had incorrect data as there was only a few of them, the screen shot above shows the missing data handling from XLMiner which deleted records with missing data in any of the columns. Looking at the scatterplot matrix is not necessary as this is categorical data and dummies were not needed as this is a classification task.

Part E-I)

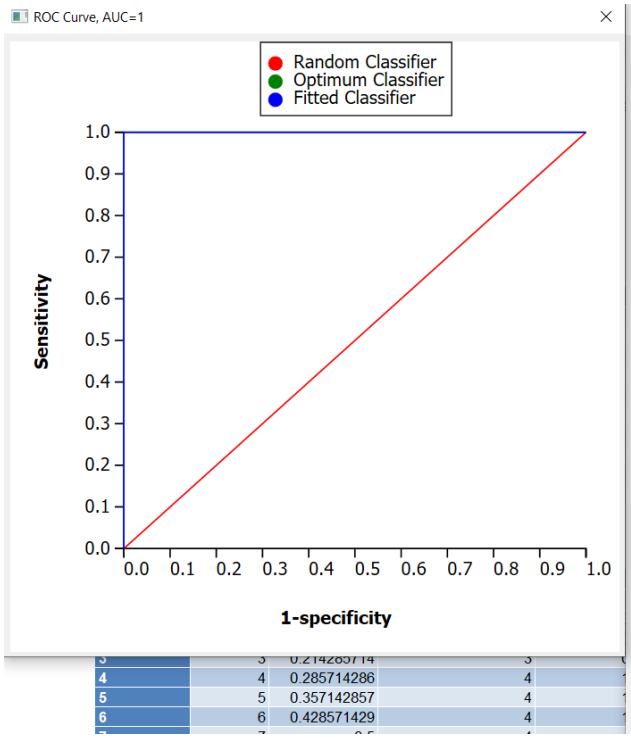
	A	B	C	D	E	F
1		Partitioning	Training ACC	Validation ACC	Training Error	Validation Error
2	Model 1	90/10	94.262295	100	5.7377	0
3	Model 2	80/20	96.33027	88.88888	3.6697	11.11111
4						
5						

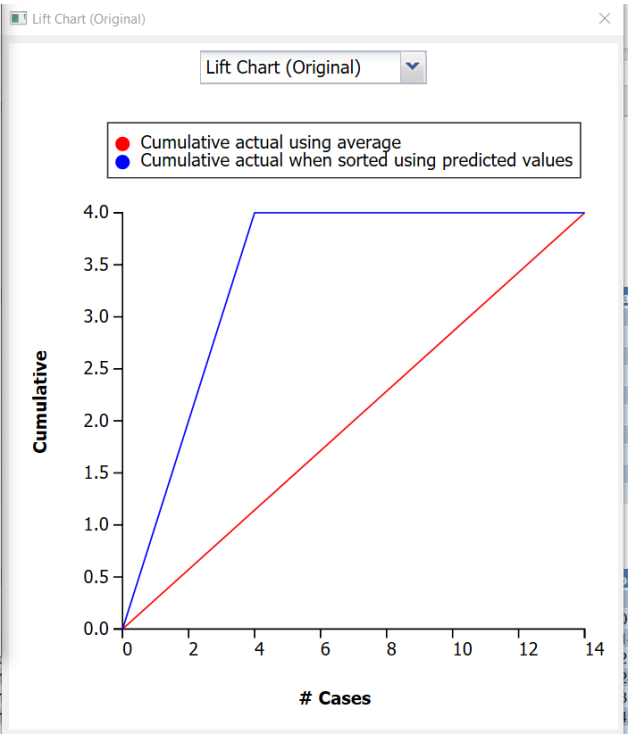
Model 1 was chosen as the best model due to having the highest validation accuracy and high ROC value (validation) as well as other metrics such as low error and high precision and sensitivity.

