Breast Cancer Detection

By: Jason Leonard (Project Lead), Mamata Khadka, Susmitha Abbaiahvari

Dataset Context

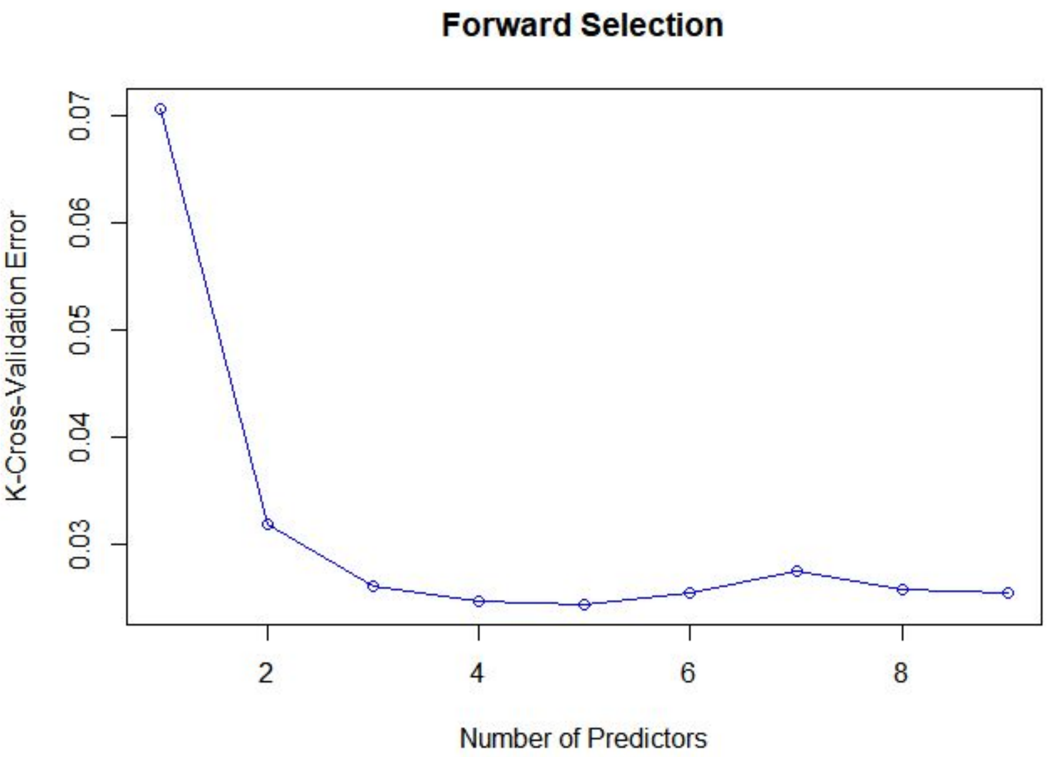
- Consists of 683 rows/observations **(*n* = 683)**
- 9 predictor variables **(*p* = 9)**
- Outcome Variable - ***Class***
  - Benign
  - Malignant
  - Type : categorical

Model Selection

Forward Selection

	Clump.Thickness	Uniformity.of.Cell.Size	Uniformity.of.Cell.Shape	Marginal.Adhesion	Single.Epithelial.Cell.Size	Bare.Nuclei	Bland.Chromatin	Normal.Nucleoli	Mitoses
1						x			
2		x				x			
3	x	x				x			
4	x	x				x		x	
5	x	x				x	x	x	
6	x	x	x			x	x	x	
7	x	x	x	x		x	x	x	
8	x	x	x	x	x	x	x	x	
9	x	x	x	x	x	x	x	x	x

