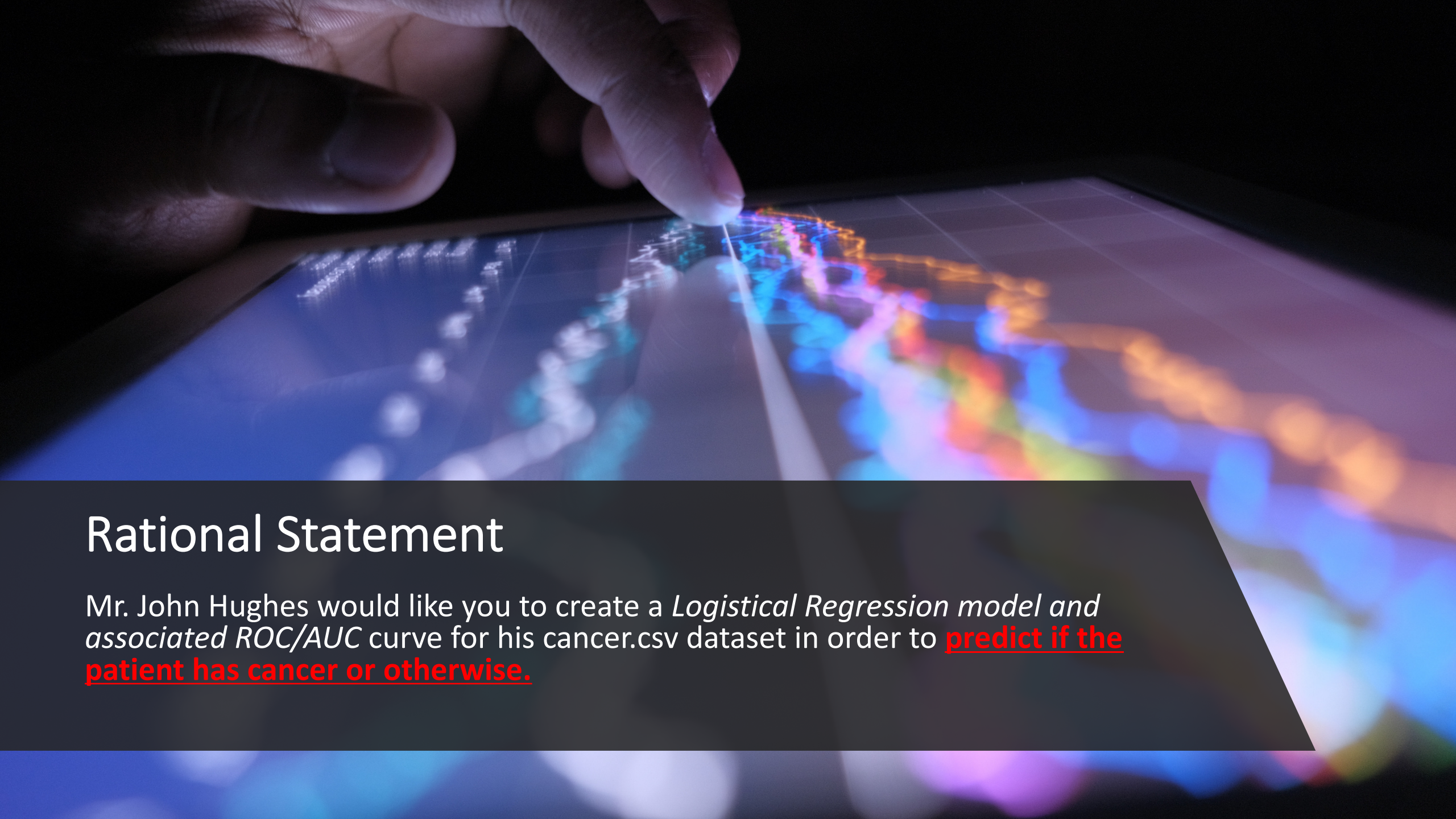
Logistical Regression

Prepared By: Rana Khan



Rational Statement

Mr. John Hughes would like you to create a *Logistical Regression model and associated ROC/AUC* curve for his cancer.csv dataset in order to **predict if the patient has cancer or otherwise.**

Independent Variable

- ID - ID number
- Clump Thickness - 1-10
- UofCSize - Uniformity of Cell Size 1-10
- UofShape - Uniformity of Cell Shape 1-10
- Marginal Adhesion - 1-10
- SECSize - Single Epithelial Cell Size 1-10
- Bare Nuclei - 1-10
- Bland Chromatin - 1-10
- Normal Nucleoli - 1-10
- Mitoses - 1-10

Dependent Variable

- Benign (i.e. No Cancer) - 2
- Malignant (i.e. Cancer) - 4

Navigation Synopsis

Copy cancer.csv into Pythondata2204 directory. Ensured the file is called cancer.csv

Launched Jupyter NoteBook

Navigated to Pythondata2204 Directory

Created a new Python NoteBook by clicking on “New Drop Down and Choose “Python3”

Renamed NoteBook by Clicking on File->Rename

Entered “Assignment#2-Logistic Regression” and Clicked OK

Learning Curve for the Logistical Regression

Learning Curve for the Logistical Regression

Bias Variance Trade-Off

Average Bias: 0.02
Average Variance: 0.01

1. There is not much variation between train and validation.
2. Recall for training and validation is in the range between 0.9 to 1.0 so it is proven that is less bias found.

