Cervical Cancer Risk Classification

Dataset from Kaggle

Erica Chio

Background

Cervical Cancer

- A malignant tumor of the lower-most part of the uterus
- Can be prevented with HPV vaccine and PAP smear screening
- Around 11,000 new cases each year in the US
- 4,000 deaths in the US; 300,000 worldwide

Dataset

- Dataset obtained from UCI Repository
- Contains a list of risk factors for cervical cancer that leads to a biopsy examination

Project Idea

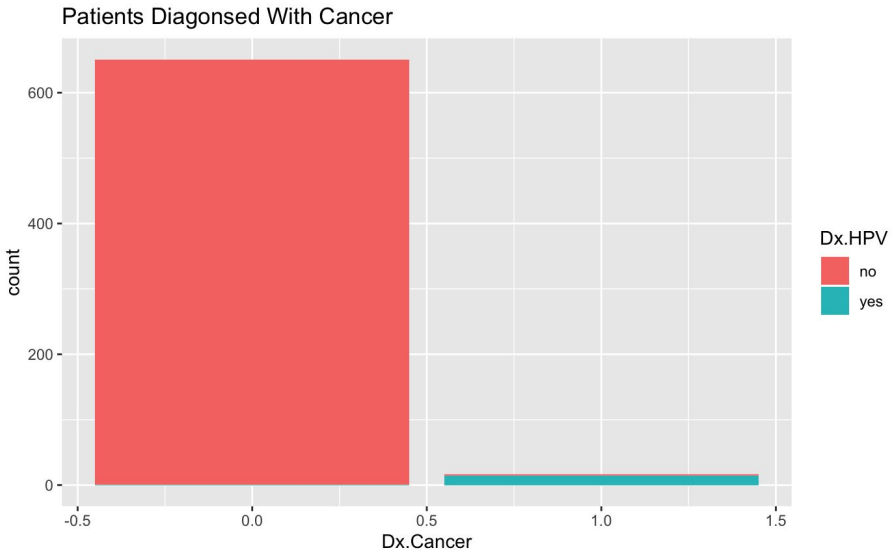
- ~600 observations, 38 variables
- Goal: Create a Model to predict the likelihood of getting tested cervical cancer
- Variables of Interest:
  - DxCancer (diagnosis of cancer)
  - DxHPV
  - Hinselmann
  - Schiller
  - Citology
  - Biopsy

Imbalance of Data (Use of SMOTE)

Options to Fix Imbalance:

- Undersampling
  - Reduce majority class
- Oversampling
  - Replicated minority class
  - No information loss

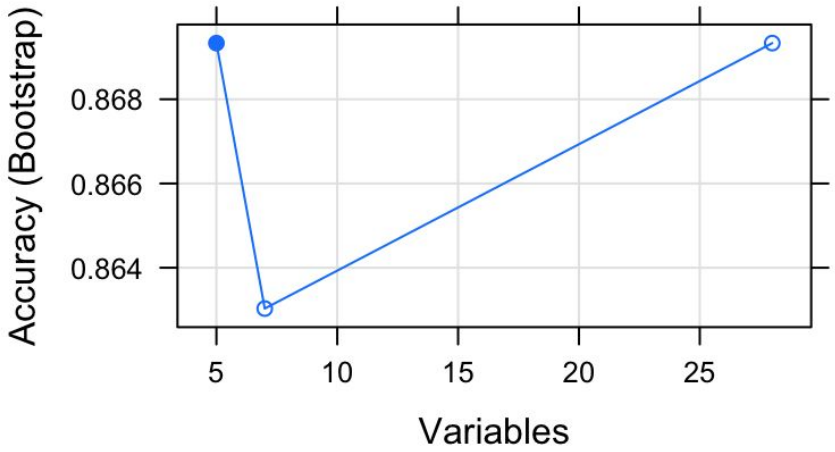
SMOTE - Synthetic Minority Oversampling Technique



Feature Selection (RFE)

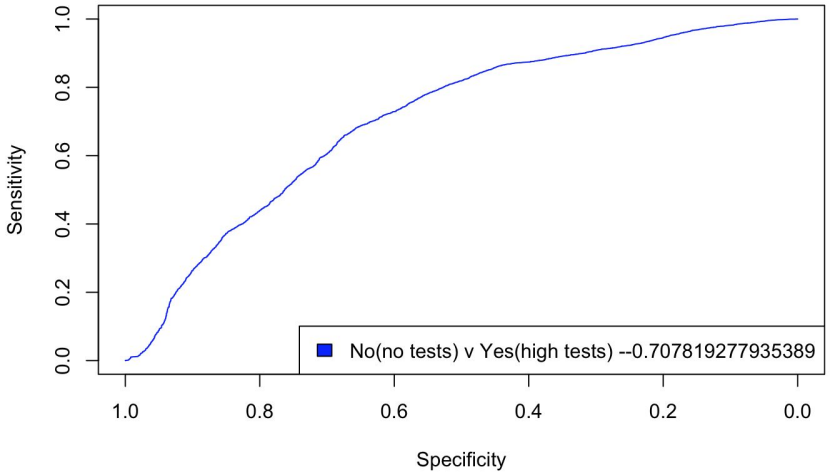
Features Selected (top 5):

1. SmokePacksPerYear
2. SmokeYears
3. Age
4. Smokes
5. HormonalContraceptivesYears

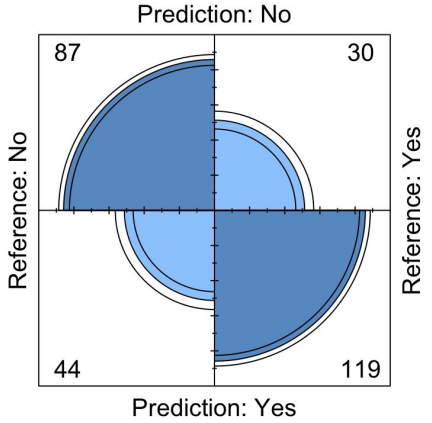


Logistic Regression with TOP 5

```
               Accuracy : 0.7357
                 95% CI : (0.68, 0.7864)
    No Information Rate : 0.5321
    P-Value [Acc > NIR] : 2.139e-12

                  Kappa : 0.4658

 Mcnemar's Test P-Value : 0.1307

            Sensitivity : 0.6641
            Specificity : 0.7987
         Pos Pred Value : 0.7436
         Neg Pred Value : 0.7301
             Prevalence : 0.4679
         Detection Rate : 0.3107
   Detection Prevalence : 0.4179
      Balanced Accuracy : 0.7314
```

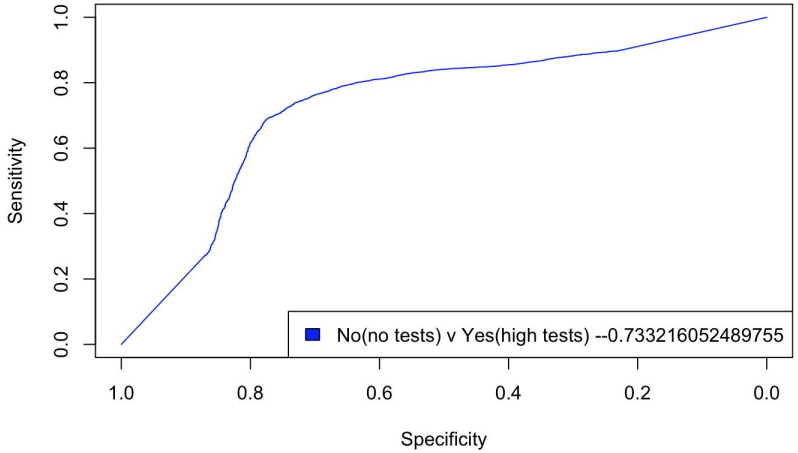


Confusion Matrix for Logistic Regression (top 5 features)