Forward Selection

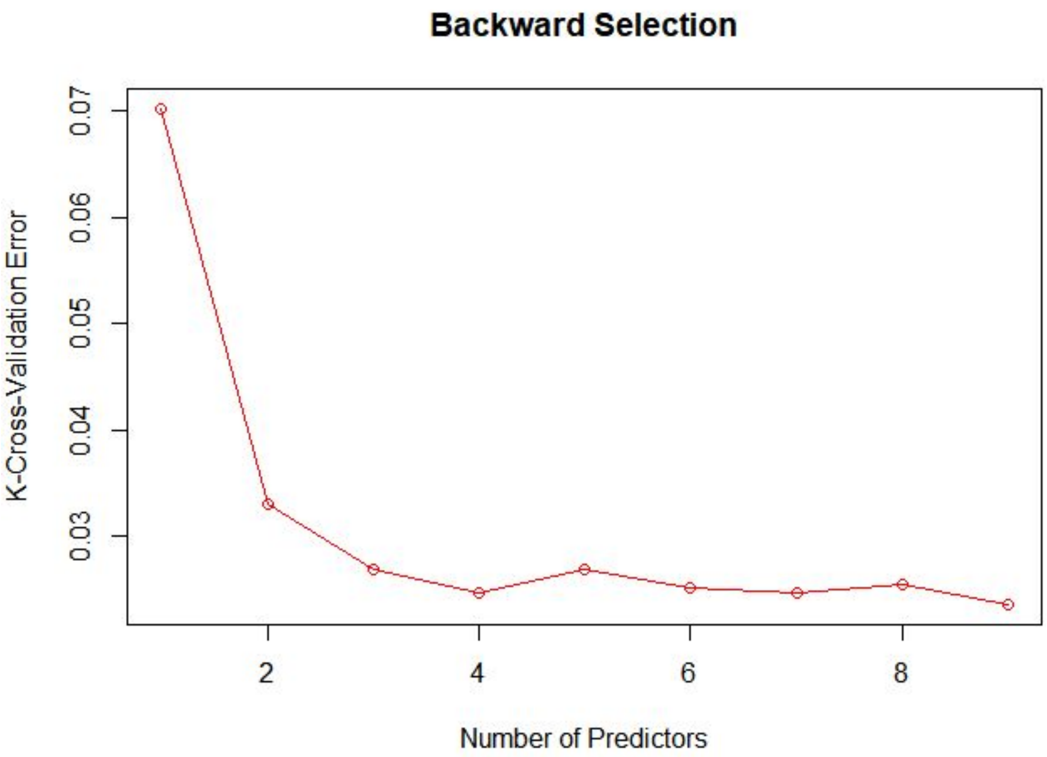


Number of Predictors	Error Rate
1	0.07055592
2	0.03190405
3	0.02612898
4	0.02461881
5	0.02440758
6	0.02540416
7	0.02743437
8	0.02581996
9	0.02550862

Backward Selection

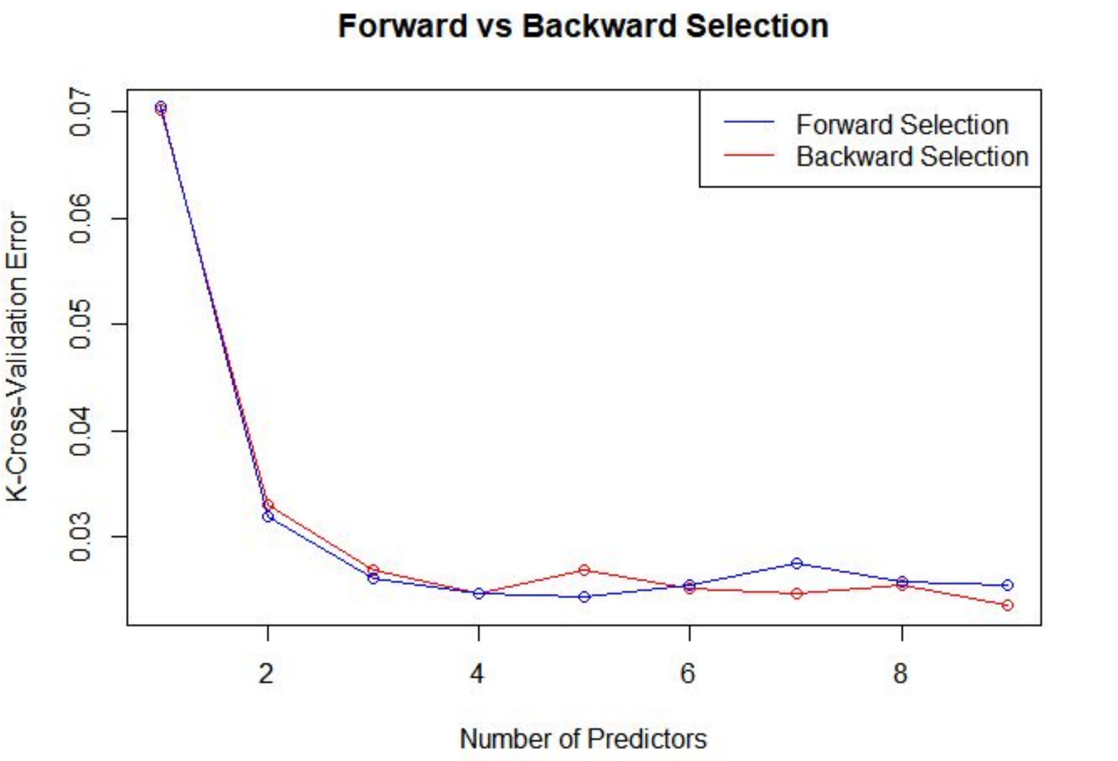
	Clump.Thickness	Uniformity.of.Cell.Size	Uniformity.of.Cell.Shape	Marginal.Adhesion	Single.Epithelial.Cell.Size	Bare.Nuclei	Bland.Chromatin	Normal.Nucleoli	Mitoses
1						x			
2		x				x			
3	x	x				x			
4	x	x				x		x	
5	x	x				x	x	x	
6	x	x	x			x	x	x	
7	x	x	x	x		x	x	x	
8	x	x	x	x	x	x	x	x	
9	x	x	x	x	x	x	x	x	x

Backward Selection



Number of Predictors	Error Rate
1	0.07018557
2	0.03304954
3	0.02694649
4	0.02463429
5	0.02686422
6	0.02508366
7	0.02467253
8	0.02547250
9	0.02361484

Forward vs Backward Selection



Ridge Regression

- **Accuracy:** `0.9692533`
- **Error Rate:** `0.03074671`
- **Optimal Lambda:** `0.05336699`
- **Confusion Matrix:**

	Benign	**Malignant**
Benign	435	12
Malignant	9	227

Lasso

- **Accuracy:** 0.966325
- **Error Rate:** 0.03367496
- **Optimal Lambda:** 0.01204504
- **Confusion Matrix:**

	Benign	**Malignant**
Benign	435	14
Malignant	9	225

Relaxing Linearity

Our Approach

- Experiment with 4 Variable Model from Forward & Backward Selection
- Experiment with All Variable Models
- Obtain Benchmarks from Logistic Regression models
- Use Validation Set approach to limit variability in prediction results
    - 70-30 Training Testing Split

4 Variable Model - Benchmark

- **Logistic Regression Model w/ 4 Variables**
  - Bare.Nuclei, Uniformity.of.Cell.Size, Clump.Thickness, and Normal.Nucleoli
- **Accuracy**: 0.9804878
- **Error Rate**: 0.0195122
- **Confusion Matrix**:

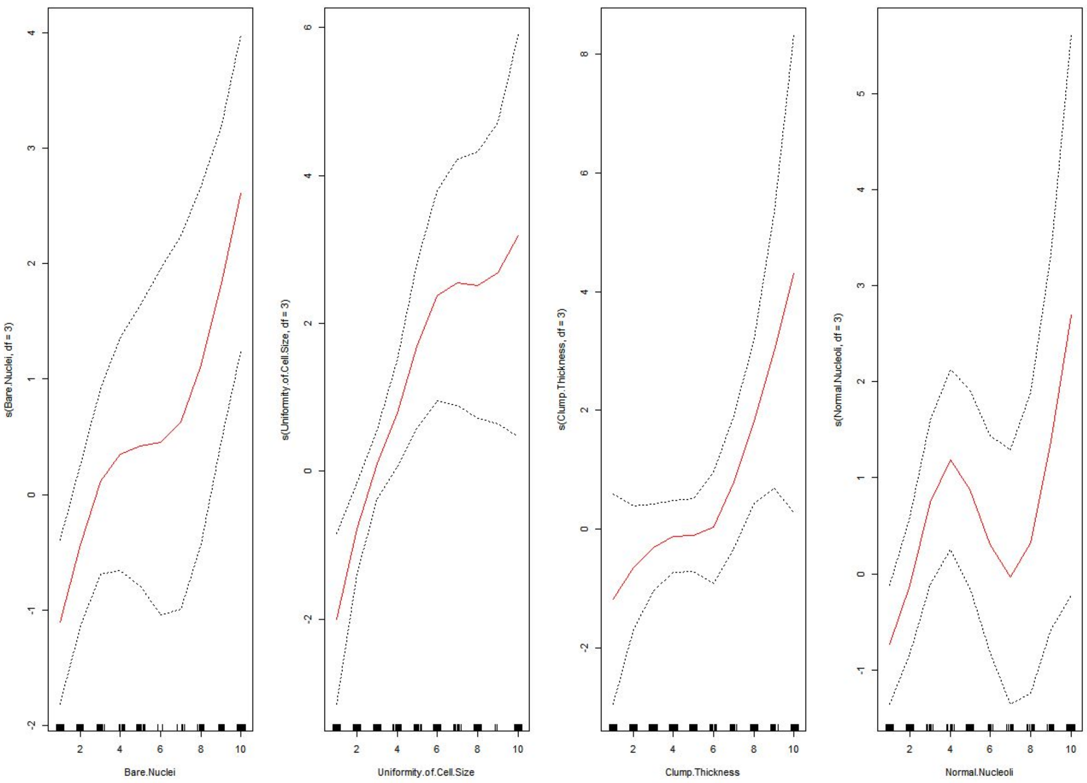
	Benign	**Malignant**
Benign	131	2
Malignant	2	70

4 Variable Model - All Variables w/ Splines

- **All 4 Variables have Splines w/ 3 Degrees of Freedom**
- **Accuracy**: `0.9707317`
- **Error Rate**: `0.02926829`
- **Confusion Matrix**:

	Benign	**Malignant**
Benign	131	4
Malignant	2	68

4 Variable Model - All Variables w/ Splines

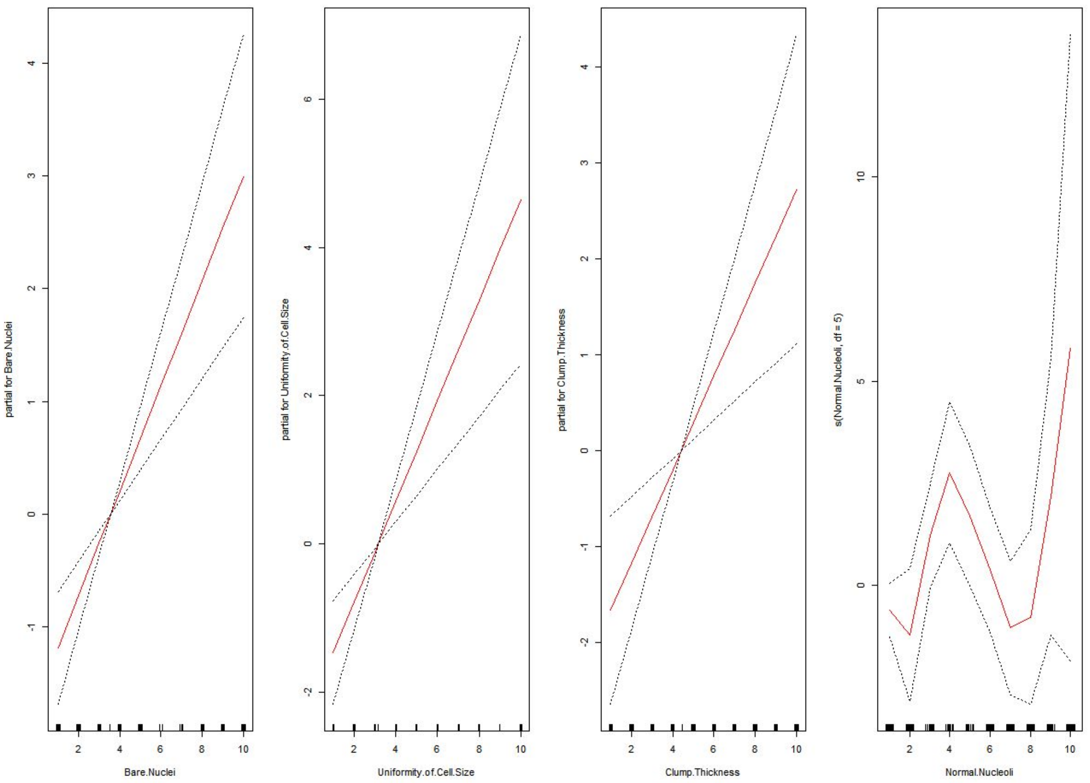


4 Variable Model - Some Variables w/ Splines

- **Only** `Normal.Nucleoli` **has a Spline w/ 5 Degrees of Freedom**
- **Accuracy**: `0.9853659`
- **Error Rate**: `0.01463415`
- **Confusion Matrix**:

	Benign	**Malignant**
Benign	131	1
Malignant	2	71

4 Variable Model - Some Variables w/ Splines



9 Variable Model - Benchmark

- **Logistic Regression Model w/ ALL variables**
- **Accuracy**: `0.9609756`
- **Error Rate**: `0.03902439`
- **Confusion Matrix**:

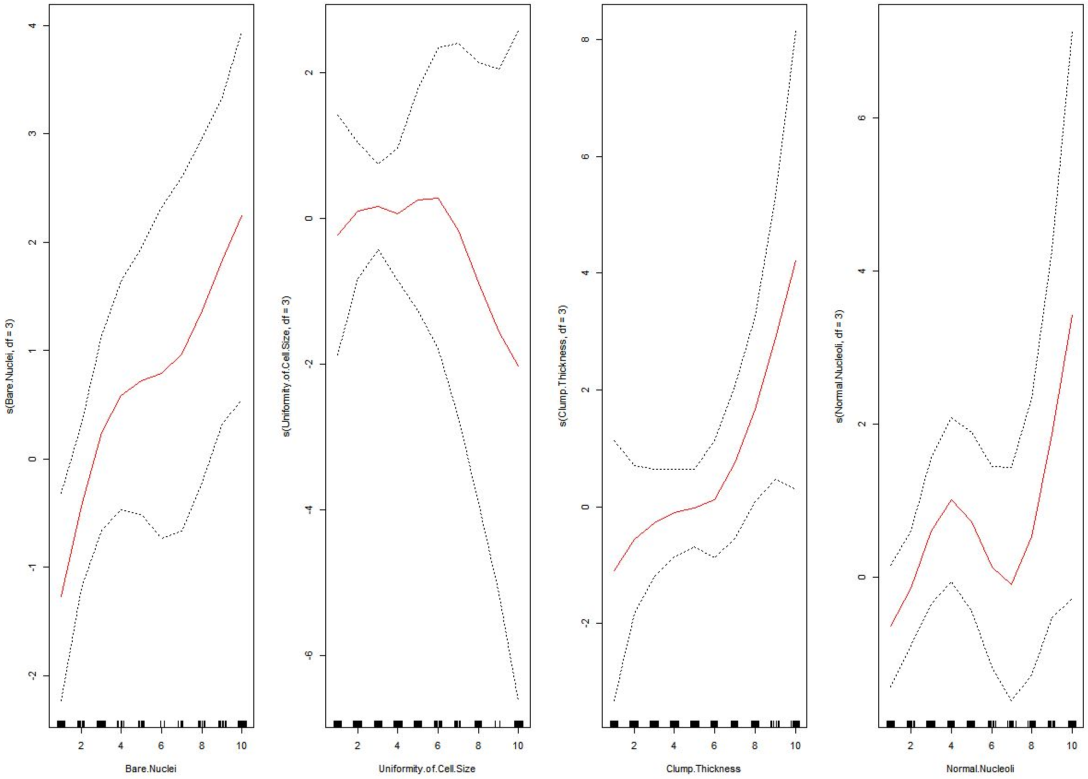
	Benign	**Malignant**
Benign	131	6
Malignant	2	66

