# **BAN210- Workshop 5**

| Total Mark: | 10 marks |
| --- | --- |
| Submission file(s): | * WS05.docx |

Please work in **groups** to complete this lab. This workshop is worth 5% of the total course grade. Please submit the submission file(s) through Blackboard. Only one person must submit for the group and only the last submission will be marked.

## **Part I: Video Tutorial**

1. Watch the fourth video in Getting Started with SAS Enterprise Miner playlist:

Video (SAS Enterprise Miner: Impute, Transform, Regression & Neural Models):

<https://youtu.be/TnWRJQb5z4c>

## **Part II: Logistic Regression**

1. Open the project you worked on in Workshop 1 & 3. Open the Retention diagram.
2. Follow the instructions on pages 3-73 to 3-77 of the SAS Advanced Business Analytics course notes. As you go through the steps, answer the following questions.

**Answer the following questions:**

* At step 2 on page 3-73, copy the Fall\_GPA plot and paste here.

Chart, histogram

Description automatically generated

* At step 5 on page 3-74,
  + What formula is used to normalize FALL\_GPA? Paste the formula here. How did the skewness change?

exp(max(Fall\_GPA-0, 0.0)/4)

A variable whose distribution is symmetric will have a value of skewness equal  
to 0.

Skewness : -0.97139

* + Paste the formula for normalizing DISTANCE. What is the effect on skewness?

sqrt(max(Distance-0.7858319691, 0.0)/3881.4065468)

Skewness : 4.026658

This variable distribution is asymmetric so the value of skewness is more than 0.

* At step 7 on page 3-74, plot the histogram of EXP\_Fall\_GPA (select X as the role), and paste here. Compare with the plot of Fall\_GPA you pasted above. What do you observe?

Chart, histogram

Description automatically generated

In Fall\_GPA the increment on X axis starts from value 0 and it is continuously increasing.

And in EXP\_Fall\_GPA the Transformed: Fall\_GPA value on x axis starts from 1 and it is increasing but it is not continuous increment in value.

* At step 9 on page 3-75, read the explanation in Help Contents for ‘Tree Surrogate’ methods. What does this setting do?

Tree Surrogate — the same as Tree Imputation, except with the addition of surrogate splitting rules. A surrogate rule is a back-up to the main splitting rule. When the main splitting rule relies on an input whose value is missing, the next surrogate is invoked. If missing values prevent the main rule and all the surrogates from applying to an observation, the main rule assigns the observation to the branch that is assigned to receive missing values.

* At step 11 on page 3-75, you selected a stepwise selection model. What does this mean? (Read the help contents for *Regression Node Model Selection Methods*.)

**Stepwise** — As in the Forward method, Stepwise selection begins, by default, with no candidate effects in the model and then systematically adds effects that are significantly associated with the target. However, after an effect is added to the model, Stepwise may remove any effect already in the model that is not significantly associated with the target.

This stepwise process continues until one of the following occurs:

* + No other effect in the model meets the Stay Significance Level.
  + The Max Steps criterion is met. If you choose the Stepwise selection method, then you can specify a Max Steps to put a limit on the number of steps before the effect selection process stops. The default value is set to the number of effects in the model. If you add interactions via the Interaction Builder, the Max Steps is automatically updated to include these terms.
  + An effect added in one step is the only effect deleted in the next step.

1. At step 19, fill in the following table for the validation set. Then use the formulas to calculate the evaluation measures for the classifier:

TP: True Positives

FP: False Positives

TN: True Negatives

FN: False Negatives

Table

Description automatically generated

A picture containing text

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Misclassification Tree | | | |
|  | Detected as 0  (outcome= 0) | Detected as 1  (outcome = 1) | Total |
| Truly 0  (target = 0) | TN= 904 | FP=14 | FP+TN = 918 |
| Truly 1  (target = 1) | FN= 83 | TP= 51 | TP+FN = 134 |
| Total | TN+FN=987 | TP+FP= 65 |  |

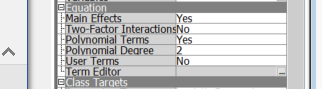
Recall (R) = TP / (TP + FN) = 0.3805

Precision (P) = TP / (TP + FP) = 0.7846

F1= 2P.R / (P + R) = 0.5124

What is the problem with this classifier?

Both Recall and Precision values are low, the model correctly classifies 38% of positive outcomes, and classifies correctly only 78.4% of total true positives. Therefore, F1 score is considered low at 0.5124.

1. Close the results and make the following change in the properties of the Regression node to allow for a quadratic polynomial term to be included in the logistic regression equation: 
2. Recalculate the F1 measure for the validation set (show your calculations). Is the performance better now? (Higher F1 values are better.)