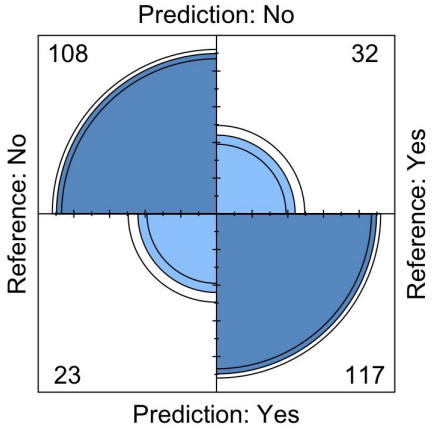


Logistic Regression with ALL features

```
                Accuracy : 0.8036
                  95% CI : (0.7521, 0.8485)
    No Information Rate : 0.5321
    P-Value [Acc > NIR] : <2e-16

                   Kappa : 0.6071

 Mcnemar's Test P-Value : 0.2807

            Sensitivity : 0.8244
            Specificity : 0.7852
         Pos Pred Value : 0.7714
         Neg Pred Value : 0.8357
             Prevalence : 0.4679
         Detection Rate : 0.3857
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.8048
```

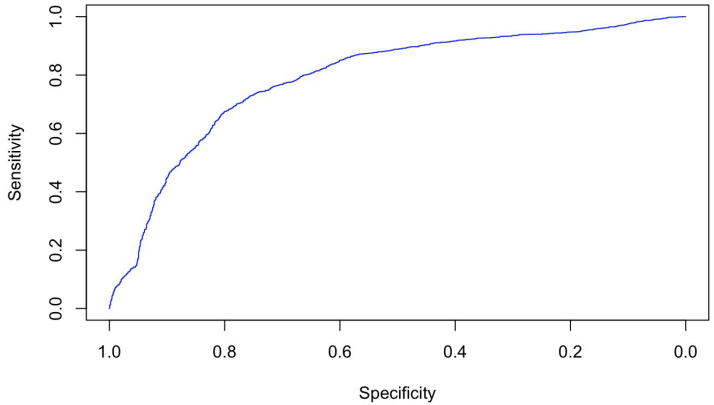


Confusion Matrix for Logistic Regression (ALL features)



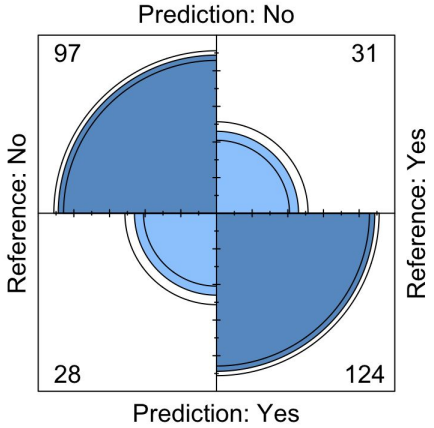
SVM

```
               Accuracy : 0.7357
                 95% CI : (0.68, 0.7864)
    No Information Rate : 0.5321
    P-Value [Acc > NIR] : 2.139e-12

                  Kappa : 0.4658

 Mcnemar's Test P-Value : 0.1307

            Sensitivity : 0.6641
            Specificity : 0.7987
         Pos Pred Value : 0.7436
         Neg Pred Value : 0.7301
             Prevalence : 0.4679
         Detection Rate : 0.3107
   Detection Prevalence : 0.4179
      Balanced Accuracy : 0.7314
```