Training Lift charts



Validation Lift charts





4	4	0.285714286	4	
5	5	0.357142857	4	
6	6	0.428571429	4	

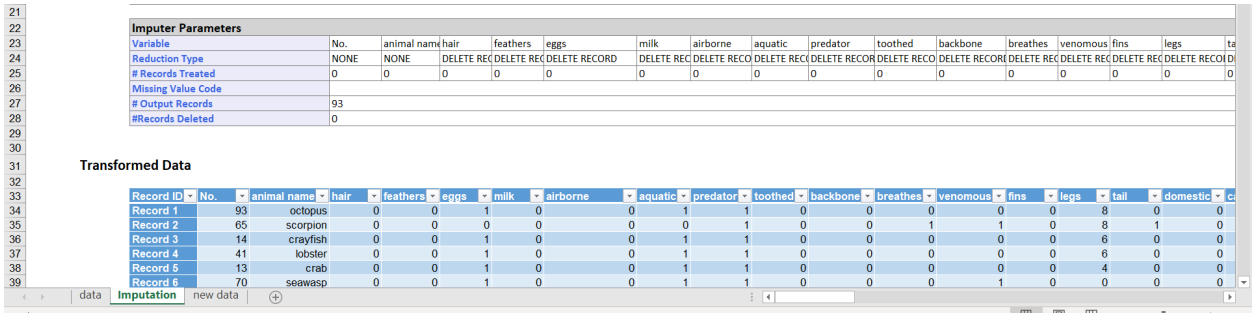
Scoring on New data

Scoring

Record ID	Prediction: Outcome class type	PostProb: 1	PostProb: 0
Record 1	0	0.174115104	0.825884896
Record 2	0	0	1
Record 3	0	0.46057652	0.53942348
Record 4	0	0	1
Record 5	0	0	1
Record 6	0	0.46057652	0.53942348
Record 7	0	0	1
Record 8	1	0.600040597	0.399959403
Record 9	1	1	0
Record 10	1	0.7629205	0.2370795
Record 11	1	0.600040597	0.399959403
Record 12	1	0.775093757	0.224906243
Record 13	1	1	0
Record 14	0	0.270303634	0.729696366

Part 2

Part C-D)



Imputer Parameters

Variable	No.	animal name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	ta
Reduction Type	NONE	NONE	DELETE REC	DELETE REC	DELETE RECORD	DELETE REC	DELETE RECO	DELETE REC	DELETE RECOR	DELETE RECO	DELETE RECOR	DELETE REC	DELETE REC	DELETE REC	DELETE RECO	D
# Records Treated	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missing Value Code																
# Output Records	93															
#Records Deleted	0															

Transformed Data

Record ID	No.	animal name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	ca
Record 1	93	octopus	0	0	1	0	0	1	1	0	0	0	0	0	8	0	0	
Record 2	65	scorpion	0	0	0	0	0	0	1	0	0	1	1	0	8	1	0	
Record 3	14	crayfish	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0	
Record 4	41	lobster	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0	
Record 5	13	crab	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0	
Record 6	70	seawasp	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	

data | Imputation | new data

Going through the original dataset, I sorted each predictor column to see if there was any incorrect of missing values. There were none and data handling in XLMiner showed that no records were deleted for missing data. Scatterplot matrix and dummy creation were not necessary as all predictors are categorical and this is a classification problem. From the predictors, the Living predictor can be omitted moving forward as all of the records for that column are the same value which could throw off the model and lead to bad performance.

Part E-I)

	A	B	C	D	E	F
1		Partitioning	Training ACC	Validation ACC	Training Error	Validation Error
2	Model 1	90/10	100	100	0	0
3	Model 2	80/20	98.6486	94.7368	1.35135	5.2631
4						
5						

Model 1 was chosen as the best model due to having the highest validation accuracy.

Scoring on new data

Scoring

Record ID	Prediction: class type	PostProb: 7	PostProb: 6	PostProb: 5	PostProb: 4	PostProb: 3	PostProb: 2	PostProb: 1
Record 1	1	0	0	0	0	0	0	1
Record 2	2	0	0	0	0	0	1	0
Record 3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Record 4	2	0	0	0	0	0	1	0
Record 5	2	0	0	0	0	0	1	0
Record 6	4	0	0	0	1	0	0	0
Record 7	1	0	0	0	0	0	0	1
Record 8	6	0	1	0	0	0	0	0

The deployed model was not able to classify starfish, I suppose that this is due to the list of predictors that are used and thus changing which ones the model uses could yield to a classification for starfish.