Predicting Cases of Cervical Cancer in the 'Hospital Universitario de Caracas'

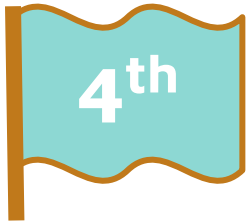
Ben Herndon-Miller and Nicole Jaiyesimi
30 November 2018

Outline

1. *Background & Motivation*
2. *Our Data*
    a. *Overview*
    b. *Cleaning*
    c. *Exploratory Data Analysis*
3. *Predictive Modeling*

Overview of Cervical Cancer

Cervical Cancer Worldwide



Most common cancer among women worldwide

570,000

New cases in 2018

6.6%

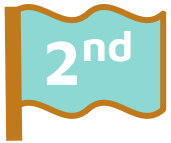
Most common cancer among women worldwide

90% of deaths occured in low & middle income countries

Cervical Cancer in Venezuela

4973 women diagnosed each year, **1789** die each year

2.3%: U.S. mortality rate **vs.** **11.5%:** Venezuela mortality rate



Most frequent cancer among Venezuelan women

Most frequent among women between 15 & 44 years old in Venezuela

Motivation

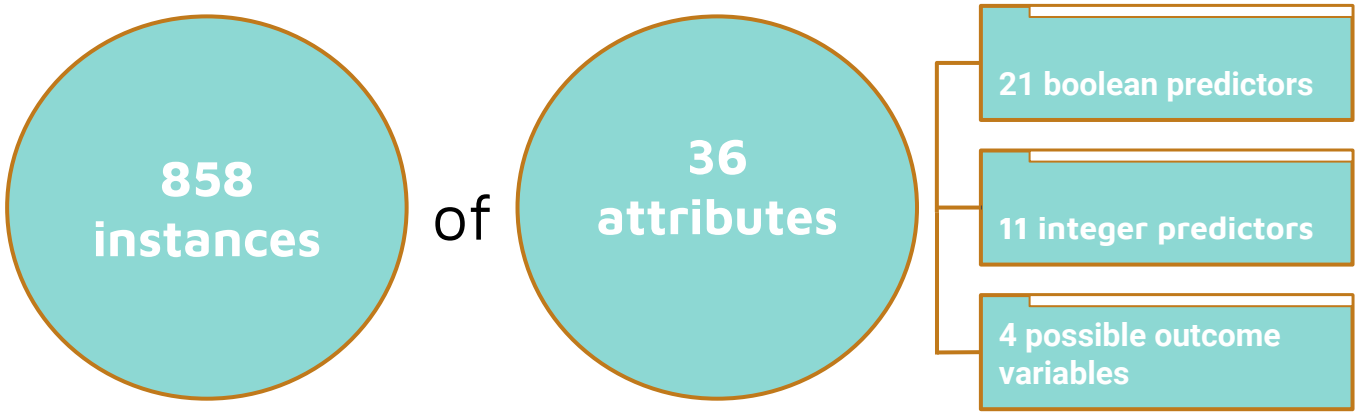
"The high mortality rate from cervical cancer globally could be reduced through a comprehensive approach that includes prevention, early diagnosis, effective screening and treatment programmes."

-World Health Organization

Our Data

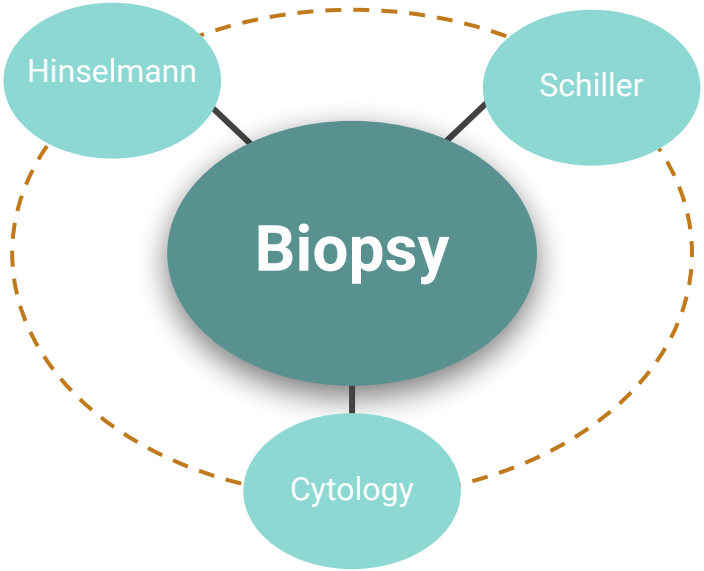
Data Overview

- Hospital Universitario - *Caracas, Venezuela* - 2017



Choosing an Outcome Variable

4 variables that could be considered the outcome variable of interest for a diagnosis of cancer:

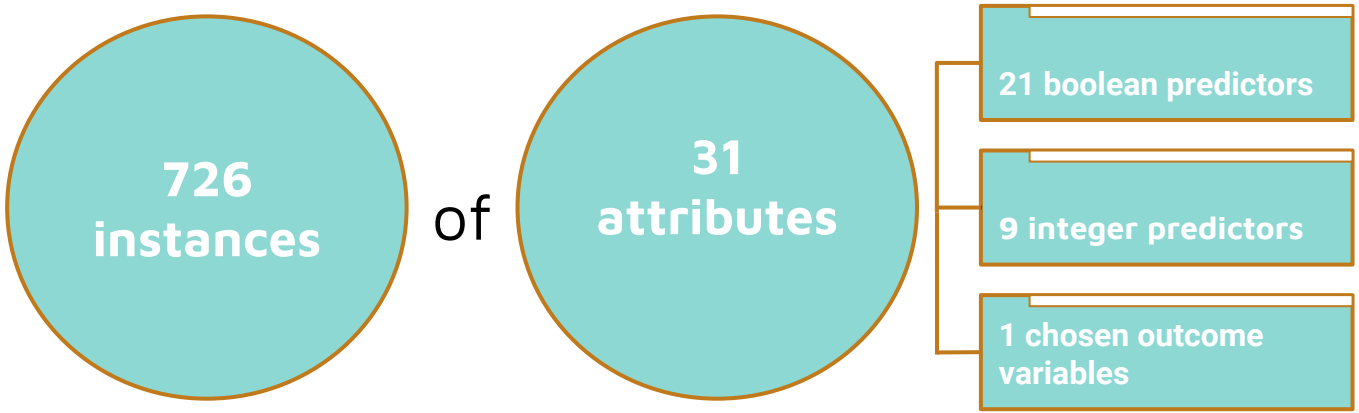


Data Coverage

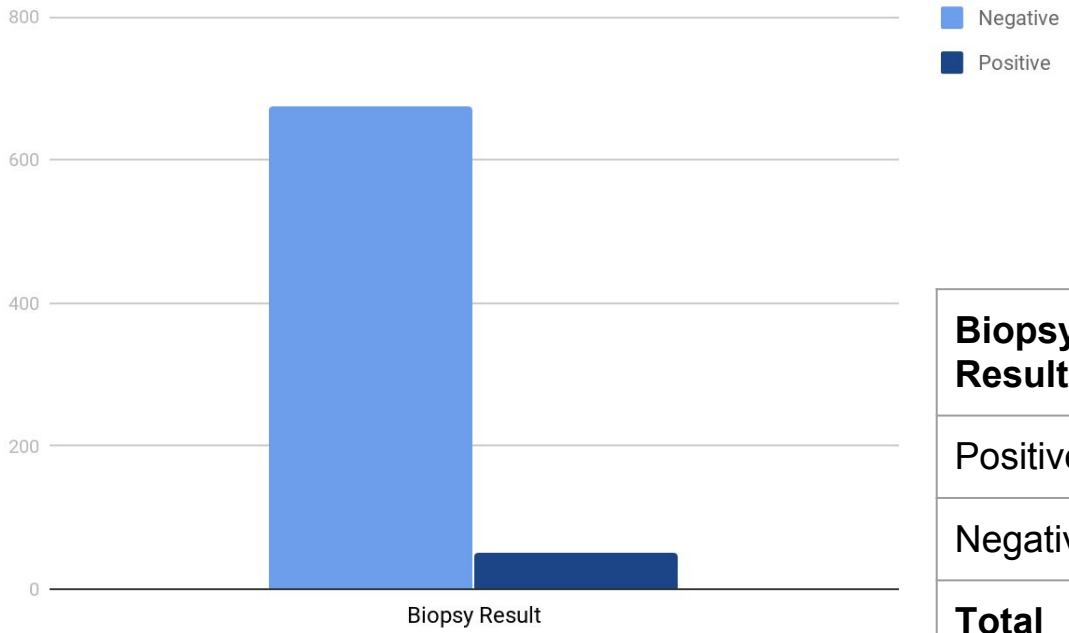
Variable	Variable_Type	Missing_Values	Percent_Missing
Age	int	0	0.0000000
Number.of.sexual.partners	int	26	0.0303030
First.sexual.intercourse	int	7	0.0081585
Num.of.pregnancies	int	56	0.0652681
Smokes	bool	13	0.0151515
Smokes..years.	int	13	0.0151515
Smokes..packs.year.	int	13	0.0151515
Hormonal.Contraceptives	bool	108	0.1258741
Hormonal.Contraceptives..years.	int	108	0.1258741
IUD	bool	117	0.1363636
IUD..years.	int	117	0.1363636
STDs	bool	105	0.1223776
STDs..number.	int	105	0.1223776
STDs.condylomatosis	bool	105	0.1223776
STDs.cervical.condylomatosis	bool	105	0.1223776
STDs.vaginal.condylomatosis	bool	105	0.1223776
STDs.vulvo.perineal.condylomatosis	bool	105	0.1223776
STDs.syphilis	bool	105	0.1223776
STDs.pelvic.inflammatory.disease	bool	105	0.1223776
STDs.genital.herpes	bool	105	0.1223776
STDs.molluscum.contagiosum	bool	105	0.1223776
STDs.AIDS	bool	105	0.1223776

Variable	Variable_Type	Missing_Values	Percent_Missing
STDs.HIV	bool	105	0.1223776
STDs.Hepatitis.B	bool	105	0.1223776
STDs.HPV	bool	105	0.1223776
STDs..Number.of.diagnosis	int	0	0.0000000
STDs..Time.since.first.diagnosis	int	787	0.9172494
STDs..Time.since.last.diagnosis	int	787	0.9172494
Dx.Cancer	bool	0	0.0000000
Dx.CIN	bool	0	0.0000000
Dx.HPV	bool	0	0.0000000
Dx	bool	0	0.0000000
Biopsy	bool	0	0.0000000

Filtered Data

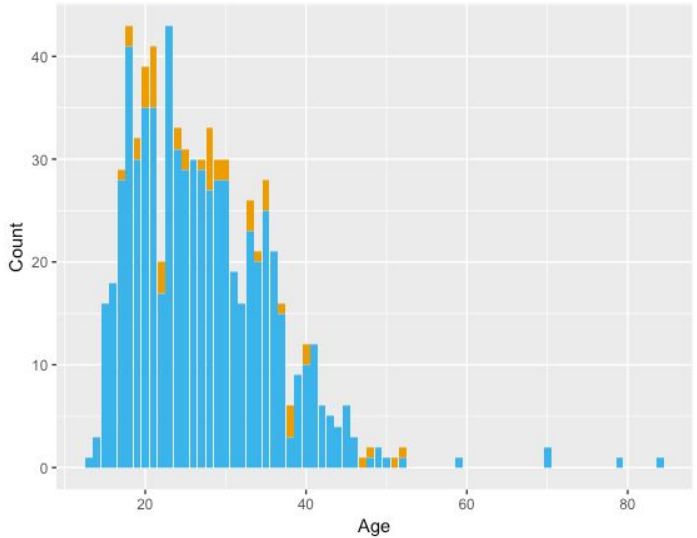


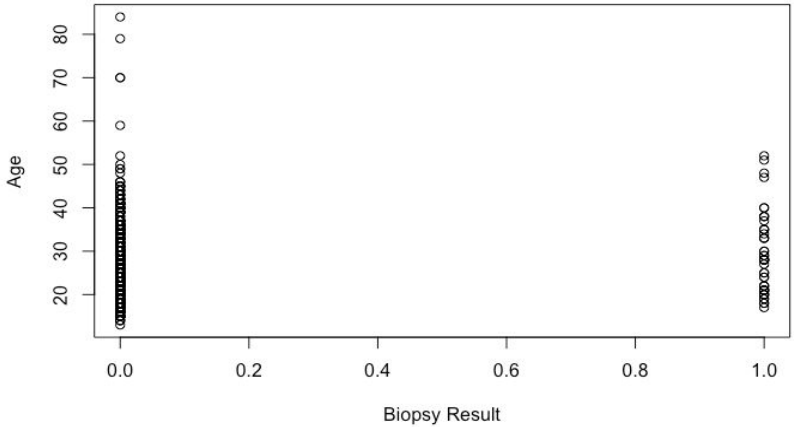
The Unbalanced Outcome Variable



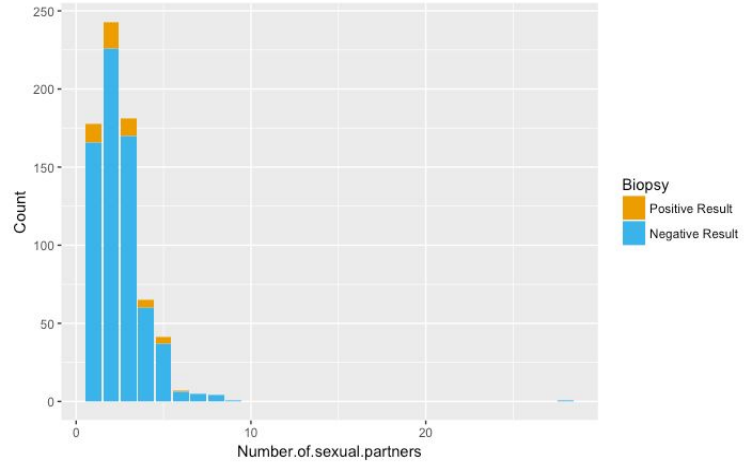
Biopsy Result	Count	Percentage
Positive	50	6.9%
Negative	676	93.1%
Total	726	100%

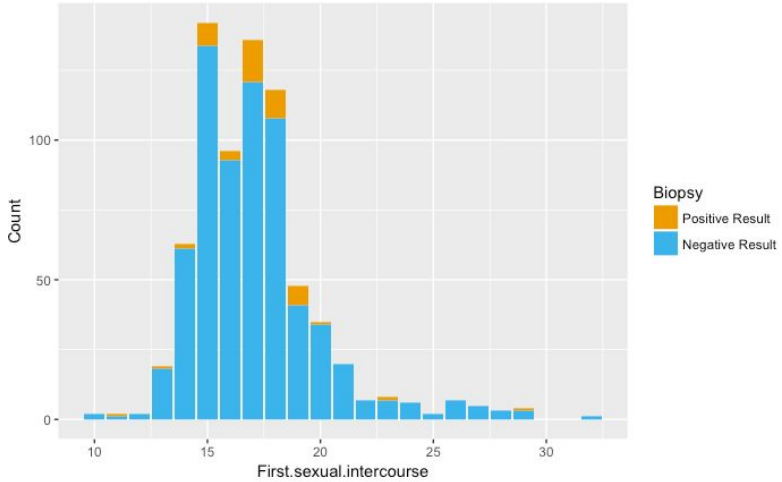
A Look at the Integer Predictors

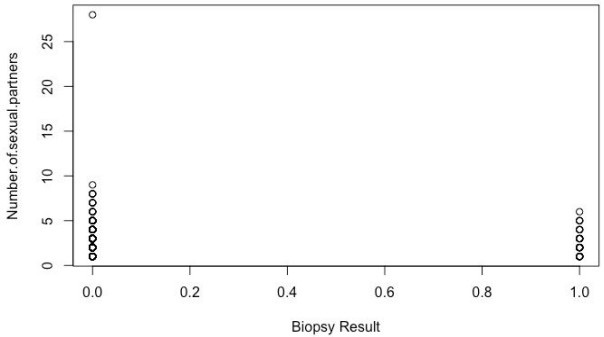


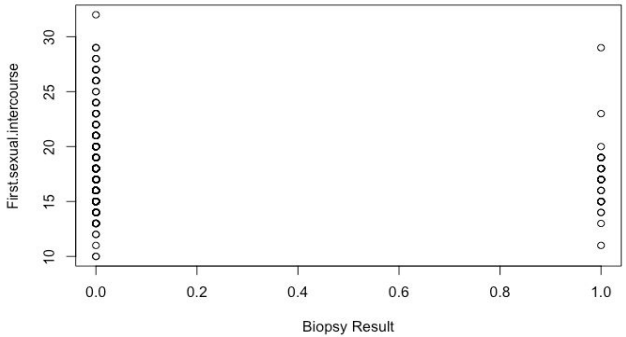


A Look at the Integer Predictors

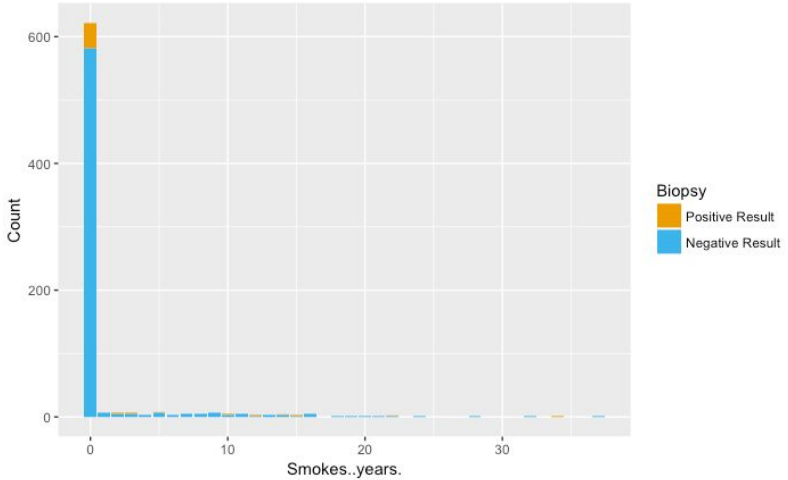


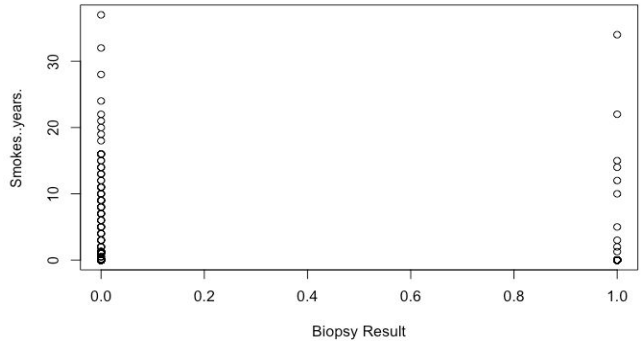


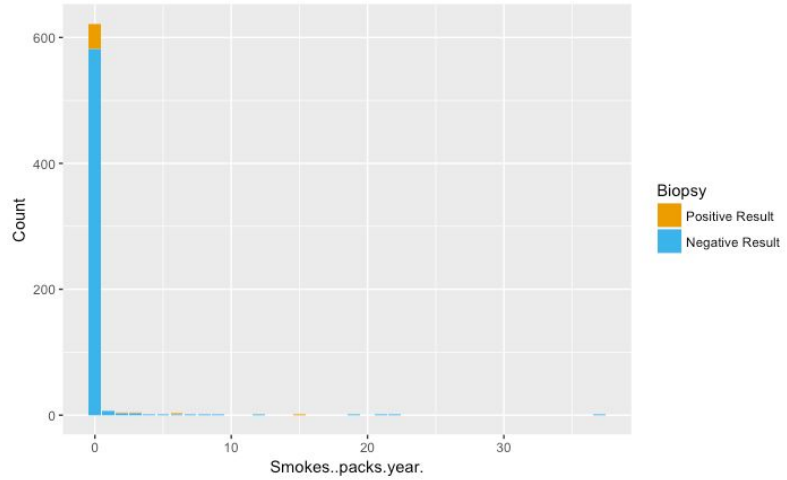


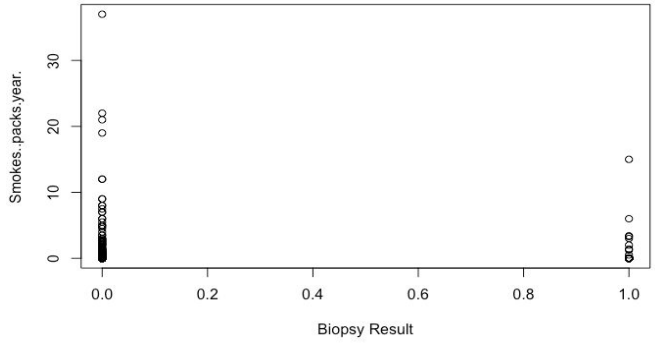


A Look at the Integer Predictors

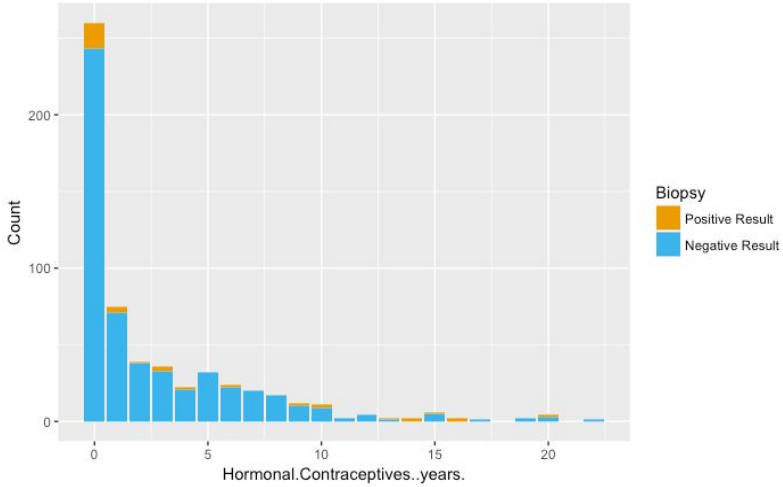


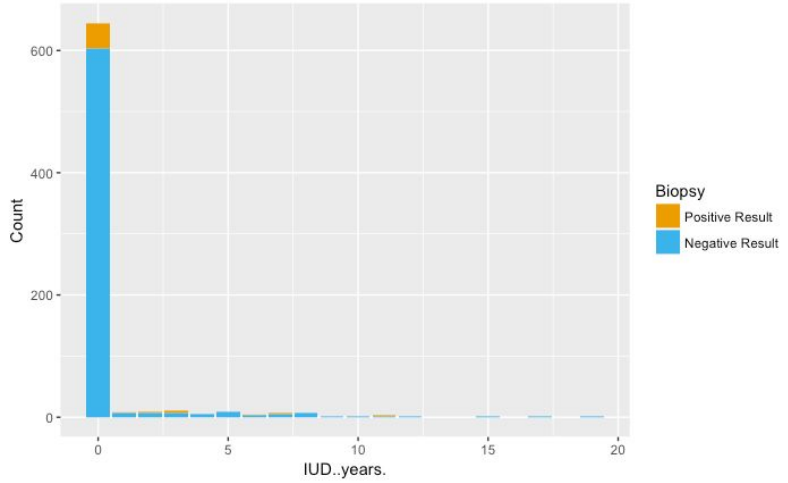


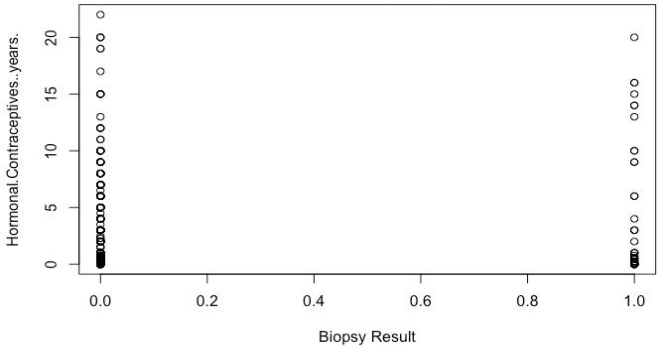


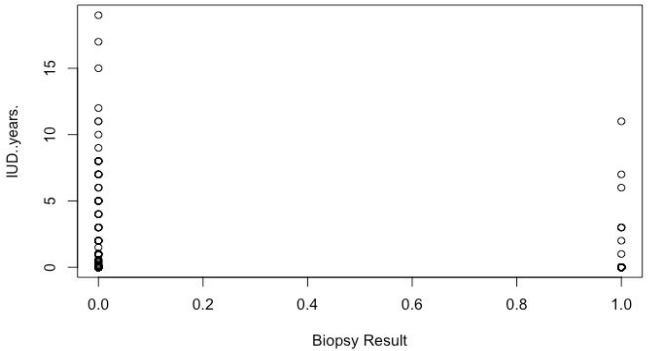


A Look at the Integer Predictors

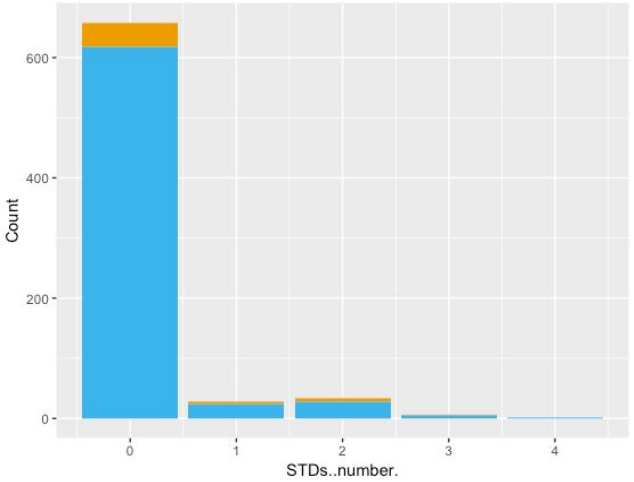


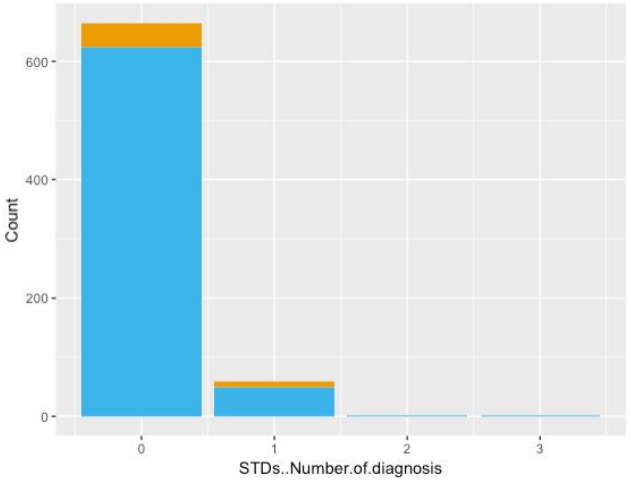


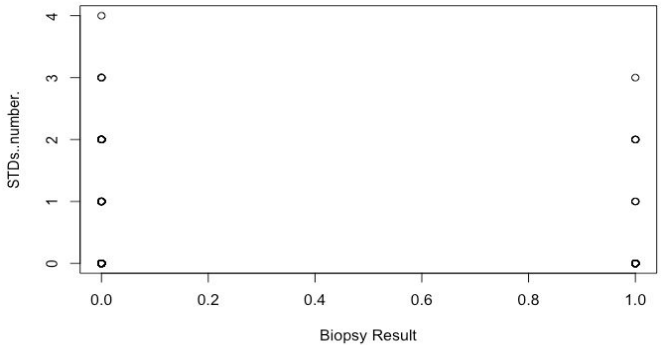


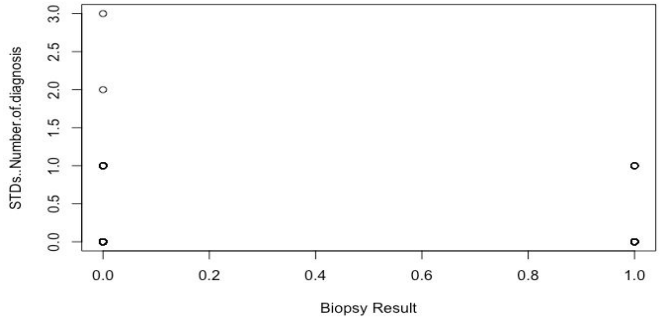


A Look at the Integer Predictors







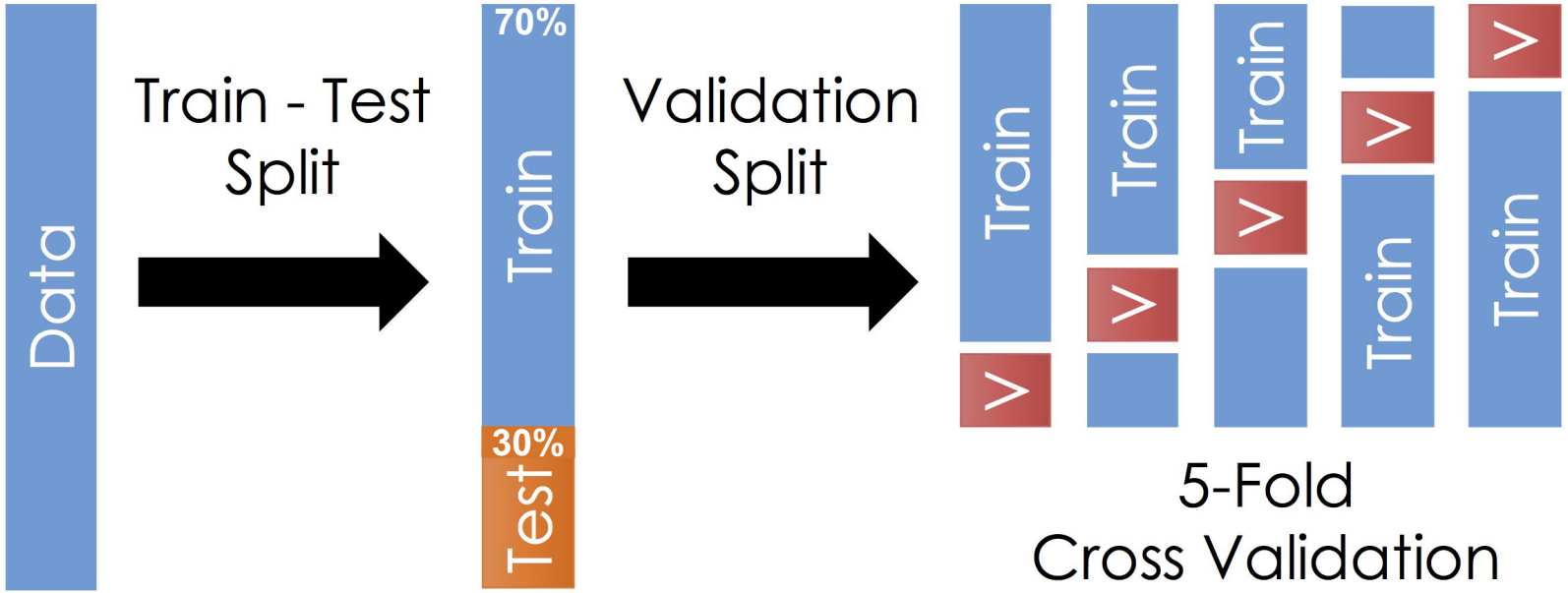


A Look at the Boolean Predictors

Predictor	Count Positive	Percent Positive
Smokes	10	20%
Hormonal.Contraceptives	33	66%
IUD	9	18%
STDs	11	22%
STDs.condylomatosis	7	14%
STDs.cervical.condylomatosis	0	0%
STDs.vaginal.condylomatosis	0	0%
STDs.vulvo.perineal.condylomatosis	7	14%
STDs.syphilis	0	0%
STDs.pelvic.inflammatory.disease	0	0%
STDs.genital.herpes	1	2%
STDs.molluscum.contagiosum	0	0%
STDs.AIDS	0	0%
STDs.HIV	4	8%
STDs.Hepatitis.B	0	0%
STDs.HPV	0	0%
Dx.Cancer	6	12%
Dx.CIN	2	4%
Dx.HPV	6	12%
Dx	6	12%

Predictive Modeling

Data Splitting and Cross-Validation



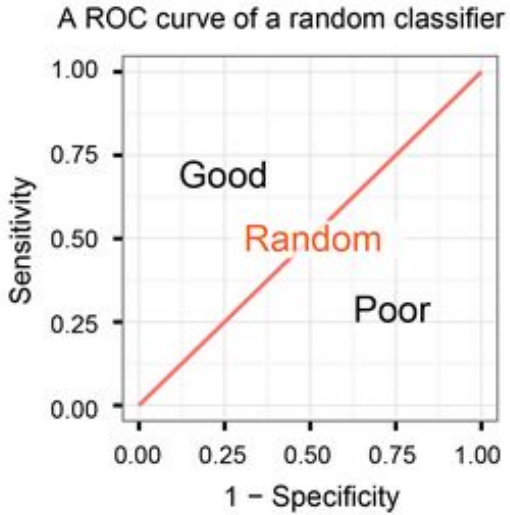
Loss Function: AUC

- *AUC: Area Under the Curve*
- *ROC: Receiver Operator Curve*
  - *Power as a function of Type 1 Error*
  - *Better than MCE for imbalanced data*

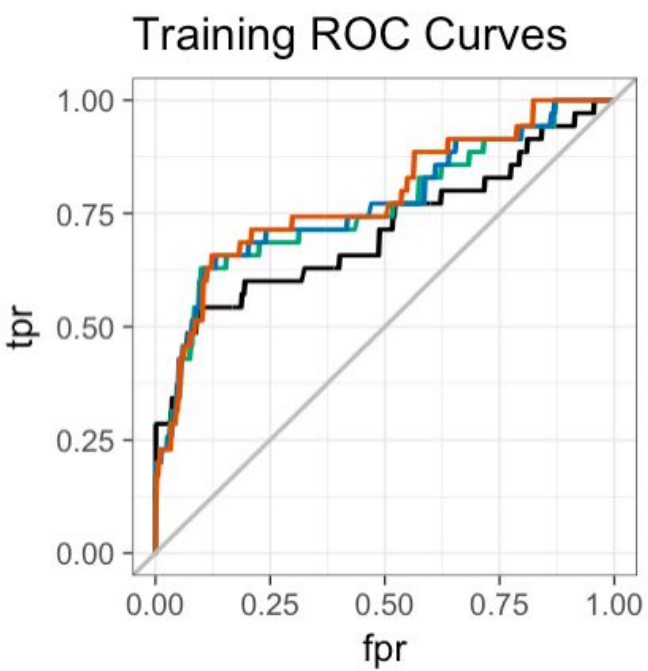
predicted → / real ↓	*Class_pos*	*Class_neg*
Class_pos	TP	FN
Class_neg	FP	TN

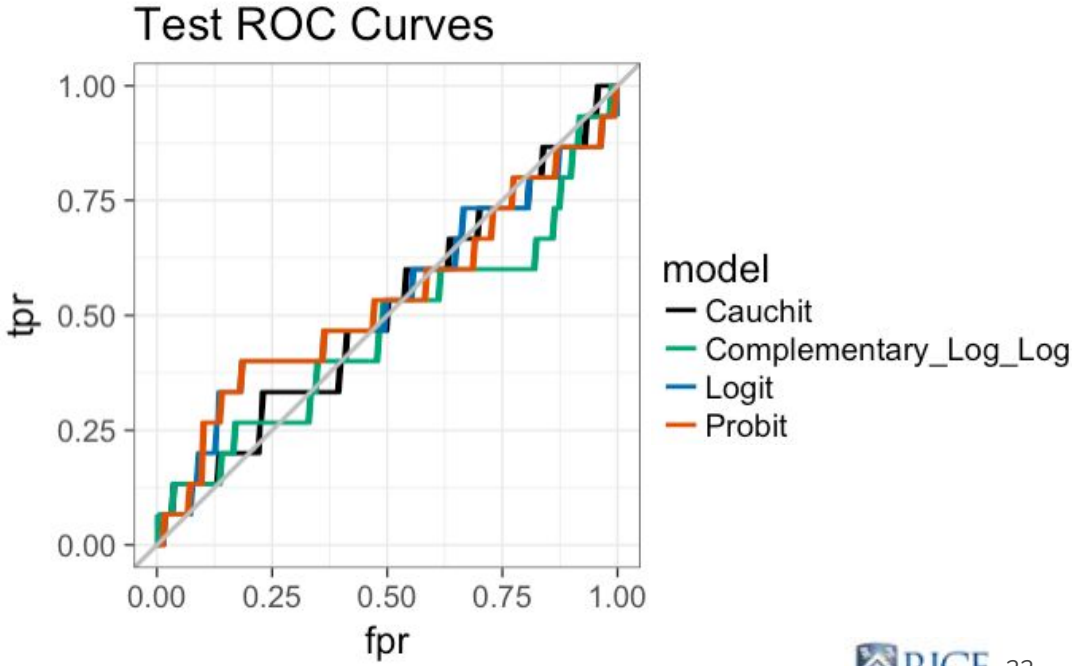
$$\text{TPR (sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR (1-specificity)} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$



Binomial Regression with Different Link Functions

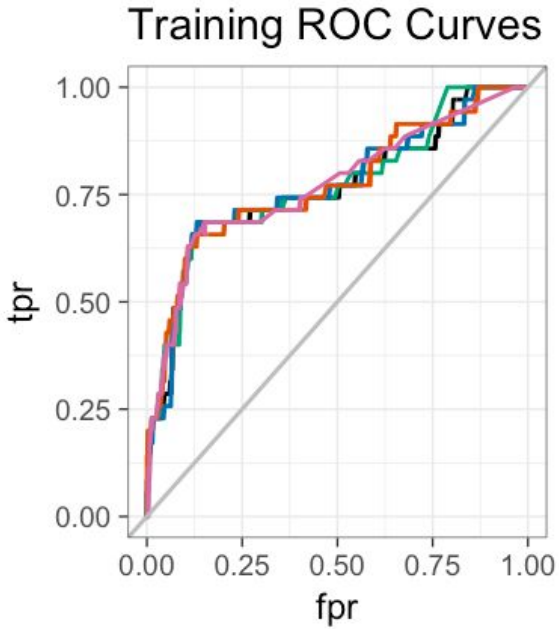


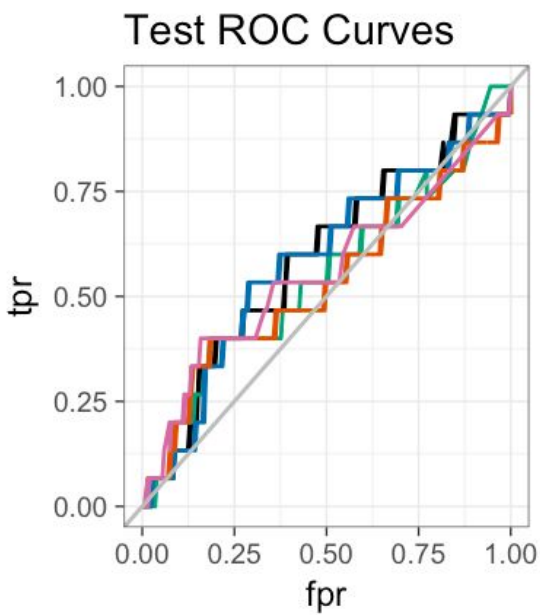


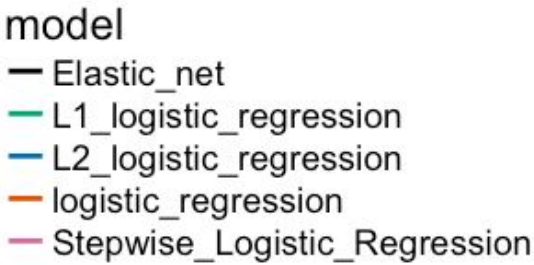
Binomial Regression with Different Link Functions

Method	Train AUC	Test AUC
Logit	0.7719409	0.5339934
Probit	0.7833333	0.5310231
Cauchit	0.7116034	0.5105611
Complementary Log-Log	0.7116034	0.4689769

Penalized Logistic Regression & Feature Selection



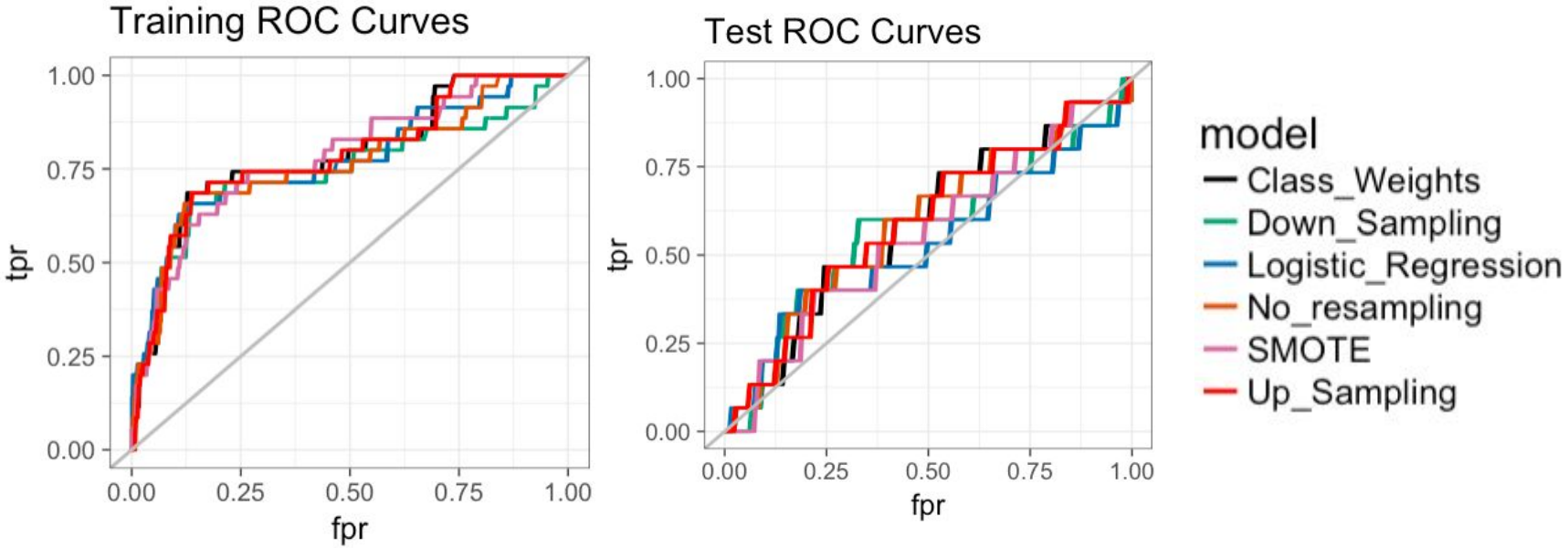




Penalized Logistic Regression & Feature Selection

Method	Train AUC	Test AUC
Elastic Net	0.7662146	0.5894389
L2 Logistic	0.7698312	0.5854785
Stepwise Logistic	0.7713984	0.5569307
L1 Logistic	0.7703436	0.5478548
Logistic	0.7719409	0.5339934

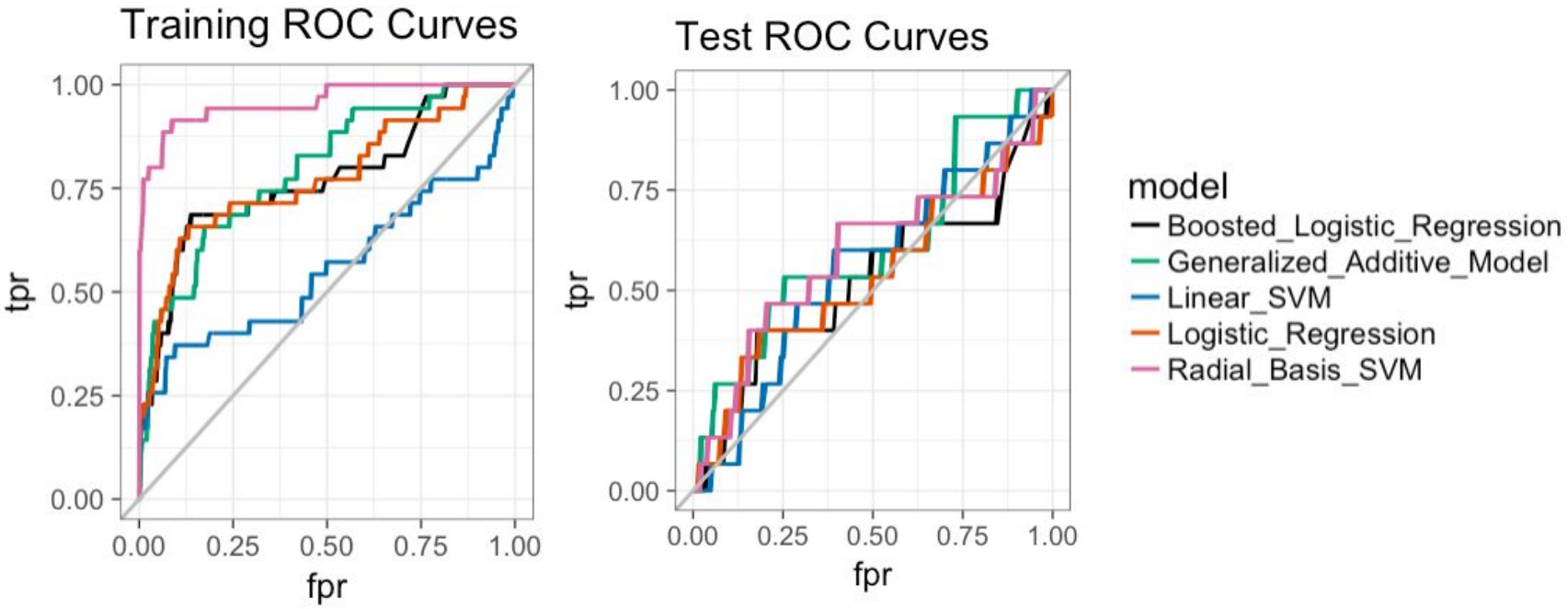
Elastic Net Logistic Regression with Re-sampling Techniques



Elastic Net Logistic Regression with Re-sampling Techniques

Method	Train AUC	Test AUC
Up-Sampling	0.7841772	0.5900990
No Re-sampling	0.7662146	0.5894389
Class Weights	0.7876733	0.5874587
Down-Sampling	0.7449367	0.5732673
SMOTE	0.7832731	0.5570957
Logistic	0.7719409	0.5339934

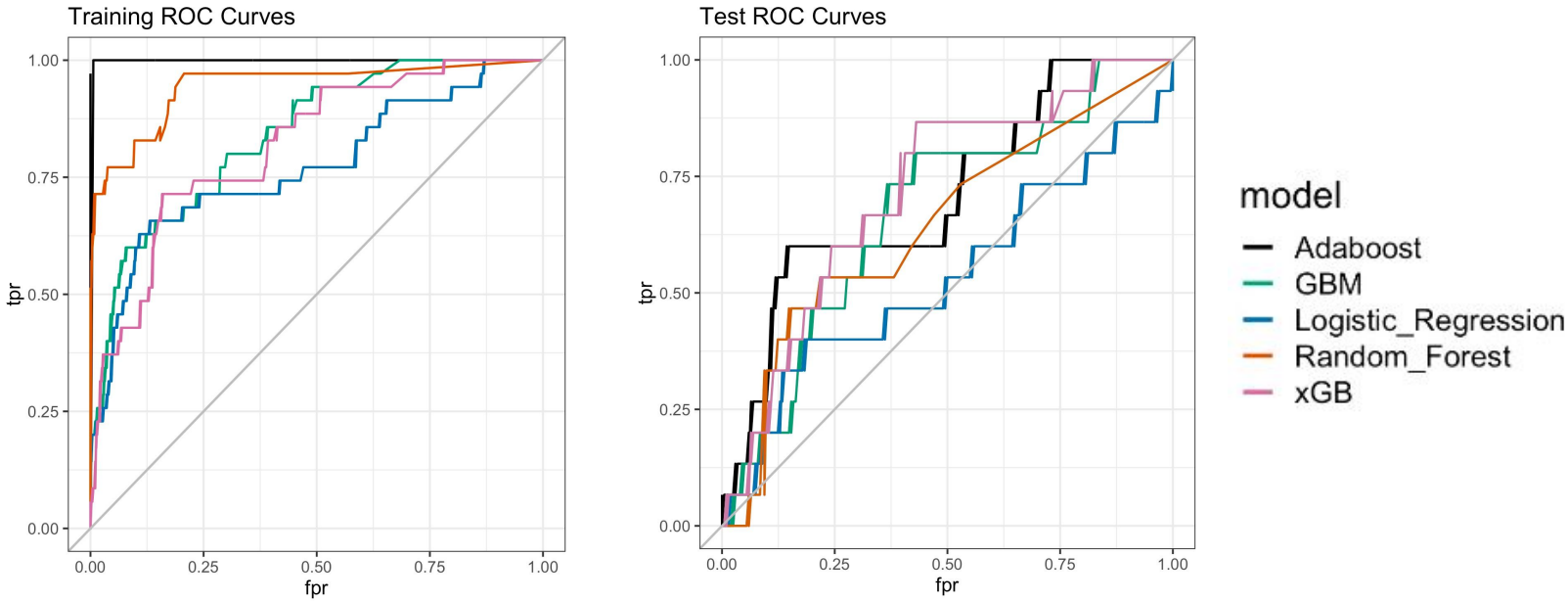
Other GLM Frameworks and SVM's



Other GLM Frameworks and SVM's

Method	Train AUC	Test AUC
Generalized Additive Model	0.7945449	0.6052805
Radial Basis SVM	0.9571730	0.5907591
Linear SVM	0.5568113	0.5594059
Logistic Regression	0.7719409	0.5339934
Boosted Logistic Regression	0.7684147	0.5264026