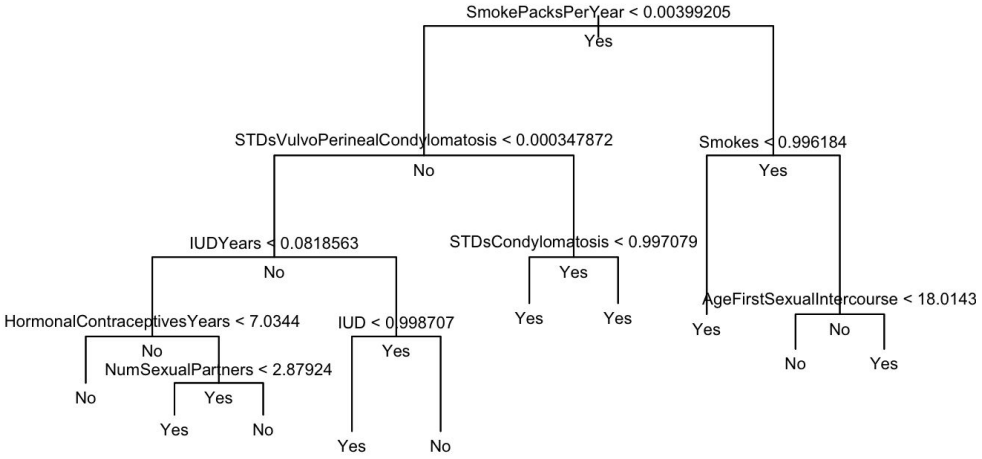


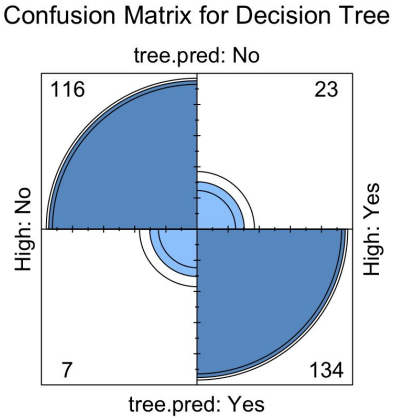
Area under ROC curve: **0.79**

Confusion Matrix for Logistic Regression (ALL features)



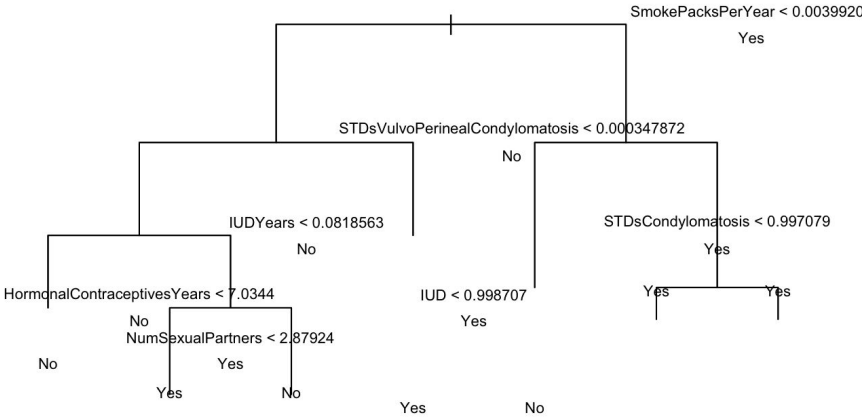
Classification Tree



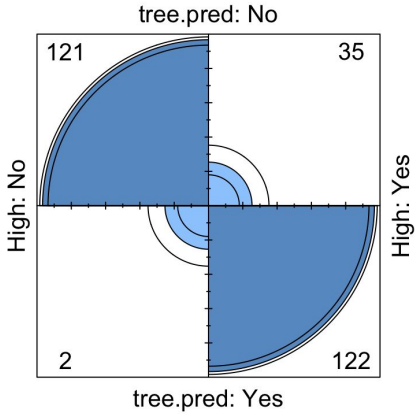



```
               Accuracy : 0.8929
                 95% CI : (0.8506, 0.9265)
    No Information Rate : 0.5607
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7855

 Mcnemar's Test P-Value : 0.00617

            Sensitivity : 0.9431
            Specificity : 0.8535
         Pos Pred Value : 0.8345
         Neg Pred Value : 0.9504
             Prevalence : 0.4393
         Detection Rate : 0.4143
   Detection Prevalence : 0.4964
      Balanced Accuracy : 0.8983
```

Classification Tree (after pruning)



Confusion Matrix for Random Forest - Smaller Forest




```
               Accuracy : 0.8679
                 95% CI : (0.8225, 0.9052)
    No Information Rate : 0.5607
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7393

 Mcnemar's Test P-Value : 1.435e-07

            Sensitivity : 0.9837
            Specificity : 0.7771
         Pos Pred Value : 0.7756
         Neg Pred Value : 0.9839
             Prevalence : 0.4393
         Detection Rate : 0.4321
   Detection Prevalence : 0.5571
      Balanced Accuracy : 0.8804
```

In Conclusion,

- **Classification Tree** before Pruning was the best model with a balanced accuracy of **0.89**

- Balancing the dataset was very important
  - Prior to balancing, models would only predict one class ➔ “50%” accuracy but didn’t learn anything
  - After oversampling, models are predicting up to ~ 80% accuracy