KNN

Section 1)

Since all the variables in this dataset are having the same range and units, no transformation is required.

Section 2)

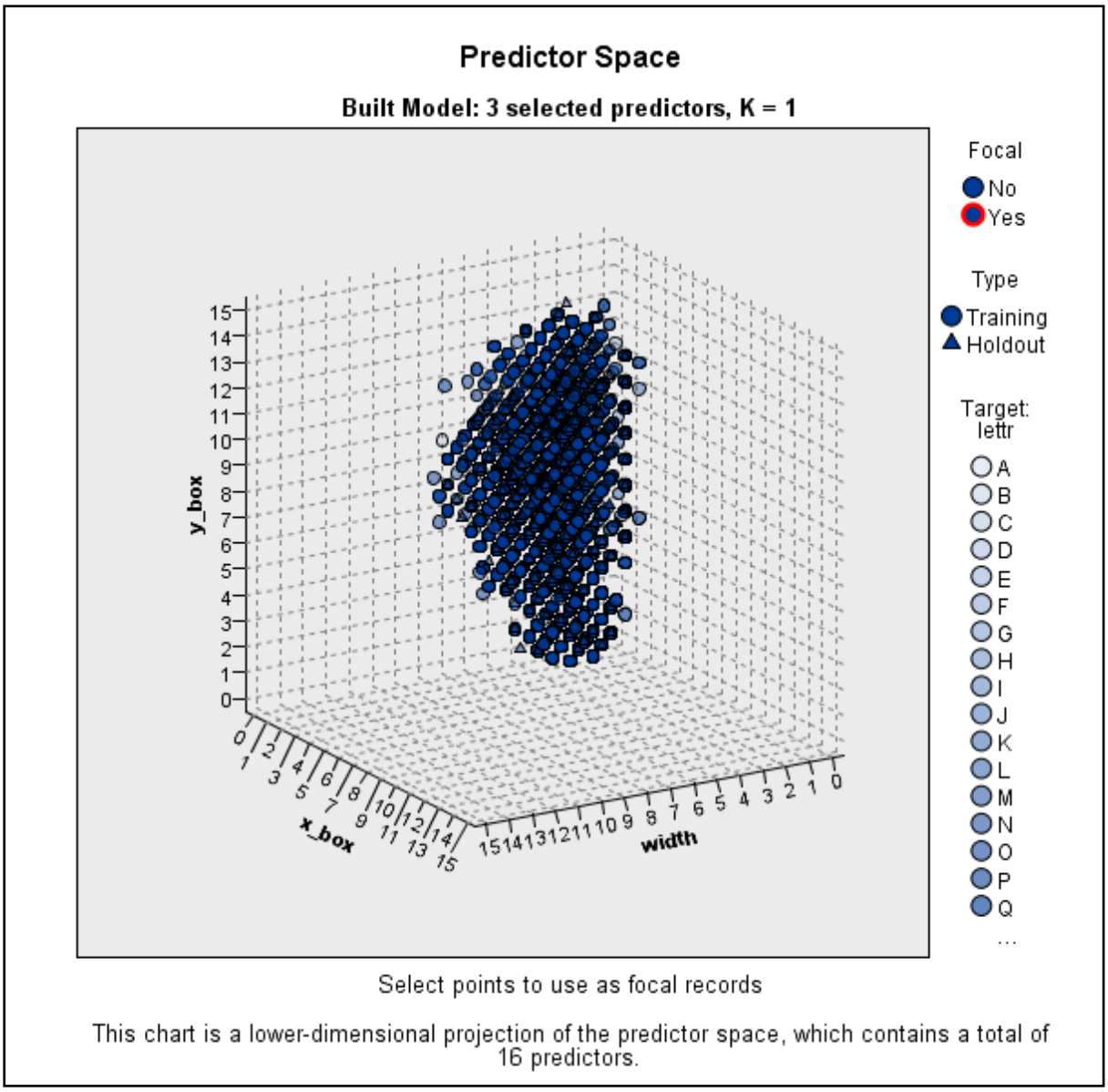
Dividing the dataset into training and testing sample with an 80:20 ratio.

When k = 1

Case Processing Summary

		N	Percent
Sample	Training	15947	79.7%
	Holdout	4052	20.3%
Valid		19999	100.0%
Excluded		1	
Total		20000	

1 record was excluded from the processing.



Classification for lettr

Overall Percent Correct = 88.0%

The classification table is not available because lettr has more than 50 categories.

Attached is the complete misclassification matrix.

knn_1.xlsx

The training set has an overall accuracy of 88%

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	12.0%
Holdout	12.4%

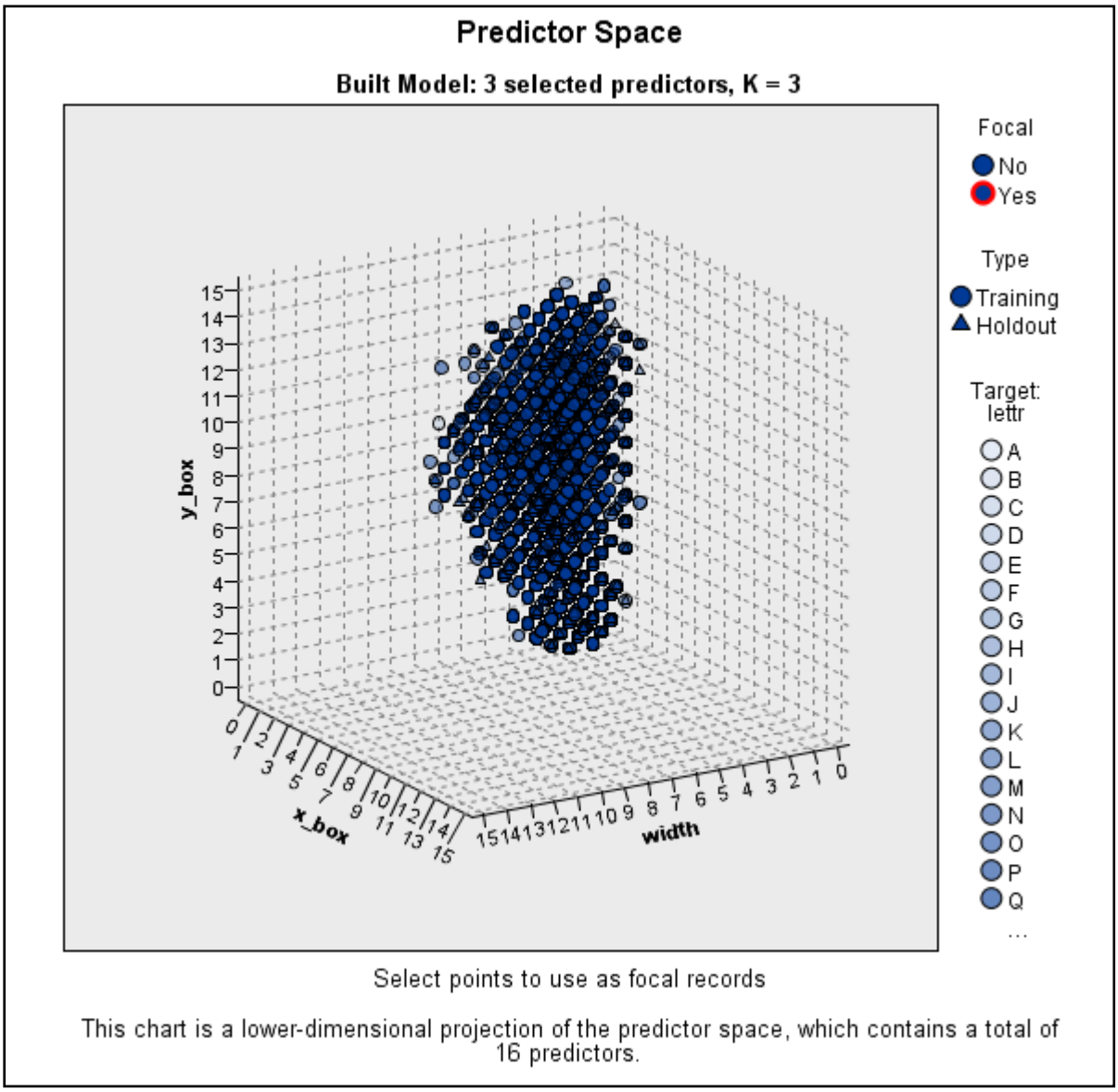
As shown in the error summary report, to the "risk" of misclassifying a letter in the training set is 12%.
The "risk" of misclassifying a letter in the testing set is approximately 12.4%.

When k = 3

Case Processing Summary

		N	Percent
Sample	Training	16019	80.1%
	Holdout	3979	19.9%
Valid		19998	100.0%
Excluded		2	
Total		20000	

2 records were excluded from the processing.



Classification for lettr

Overall Percent Correct = 86.5%

The classification table is not available because lettr has more than 50 categories.

Attached is the complete misclassification matrix.

knn_3.xlsx

The training set has an overall accuracy of 86.5%

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	13.5%
Holdout	12.4%

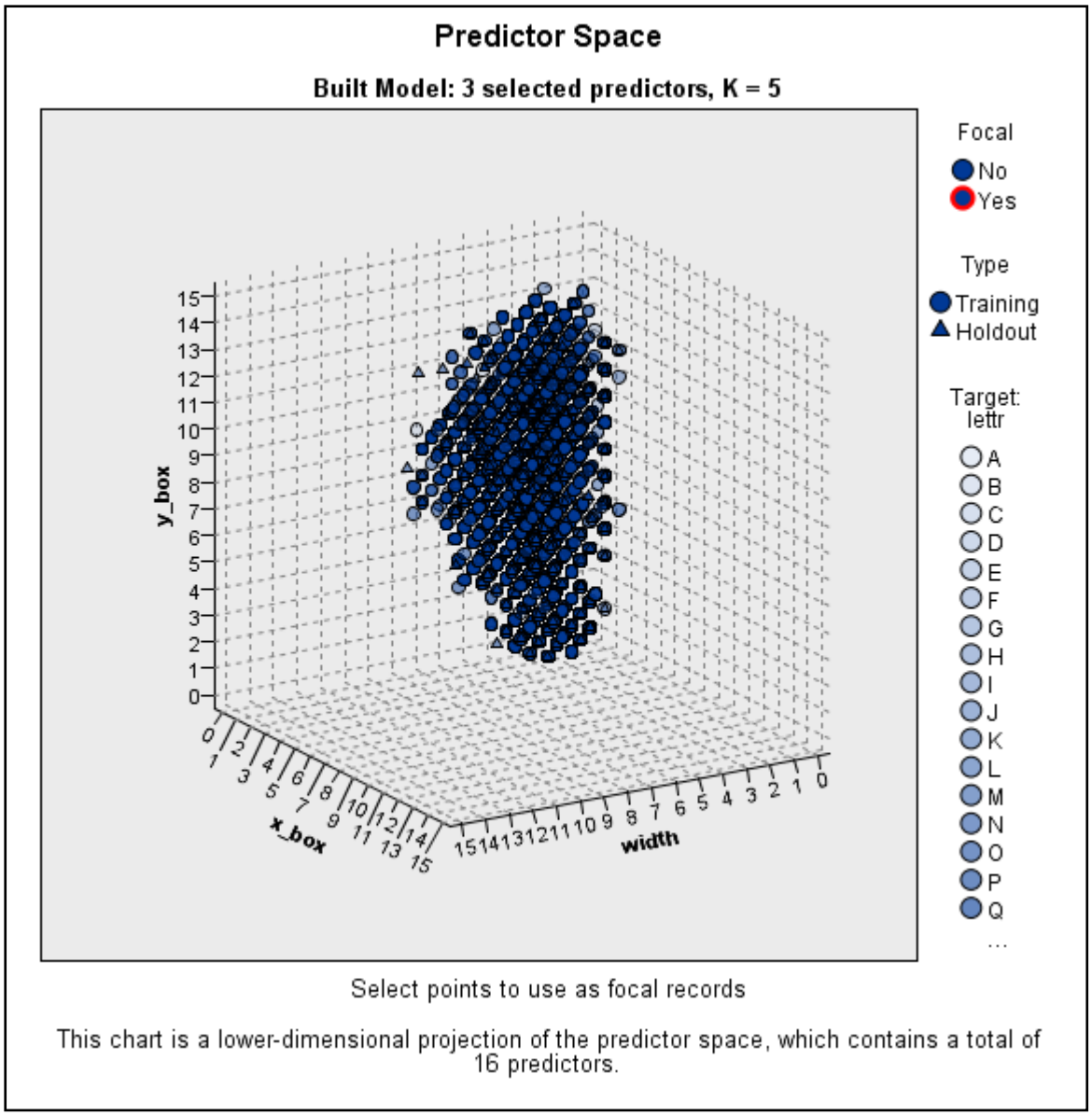
As shown in the error summary report, to the "risk" of misclassifying a letter in the training set is 13.5%.
The "risk" of misclassifying a letter in the testing set is approximately 12.4%.

When k = 5

Case Processing Summary

		N	Percent
Sample	Training	15888	79.4%
	Holdout	4112	20.6%
Valid		20000	100.0%
Excluded		0	
Total		20000	

No records were excluded from processing.



Classification for lettr

Overall Percent Correct = 87.5%

The classification table is not available because lettr has more than 50 categories.

