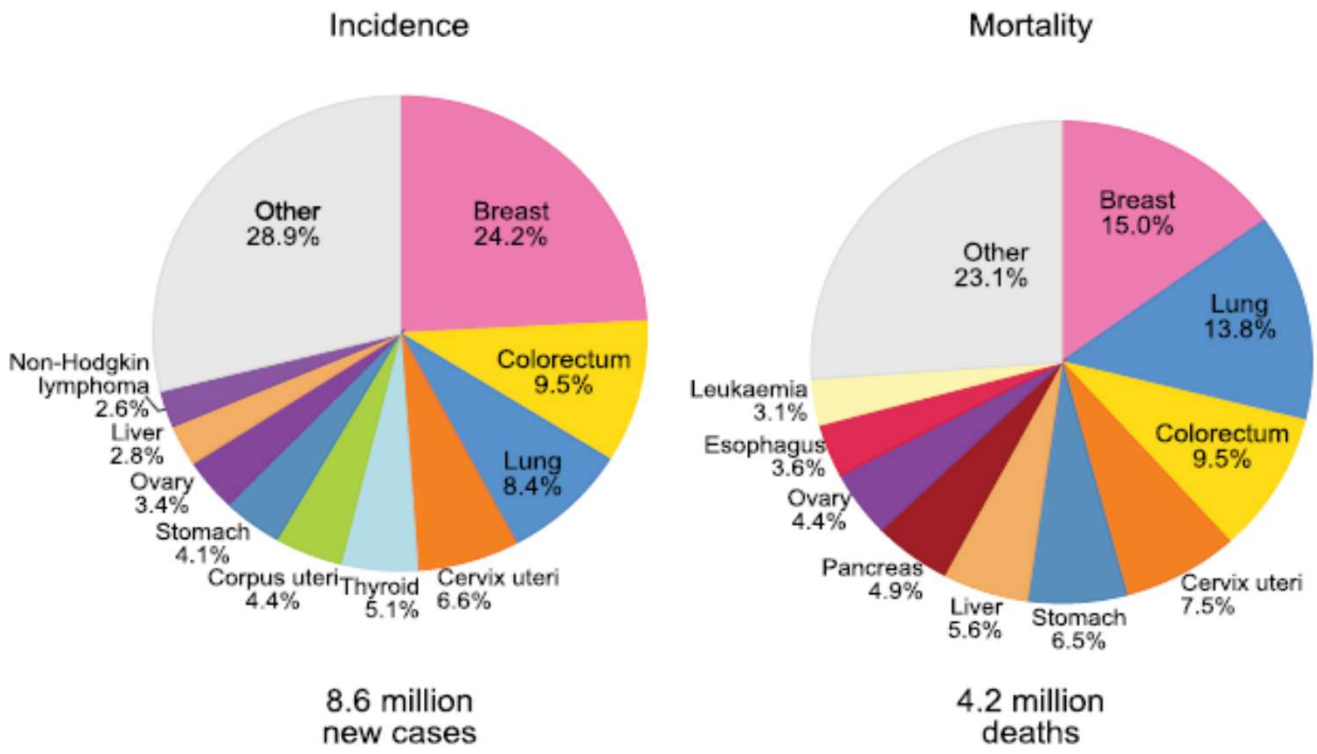
Predicting Cervical Cancer with Machine Learning

Springboard Capstone Project 1

Yi Li

Introduction



Risk Factors

- Human papillomavirus (HPV)
- Smoking
- Earlier sexual debut
- Younger age at first pregnancy
- High parity
- Long-term use of oral contraceptives
- …

Potential Client

- A risk prediction model to identify women most likely to develop cervical cancer will facilitate the cancer screening.
- Centers for Disease Control and Prevention (CDC)
- Hospitals

Data Wrangling

- Dataset: cervical cancer (Risk Factor) dataset from The UCI Machine Learning Repository
- Missing data:

  filled with median for numeric data

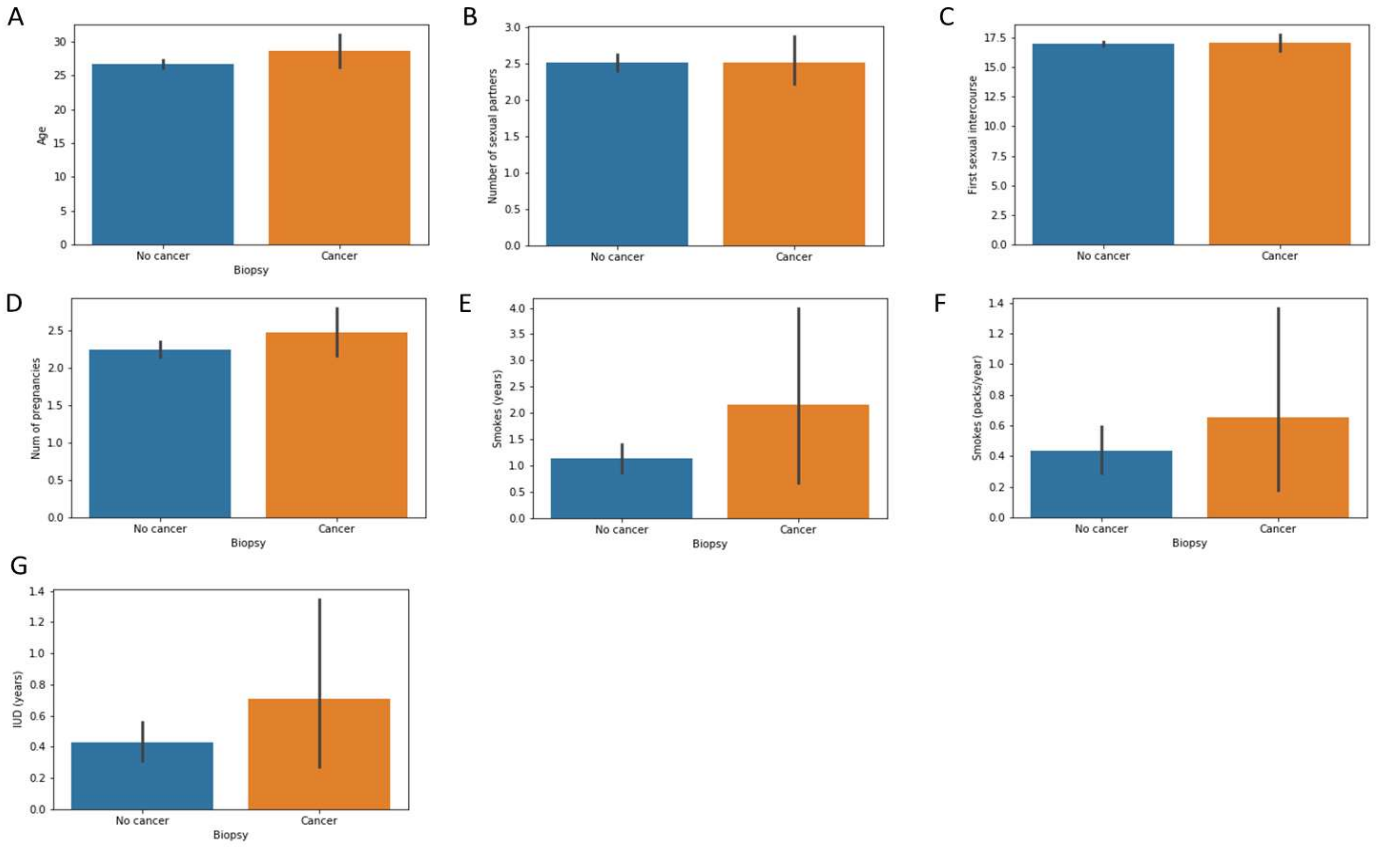
  filled mode for categorical data
- Exclude columns “STDs: Time since first diagnosis”

  and “STDs: Time since last diagnosis” , since too many missing data.

Exploratory Data Analysis and Statistical Inference

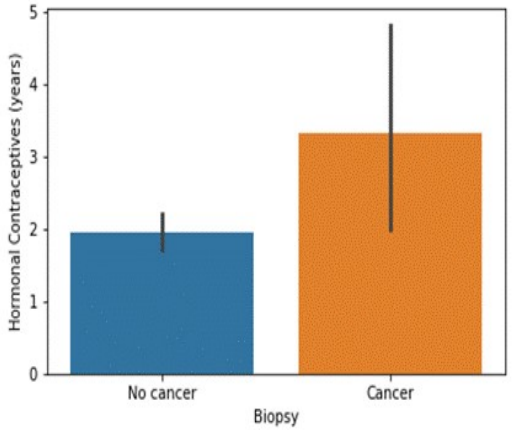
- “Biopsy” was chose as the target, the gold standard for cervical cancer diagnosis is usually biopsy result.
- Numeric features statistical inference
- Numeric features Statistical Inference 2
- Categorical features statistical inference

Numeric Features Statistical Inference

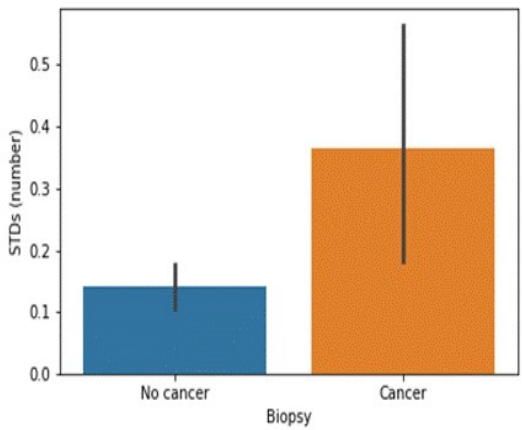


Numeric Features Statistical Inference

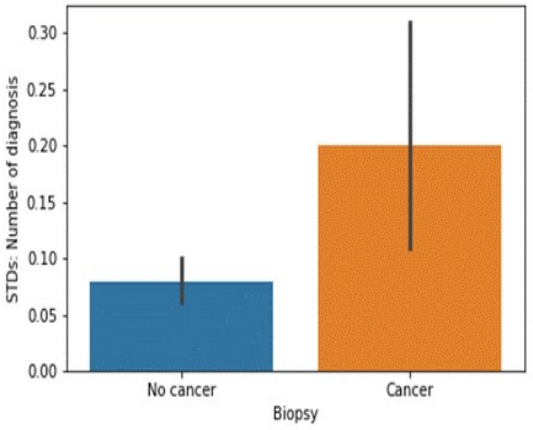
A



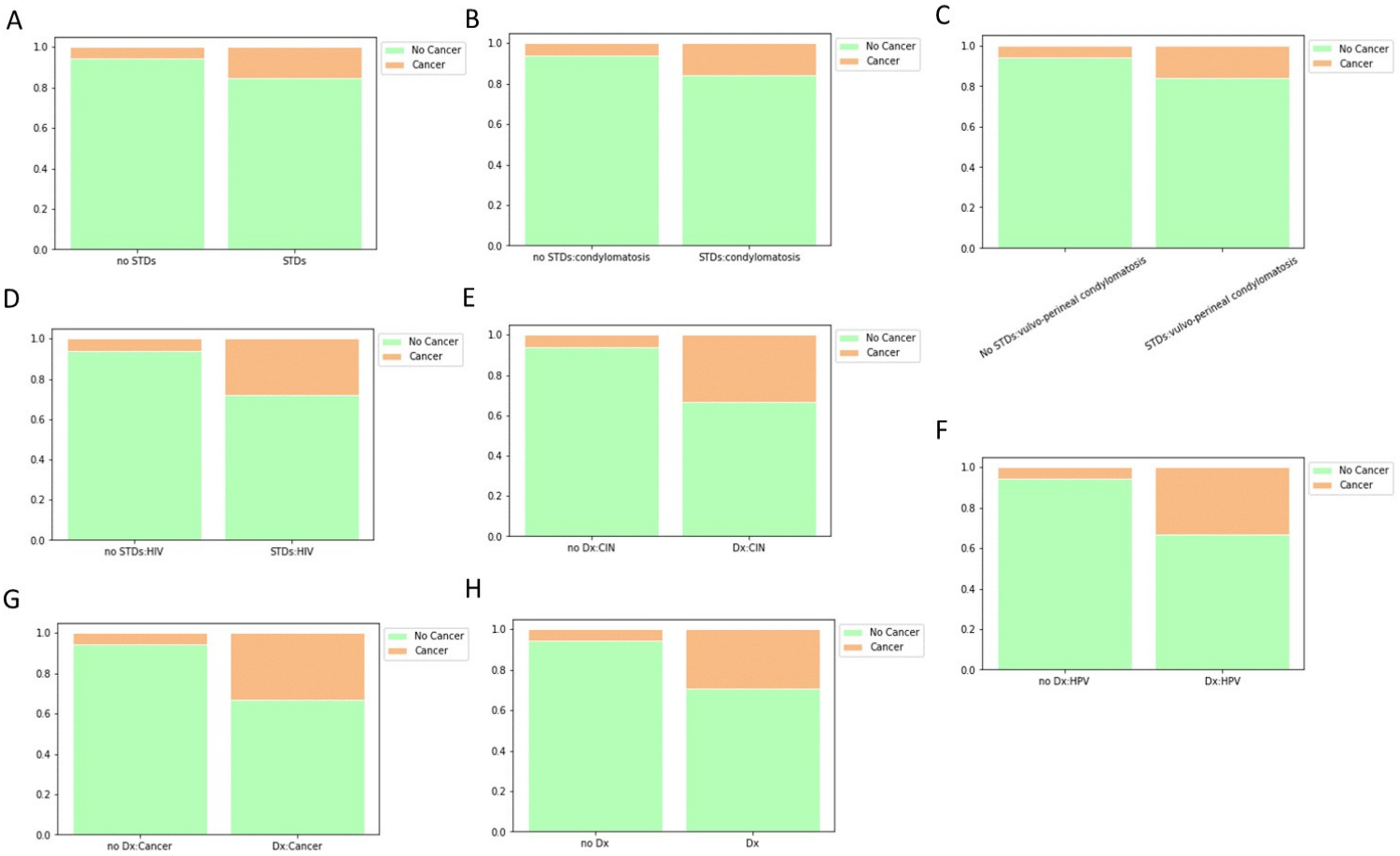
B



C



Catagorical Features Statistical Inference



Machine learning

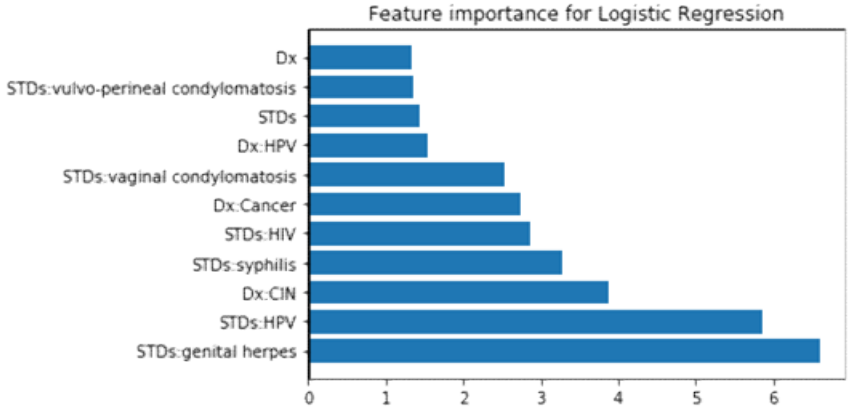
- Classification metrics

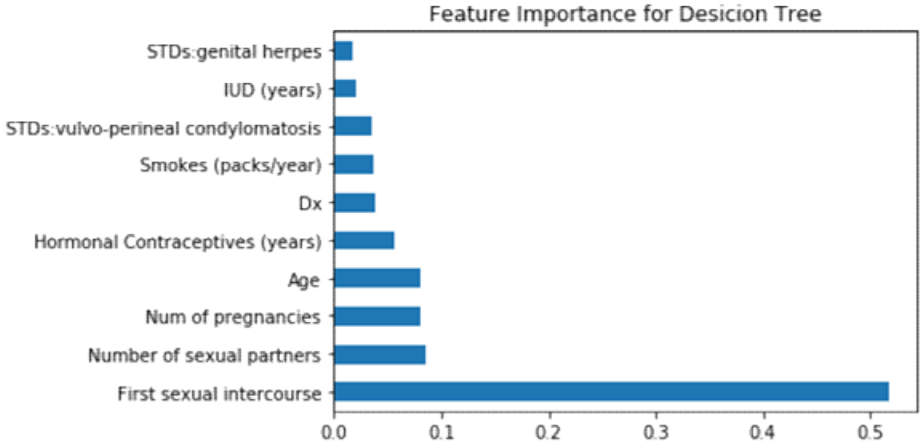
Term	Formula
