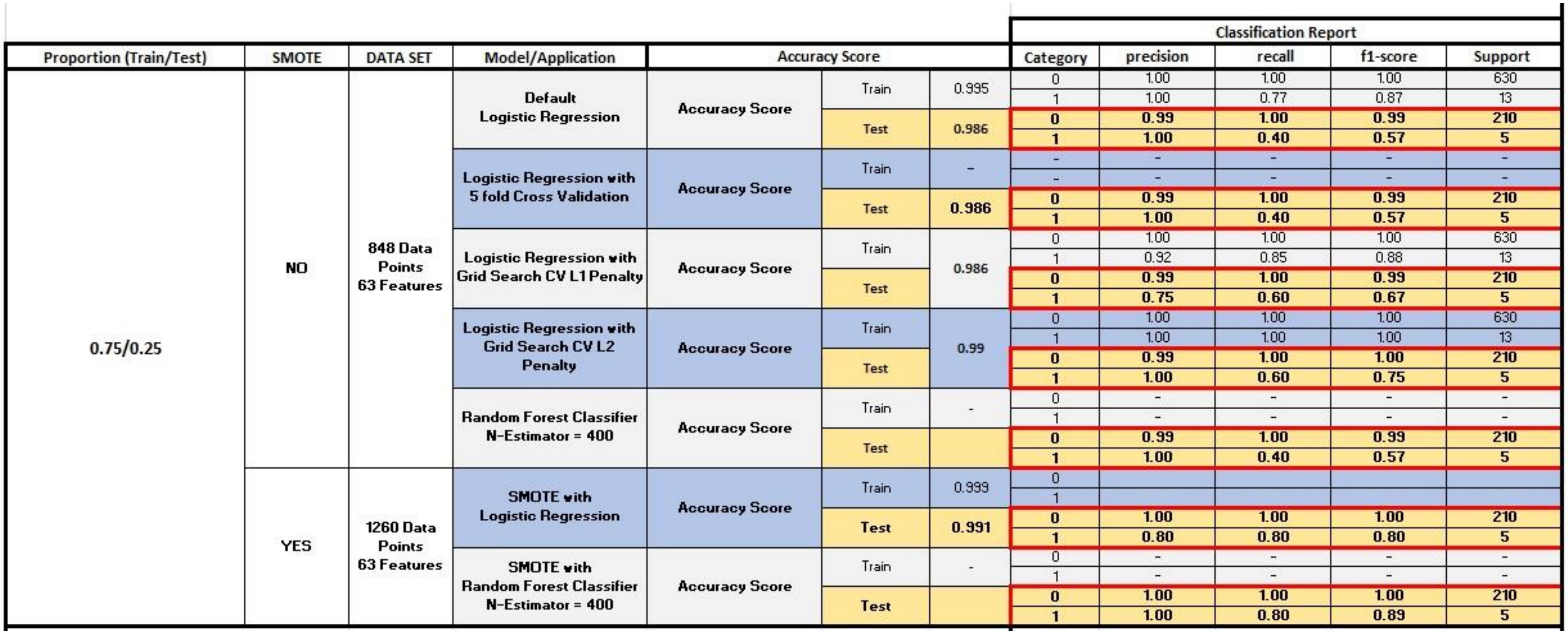
Capstone Project-1 : In-depth Analysis (Machine Learning)

1. Splitting the data into two sets (Train and Test sets with the proportion of 0.75/0.25)



							Classification Report				
Proportion (Train/Test)	SMOTE	DATA SET	Model/Application	Accuracy Score			Category	precision	recall	f1-score	Support
0.75/0.25	NO	848 Data Points 63 Features	Default Logistic Regression	Accuracy Score	Train	0.995	0	1.00	1.00	1.00	630
							1	1.00	0.77	0.87	13
					Test	0.986	**0**	**0.99**	**1.00**	**0.99**	**210**
							1	**1.00**	**0.40**	**0.57**	**5**
			Logistic Regression with 5 fold Cross Validation	Accuracy Score	Train	-	-	-	-	-	-
							-	-	-	-	-
					Test	**0.986**	**0**	**0.99**	**1.00**	**0.99**	**210**
							1	**1.00**	**0.40**	**0.57**	**5**
			Logistic Regression with Grid Search CV L1 Penalty	Accuracy Score	Train	0.986	0	1.00	1.00	1.00	630
							1	0.92	0.85	0.88	13
					Test		**0**	**0.99**	**1.00**	**0.99**	**210**
							1	**0.75**	**0.60**	**0.67**	**5**
			Logistic Regression with Grid Search CV L2 Penalty	Accuracy Score	Train	0.99	0	1.00	1.00	1.00	630
							1	1.00	1.00	1.00	13
					Test		**0**	**0.99**	**1.00**	**1.00**	**210**
							1	**1.00**	**0.60**	**0.75**	**5**
			Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
							1	-	-	-	-
					Test		**0**	**0.99**	**1.00**	**0.99**	**210**
							1	**1.00**	**0.40**	**0.57**	**5**
	YES	1260 Data Points 63 Features	SMOTE with Logistic Regression	Accuracy Score	Train	0.999	0				
							1				
					Test	**0.991**	**0**	**1.00**	**1.00**	**1.00**	**210**
							1	**0.80**	**0.80**	**0.80**	**5**
			SMOTE with Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
							1	-	-	-	-
					Test		**0**	**1.00**	**1.00**	**1.00**	**210**
							1	**1.00**	**0.80**	**0.89**	**5**

a) Cervical Cancer Risk Data set is pretty imbalanced since Dx:Cancer(response vector) consists of 1 and 0 which has a 18/840 proportion. After splitting out the data set two parts with the .25/.75 proportion and applying Logistic Regression model. We found out that our model was predicting Nan-cancer patients with %99 accuracy which was expected. But for the cancer patients, model could not get the same accuracy results and 2 patients out of 5 patients were correctly predicted as cancer but 3 patients were missed (0.4). As seen, there were only 5 cancer patients in the test set which did not allow the model to predict more accurate.

b) After applying the Logistic Regression with 5-Fold Cross Validation, we got 0.986 accuracy score. However, there was no improvement of the proper prediction of the Cancer Patients.

c) Then we applied Hyper Parameter optimization to the model with Grid Search CV both L1 and L2 (default) penalties. After applying L1 Grid Search CV optimization, the accuracy score did not change but prediction of Cancer patients increased up to %60 which means 3 out of 5 patients were predicted as Cancer.

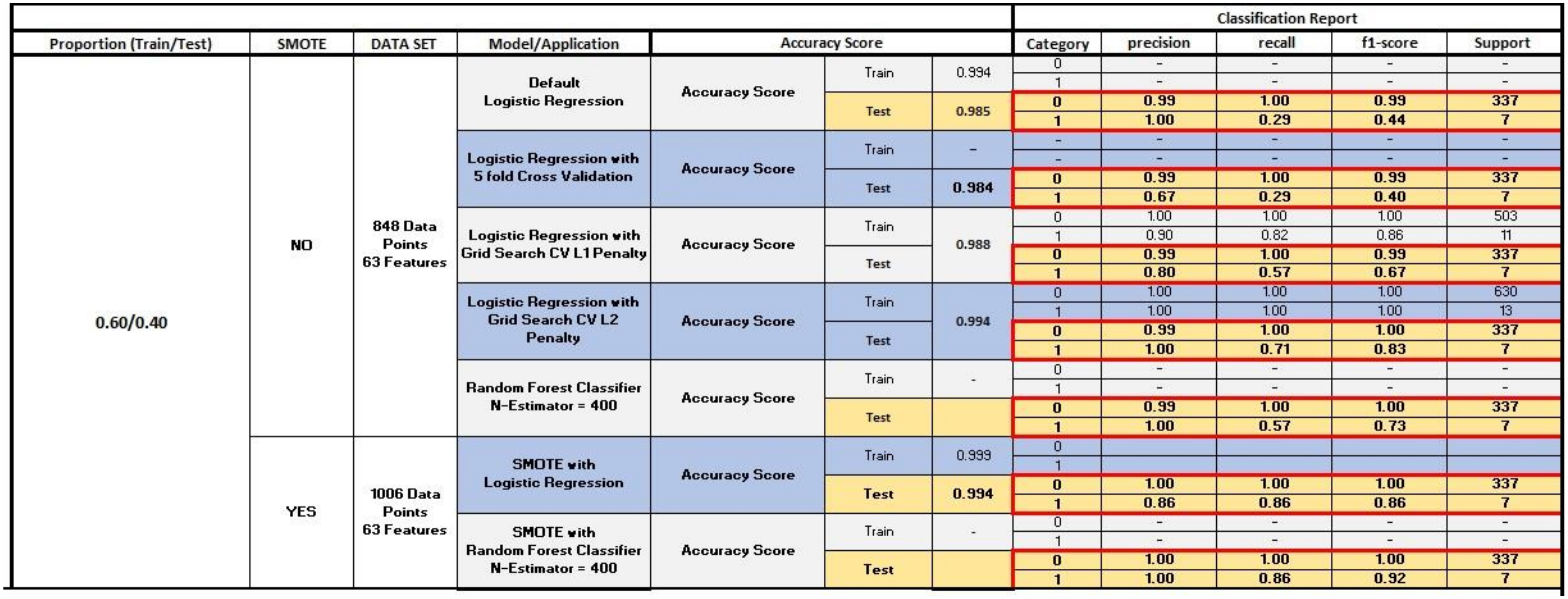
d) Once we applied L2 Grid Search CV optimization, model produced as the same result as L1 Grid Search CV on recall which was 0.6.