Table

Description automatically generated

A picture containing text

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Misclassification Tree | | | |
|  | Detected as 0  (outcome= 0) | Detected as 1  (outcome = 1) | Total |
| Truly 0  (target = 0) | TN= 896 | FP=22 | FP+TN = 918 |
| Truly 1  (target = 1) | FN= 82 | TP= 52 | TP+FN = 134 |
| Total | TN+FN=978 | TP+FP= 74 |  |

Recall (R) = TP / (TP + FN) = 0.3880

Precision (P) = TP / (TP + FP) = 0.7027

F1= 2P.R / (P + R) = 0.50

Recall value is higher than the previous model, the model correctly classifies 38.8% of positive outcomes, For Precision classifies correctly only 70.2% of total true positives which is less than previous model. F1 score is not higher than the previous model at 0.50 thus, it has a similar value as the previous model has. This model does not perform better.

## **Part III: Neural Networks**

1. Apply Variable selection as suggested in the Video mentioned above (after the **Impute** node). How many variables are selected?

Table

Description automatically generated

1. Use the same neural network configuration as mentioned in the video.
2. Recalculate the F1 measure for the validation set (show your calculations). How does the performance compare with the regression results?

Table

Description automatically generated

Text

Description automatically generated with low confidence

|  |  |  |  |
| --- | --- | --- | --- |
| Misclassification Tree | | | |
|  | Detected as 0  (outcome= 0) | Detected as 1  (outcome = 1) | Total |
| Truly 0  (target = 0) | TN= 903 | FP=15 | FP+TN = 918 |
| Truly 1  (target = 1) | FN= 81 | TP= 53 | TP+FN = 134 |
| Total | TN+FN=984 | TP+FP= 68 |  |

Recall (R) = TP / (TP + FN) = 0.3955

Precision (P) = TP / (TP + FP) = 0.7794

F1= 2P.R / (P + R) = 0.5247

The performance has slightly improved when compared with the first regression results and it performs a little worse than the regression after performing the stepwise selection process.

1. Right click on the Retention diagram and choose ‘Copy to clipboard’. Then paste the diagram here.

Graphical user interface

Description automatically generated with medium confidence

## **Part IV: KNN**

1. Start a new **project** named *W05\_KNN*.
2. Create a new **diagram** and name it as *KNN*.
3. Select the **Sample** tab and find the **File Import** node. Drag and drop the File Import node to the diagram. In property panel, under Train, select the **Import File** item and click on the properties indicated by the three dots. Choose W05.xlsx file.
4. Right click on the File Import node and click on **Edit Variables**.

* Click on Target and make sure that the role **Target** is selected and level is set to **Binary**. Then click **Explore**.
* How many rows were imported? How many of these rows have Target = 1?
* Close the explore window. Then select X1 and X2, then choose **Explore**. Are the distributions of X1 and X2 normal?
* Close and press OK.

1. Select the **Sample** tab and find the **Data Partition** node. Drag and drop the Data Partition node to the diagram and connect to the File Import node. In property panel, choose an 80:20 split for training versus validation set allocations (and zero for test set). Run and see the results.

* How many data samples are in the training set? What percentage has target = 1?

1. Close the Results window.
2. Select the model tab and find the **MBR** node. The Memory-Based Reasoning node uses a k-nearest neighbor algorithm to categorize or predict observations. Drag and drop the node to the diagram and connect to the data partition node.
3. Check the property panel.

* What is the default method under Train?
* Change the method to **scan**. This method scans through every observation in the data set and calculates its distance to the new observation.
* What is the default value for k (number of neighbors)?
* Change the value for **Weighted** to **No**.

1. Run the node and see the results.

* Under Fit statistics, what is the mean squared error for training versus validation set?
* In the Output window, find the number of FN, FP, etc. for the validation set and paste here.
* Check the run duration in the property panel and paste here.

1. Change the value of k to 1. Run again.

* What is the warning? What is the MSE values this time?
* Paste the number of FN, FP, etc. for validation set here. How did the results change?
* Check the run duration in the property panel and paste here.

1. Find the **Principal Component** node under Modify tab. Drag and drop into the diagram and place it between the **Data Partition** and the **MBR** nodes. Run this node and check the output window in the results.

* What are the Eigen values?
* What percentage of the energy is explained by the first principal component?

1. Change k (number of nearest neighbors) back to 16 in the MBR node and run again. Paste the FN, FP, etc. counts here. Did the results improve?
2. Right click on the diagram and choose ‘Copy to clipboard’. Then paste the diagram here.

## **Part V: Group Work**

1. Add this declaration to your file:

We, ------------ (mention your names), declare that the attached assignment is our own work in accordance with the Seneca Academic Policy. We have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. We have not distributed our work to other students.

1. Specify what each member has done towards the completion of this work:

|  | Name | Task(s) |
| --- | --- | --- |
| 1 | Jigar Parmar | Part 1 2 3 |
| 2 | Nicky Rathod | Part 1 2 3 |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |