|  |
| --- |
| Breast Cancer Detection |

You belong to the data team at a local research hospital. You've been tasked with developing a means to help doctors diagnose breast cancer. You've been given data about biopsied breast cells; where it is benign (not harmful) or malignant (cancerous).

* What features of a cell are the largest drivers of malignancy? Build a model that predicts whether a given biopsied breast cell is benign or malignant.
* What features drive your false positive rate for your model you derived above, what features drive your false negative rate?
* How would a physician use your product?
* There is a non-zero cost in time and money to collect each feature about a given cell. How would you go about determining the most cost-effective method of detecting malignancy?

In addition to well-documented, extendable, and reusable code we are looking for a coherent data-story. This is a deliberately open-ended question that provides a chance to showcase your EDA, analysis, and presentation skills. The expectation is that you should spend no more than 4 hours on this task, and no more than 1 one preparing a presentation.

|  |  |
| --- | --- |
| Data | |
| Name | Range or Description |
| Sample code number | id number |
| Clump Thickness | 1 - 10 |
| Uniformity of Cell Size | 1 - 10 |
| Uniformity of Cell Shape | 1 - 10 |
| Marginal Adhesion | 1 - 10 |
| Single Epithelial Cell Size | 1 - 10 |
| Bare Nuclei | 1 - 10 |
| Bland Chromatin | 1 - 10 |
| Normal Nucleoli | 1 - 10 |
| Mitoses | 1 - 10 |
| Class | (4 for benign, 2 for malignant) |