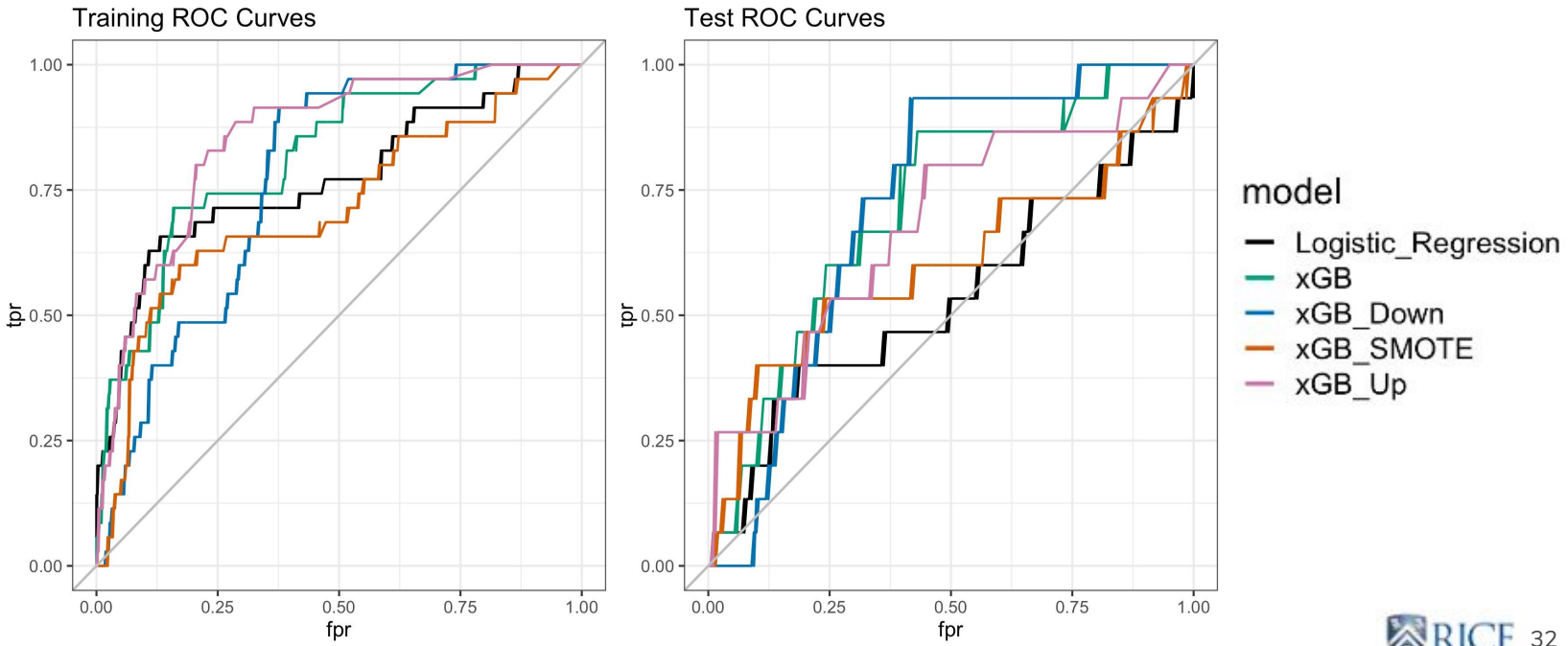
Modern Machine Learning Classifiers



Modern Machine Learning Classifiers

Method	Train AUC	Test AUC
eXtreme Gradient Boosting	0.8171187	0.7186469
Adaboost	0.9999096	0.7089109
Gradient Boosting Machine	0.8404461	0.6714521
Random Forest	0.9440024	0.6448845
Logistic Regression	0.7719409	0.5339934

Extreme Gradient Boosting with Re-sampling Techniques



Extreme Gradient Boosting with Re-sampling Techniques

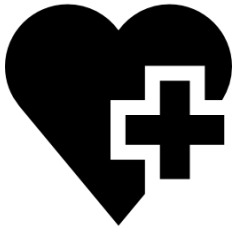
Method	Train AUC	Test AUC
xGB Down-Sampling	0.7710368	0.7254125
xGB	0.8171187	0.7186469
xGB Up-Sampling	0.8543701	0.6816832
xGB SMOTE	0.7114527	0.6037954
Logistic Regression	0.7719409	0.5339934

Conclusion

Measures of Variable Importance

First_sexual_intercourse	100.000000	Smokes_years	0.000000
Hormonal_Contraceptives_years	80.522084	Hormonal_ContraceptivesX1	0.000000
Age	60.392393	IUDX1	0.000000
Number_of_sexual_partners	42.861124	STDs_condylomatosisX1	0.000000
Num_of_pregnancies	39.922247	STDs_syphilisX1	0.000000
SmokesX1	18.271060	STDs_HIVX1	0.000000
STDsX1	10.379284	STDs_Number_of_diagnosis	0.000000
Smokes_packs_year	8.540733	Dx_CancerX1	0.000000
STDs_number	6.072748	Dx_HPVX1	0.000000
IUD_years	0.830163	DxX1	0.000000

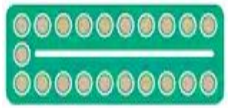
All Together



First Sexual Intercourse

-sex before age 18, sex with multiple partners and sex with someone who has had multiple partners are all known risk factors for cervical cancer and HPV

Cancer Treatment Centers of America



Hormonal Contraceptives Years

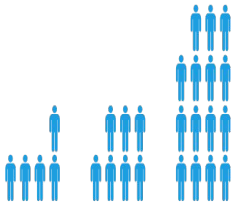
-Over five years of use = increased risk of cervical cancer
- This risk returns to normal a few years after stopping the pill

Cancer Treatment Centers of America

Age

-35 - 44 = most frequent age range
-rarely develops in women younger than 20

American Cancer Society



Number of Sexual Partners

-see first sexual intercourse

Cancer Treatment Centers of America

All Together

Number of Pregnancies

- 3 or more full-term pregnancies /1st full-term pregnancy before 17 = twice as likely to get cervical cancer.

Cancer Treatment Centers of America

Smoking

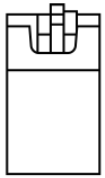
-A woman who smokes doubles her risk of cervical cancer.

Cancer Treatment Centers of America

STDs

-immune system is important in destroying cancer cells and slowing their growth and spread
-Chlamydia & HPV both linked to cervical cancer

American Cancer Society



Packs per Year

-See smoking

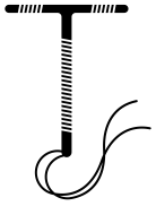
Cancer Treatment Centers of America

All Together

STDs Number

-see first sexual intercourse

Cancer Treatment Centers of America



IUD Years

-Lower risk of cervical cancer compared to pills
-There may be an association between IUD use and sexual activity, leadinging to an association with cervical cancer

American Cancer Society

Thank You