e) When we applied Random Forest Classifier to increase the prediction of our model, recall did not change.

f) In order to increase 1 (positive) samples in Dx:Cancer to get more accurate prediction model, we applied Synthetic Minority Oversampling Technique(SMOTE) to the train set and created 1260 (consisting of equal positive and negative samples) data points. Then we applied Logistic Regression and Random Forest Classifier to increase the accuracy of prediction. This time our model started to predict Cancer patients with %80 accuracy and only 1 out of 5 patients is missed by model.

g) Since we do not have enough positive sample in our data set, only 18, we will tweak with the train and test set proportion to get more accurate model. We will give test set more positive (1) samples and apply SMOTE to train set to get better prediction.

2. Splitting the data into two sets (Train and Test sets with the proportion of 0.60/0.40)



							Classification Report				
Proportion (Train/Test)	SMOTE	DATA SET	Model/Application	Accuracy Score			Category	precision	recall	f1-score	Support
0.60/0.40	NO	848 Data Points 63 Features	Default Logistic Regression	Accuracy Score	Train	0.994	0	-	-	-	-
							1	-	-	-	-
					Test	0.985	**0**	**0.99**	**1.00**	**0.99**	**337**
							1	**1.00**	**0.29**	**0.44**	**7**
			Logistic Regression with 5 fold Cross Validation	Accuracy Score	Train	-	-	-	-	-	-
							-	-	-	-	-
					Test	**0.984**	**0**	**0.99**	**1.00**	**0.99**	**337**
							1	**0.67**	**0.29**	**0.40**	**7**
			Logistic Regression with Grid Search CV L1 Penalty	Accuracy Score	Train	0.988	0	1.00	1.00	1.00	503
							1	0.90	0.82	0.86	11
					Test		**0**	**0.99**	**1.00**	**0.99**	**337**
							1	**0.80**	**0.57**	**0.67**	**7**
			Logistic Regression with Grid Search CV L2 Penalty	Accuracy Score	Train	0.994	0	1.00	1.00	1.00	630
							1	1.00	1.00	1.00	13
					Test		**0**	**0.99**	**1.00**	**1.00**	**337**
							1	**1.00**	**0.71**	**0.83**	**7**
			Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
							1	-	-	-	-
					Test		**0**	**0.99**	**1.00**	**1.00**	**337**
							1	**1.00**	**0.57**	**0.73**	**7**
	YES	1006 Data Points 63 Features	SMOTE with Logistic Regression	Accuracy Score	Train	0.999	0				
							1				
					Test	**0.994**	**0**	**1.00**	**1.00**	**1.00**	**337**
							1	**0.86**	**0.86**	**0.86**	**7**
			SMOTE with Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
							1	-	-	-	-
					Test		**0**	**1.00**	**1.00**	**1.00**	**337**
							1	**1.00**	**0.86**	**0.92**	**7**

a) After changing the proportion of the train and test size, once we applied Logistic Regression model the accuracy of predicting Cancer patients decreased up to 0.29 in our test model which means 2 out of 7 patients were predicted accurately as Cancer

b) After applying Logistic Regression with 5-Fold Cross Validation with, there was no improvement of the prediction.

c) Then we applied Hyper Parameter Tuning to the model with Grid Search CV both L1 and L2 (default) penalties. After applying L1 Grid Search CV regularization, the prediction of Cancer patients increased up to %57 which means 4 out of 7 patients were predicted as Cancer.

d) Once we applied L2 Grid Search CV regularization, prediction accuracy of Cancer patients increased up to (0.71) and model predicted 5 out of 7 patients as Cancer.

e) Then we applied Random Forest Classifier to increase the prediction of our model. But recall decreased to 0.57. Our model predicted 4 out of 7 patients as Cancer.

f) In order to increase 1 (positive) samples in Dx:Cancer to get more accurate prediction model, we applied Synthetic Minority Oversampling Technique(SMOTE) to the train set and created 1006 (consisting of equal positive and negative samples) data points. Then we applied Logistic Regression and Random Forest Classifier to increase the accuracy of prediction. This time our model started to predict Cancer patients with %85.7 accuracy and only 1 out of 7 patients was missed by model.

3. Conclusion:

After changing the proportion of Train and Test Set, our model's prediction accuracy almost increased up to %6 and 6 out of 7 patients are also predicted as Cancer correctly. If we had more Cancer patient samples in data set, we would have train our model better and get more accurate predictions. To said that we would advise the customer to get more Cancer samples to have better predictions.