Accuracy	(TP + TN)/(P+N)
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F-measure	(2 x recall x precision) / (recall+precision)

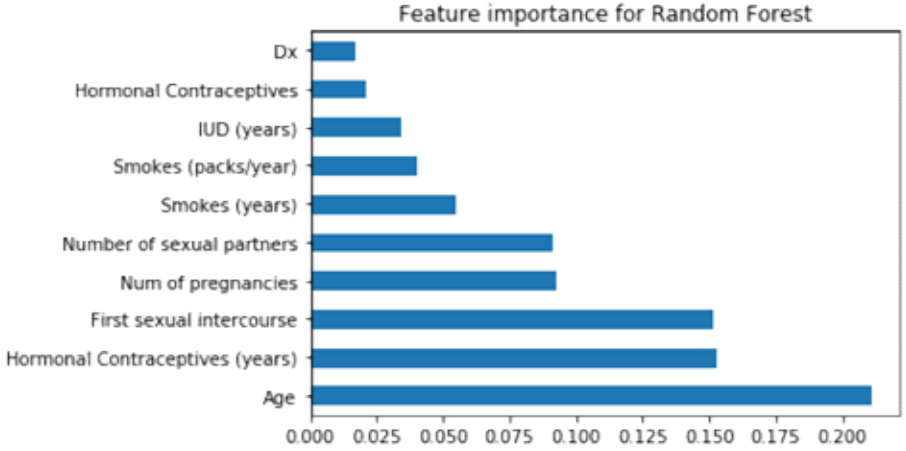
Machine Learning Models Comparison

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.860	0.913	0.900	0.953
Recall	0.222	0.222	0.444	0.333
Precision	0.105	0.200	0.222	0.6
F1	0.143	0.210	0.296	0.428

Feature Selection







Classifier /Performance	Logistic regression	Decision Tree	Random Forest
Accuracy	0.948	0.907	0.948
Recall	0.111	0.667	0.222
Precision	0.5	0.316	0.5
F1	0.182	0.429	0.428

Class Imbalance

- Upsampling

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.744	0.779	0.907	0.959
Recall	0.556	0.333	0.222	0.222
Precision	0.111	0.0857	0.182	1
F1	0.185	0.136	0.200	0.363

- Downsampling

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.622	0.738	0.517	0.639
Recall	0.444	0.555	0.667	0.777
Precision	0.0625	0.182	0.0698	0.104
F1	0.110	0.109	0.126	0.184

- SMOTE

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.779	0.756	0.860	0.942
Recall	0.556	0.333	0.222	0.222
Precision	0.128	0.0769	0.105	0.4
F1	0.208	0.125	0.143	0.286

Summary and ongoing works

- All classifiers suffered with low Recall, precision and F1 scores.
- Decision Tree with feature selection has the best overall performance.
- Feature selection for the upsampled and synthetic samples
- ROC Curves and Precision-Recall curves
- XGboost

Acknowledgements

- My springboard mentors
- Springboard staff and community