9 Variable Model - All Variables w/ Splines

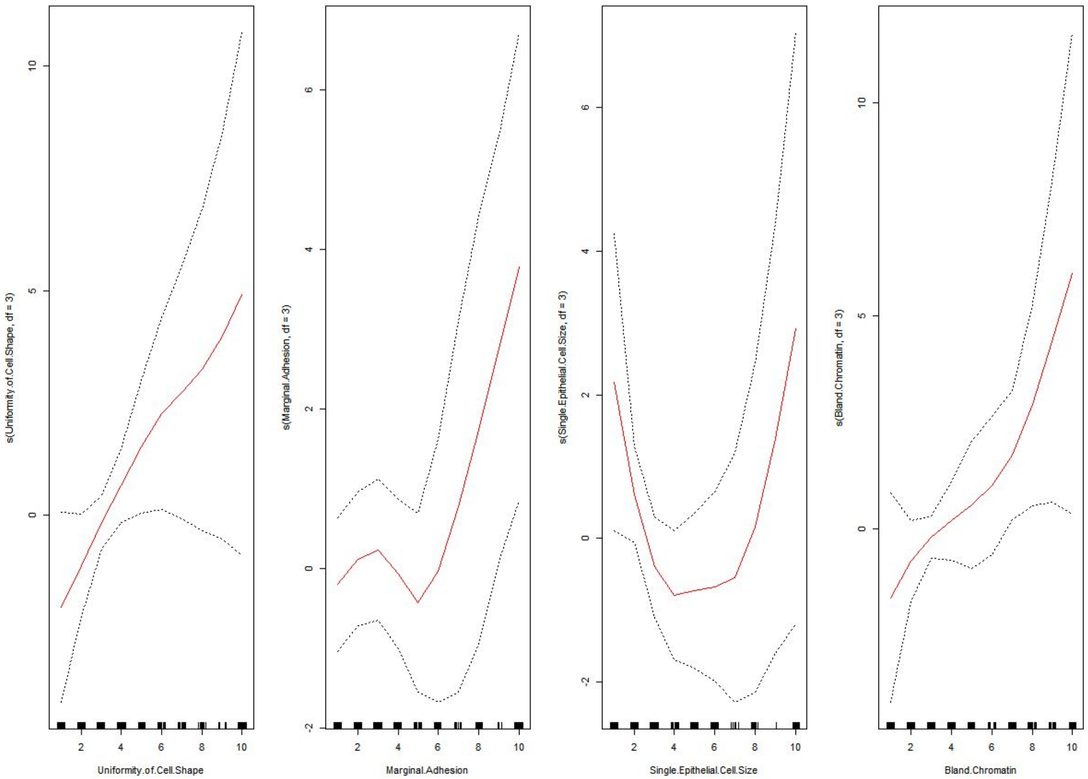
- **ALL Variables have Splines w/ 3 Degrees of Freedom**
- **Accuracy**: `0.9658537`
- **Error Rate**: `0.03414634`
- **Confusion Matrix**:

	Benign	**Malignant**
Benign	131	5
Malignant	2	67

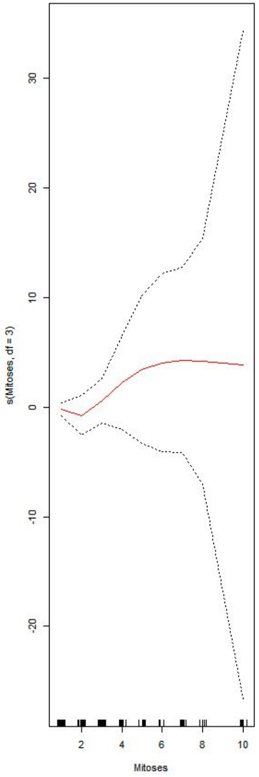
9 Variable Model - All Variables w/ Splines



9 Variable Model - All Variables w/ Splines



9 Variable Model - All Variables w/ Splines

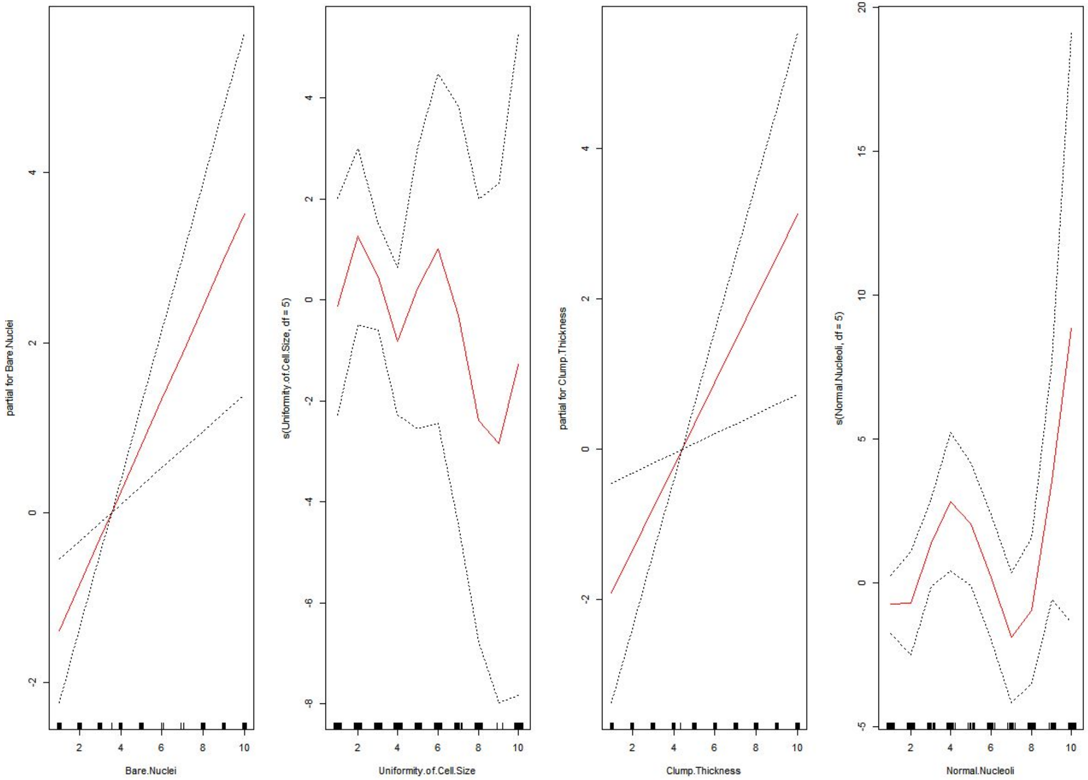


9 Variable Model - Some Variables w/ Splines

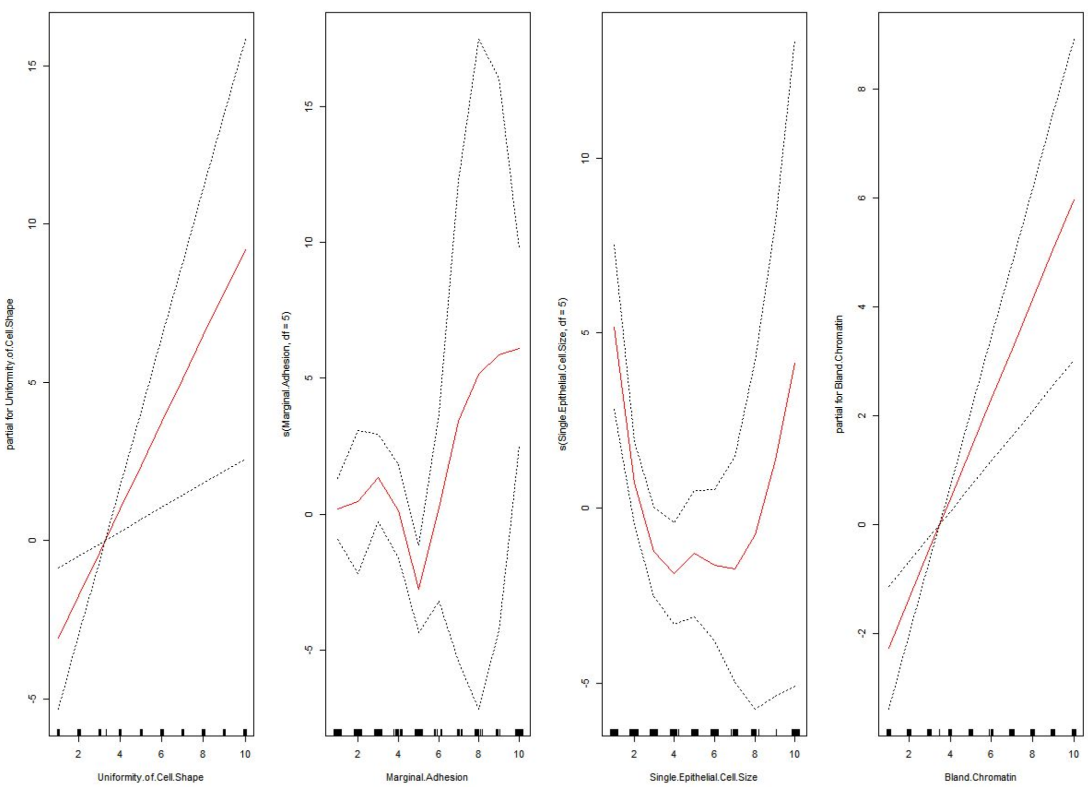
- **Some Variables Splines w/ 5 Degrees of Freedom**
- **Accuracy**: `0.9560976`
- **Error Rate**: `0.04390244`
- **Confusion Matrix**:

	Benign	**Malignant**
Benign	127	3
Malignant	6	69

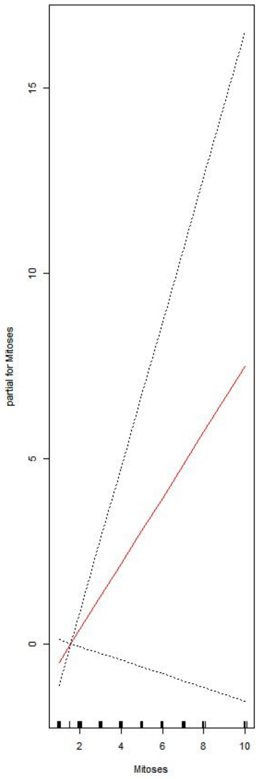
9 Variable Model - Some Variables w/ Splines



9 Variable Model - Some Variables w/ Splines



9 Variable Model - Some Variables w/ Splines



The Results

- Accuracy Comparison of Generated Models

	4 Variable Model Accuracy	9 Variable Model Accuracy
Benchmark Model (Logistic Regression)	0.9804878	0.9609756
Model w/ Splines for all Predictors	0.9707317	0.9658537
Model w/ Splines for Some Predictors	0.9853659	0.9560976

Questions?