Classification Report metrics for the Optimized Logistical Regression Model

Classification Report metrics for the Optimized Logistical Regression Model

Confusion Matrix

```
[[89  0]
[ 3 45]]
```

Classification Report

	precision	recall	f1-score	support
Outcome 0	0.97	1.00	0.98	89
Outcome 1	1.00	0.94	0.97	48
accuracy			0.98	137
macro avg	0.98	0.97	0.98	137
weighted avg	0.98	0.98	0.98	137

Key Insights

Accuracy:
Accuracy is the ratio of correct predictions to total predictions made. In this case, accuracy is 98%

Precision:
Precision is about being precise, i.e., how accurate your model is. In other words, we can say, when a model makes a prediction, how often it is correct. In this case, precision is 98%

Recall:

The recall is the ratio of correctly predicted positive observations to all observations in the actual class. In this case, recall 97%

F1 score:

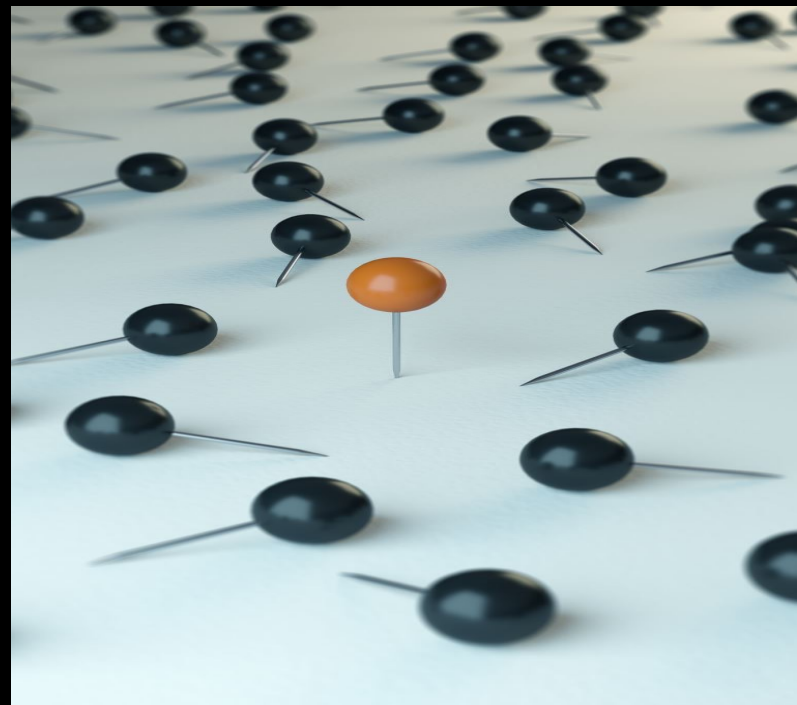
F1 Score is the weighted average of Precision and Recall. In our case, F1 score 98%

Key insights from ROC/AUC Curve (Optimized Model)

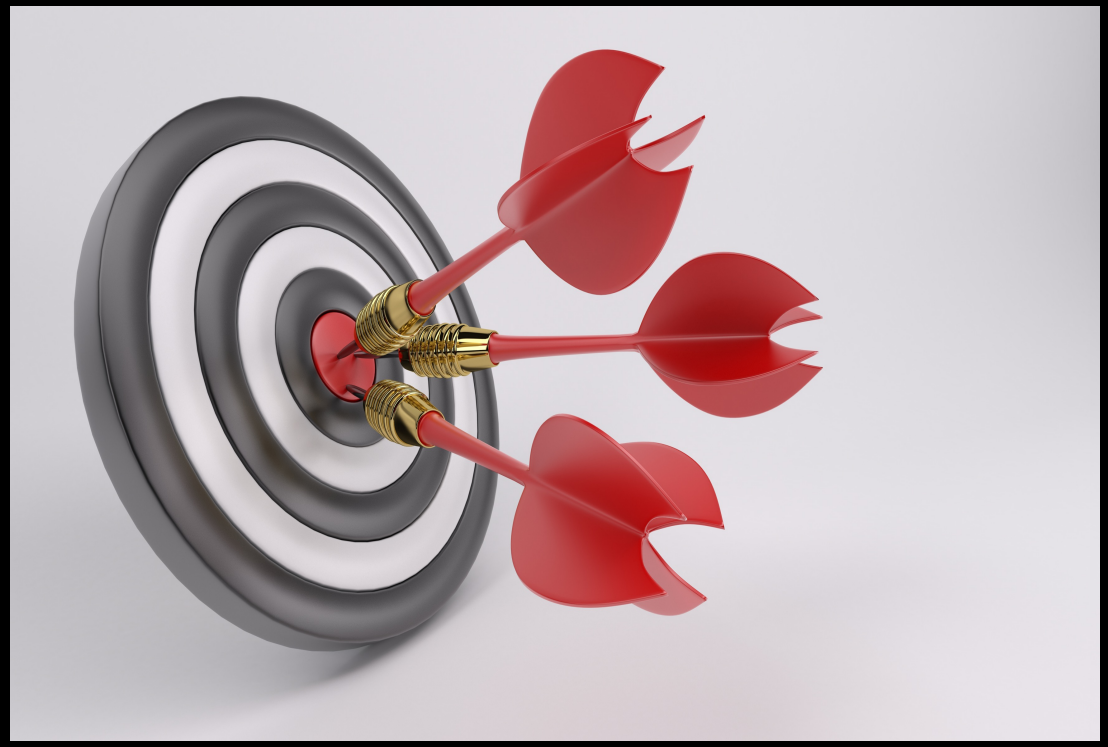
Key insights from ROC/AUC Curve (Optimized Model)

1. Higher the **AUC**, the better the model is at predicting 2 classes like 2 and 4 classes as 4.

2. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.



Recommendations



Based on the outcome of the model, the followings are the **recommendations** which can be furnished to Mr. John Hughes:

- To enhance the model performance, more data can be requested
- Perform with another set of algorithms
- Adapt the methodology of transformation

Note: Python HTML file is attached for reference