Attached is the complete misclassification matrix.

knn_5.xlsx

The training set has an overall accuracy of 87.5%

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	12.5%
Holdout	12.2%

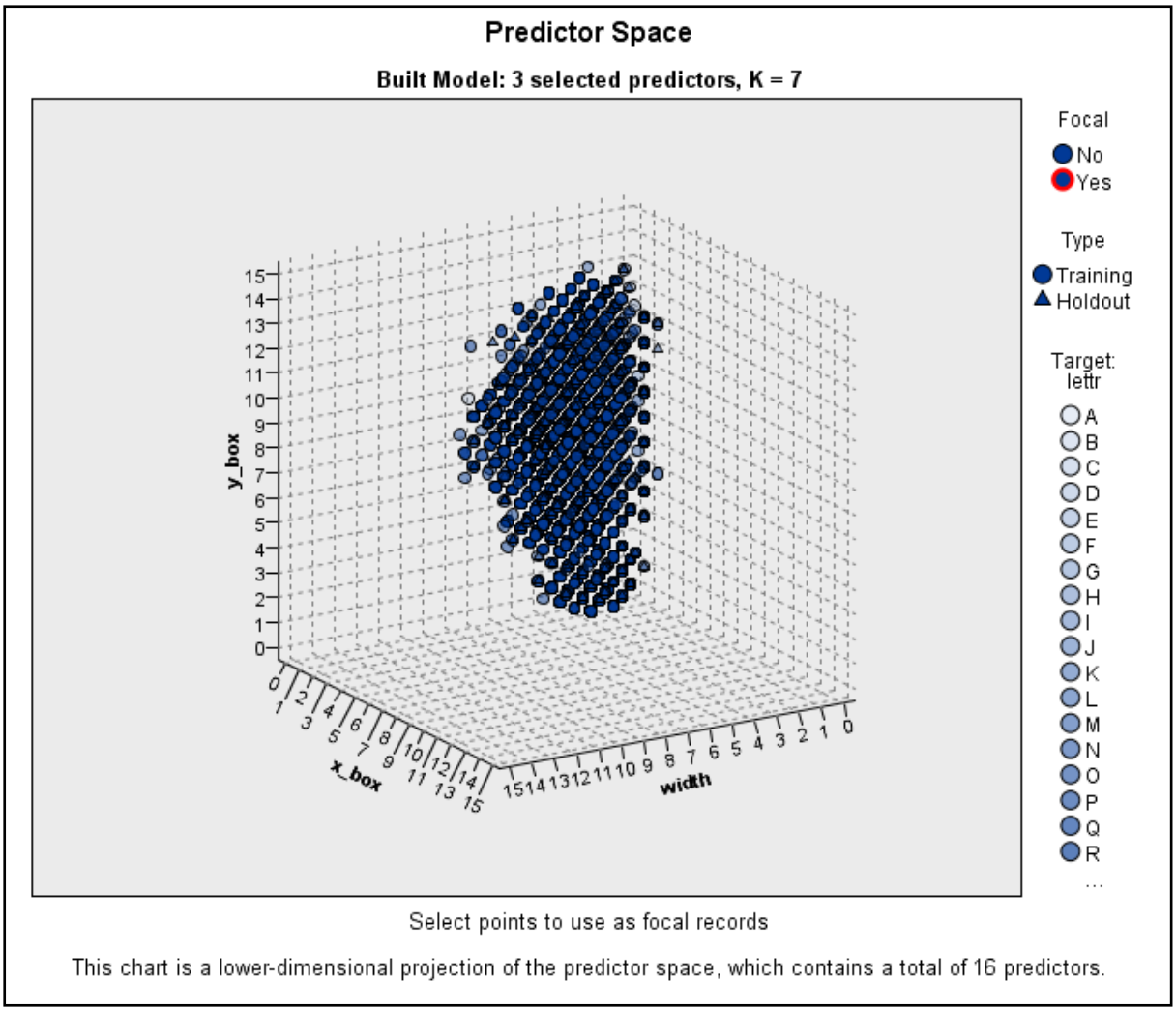
As shown in the error summary report, to the "risk" of misclassifying a letter in the training set is 12.5%.
The "risk" of misclassifying a letter in the testing set is approximately 12.2%.

When k = 7

Case Processing Summary

		N	Percent
Sample	Training	15965	79.8%
	Holdout	4034	20.2%
Valid		19999	100.0%
Excluded		1	
Total		20000	

1 record was excluded from processing.



Classification for lettr

Overall Percent Correct = 87.2%

The classification table is not available because lettr has more than 50 categories.

Attached is the complete misclassification matrix.

knn_7.xlsx

The training set has an overall accuracy of 87.2%

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	12.8%
Holdout	12.8%

As shown in the error summary report, to the "risk" of misclassifying a letter in the training set is 12.8%.
The "risk" of misclassifying a letter in the testing set is approximately 12.8%.

Section 3)

The KNN model produces a misclassification value of around 12 – 14%. The classification for KNN model varies from ~ 86 – 88 % for different values of k.
The decision tree in the previous section has a misclassification range of 12 – 28% on the training set while 18 – 31% on the testing set.
The accuracy prediction ranges from 71 – 88% on the training set and 69 – 82% on the testing set.

It turns out that KNN has a better accuracy and misclassification rate for this dataset.