Data Science Capstone Project: Cervical Cancer Predictive Modeling

Presented by Sofy Weisenberg
February 2020

Springboard Data Science Career Track

Presentation Outline

- Introduction: Motivations and Dataset
- Exploratory Data Analysis
- Statistical Inference
- Machine Learning Models
- Conclusions

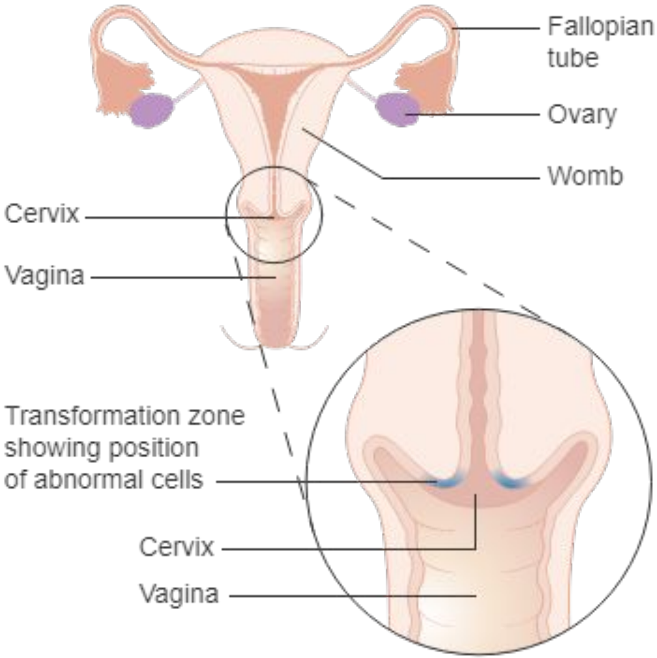
Introduction: Motivations and Dataset

Problem Description

- Despite progress in cervical cancer screening rates in recent decades, healthcare providers are still looking for ways to **predict and prevent the disease**.

- A predictive model based on a patient's medical history would potentially address the gap in current screenings.

- Such a model would give caregivers some relevant information for better **clinical decision making** and earlier diagnosis/intervention.



Data Set

- The dataset used for this model was collected at 'Hospital Universitario de Caracas' in **Caracas, Venezuela**.
- It comprises demographic information, habits, and historic medical records of **858 patients**. Several patients decided not to answer some of the questions because of privacy concerns.
- A total of **36 attributes** were collected for each patient: 35 explanatory attributes and 1 target attribute. Target is cervical biopsy result (definitive outcome for cervical cancer diagnosis).

Data set available on Kaggle and UCI.

Features and Data Types

- The following features are broken down into the 35 explanatory attributes. Categorical features are split into individual boolean attributes for each category.

Feature	Type	Feature	Type
Age	int	Sexually transmitted diseases (STDs) (yes/no)	bool
Number of sexual partners	int	Number of STDs	int
Age of first sexual intercourse	int	Diagnosed STDs	categorical x 12
Number of pregnancies	int	STDs (years since first diagnosis)	int
Smokes (yes/no)	bool	STDs (years since last diagnosis)	int
Smokes (years)	int	Previous cervical diagnosis (yes/no)	bool
Smokes (packs)	int	Previous cervical diagnosis (years)	int
Hormonal contraceptives (yes/no)	bool	Previous cervical diagnosis	categorical x 4
Hormonal contraceptives (years)	int	Hinselmann (colposcopy screening using acetic acid)	bool
Intrauterine device (IUD) (yes/no)	bool	Schiller (colposcopy screening using Lugol iodine)	bool
IUD (years)	int	Cytology (cellular microscopy screening/‘Pap’ smear)	bool

Limitations of the Dataset

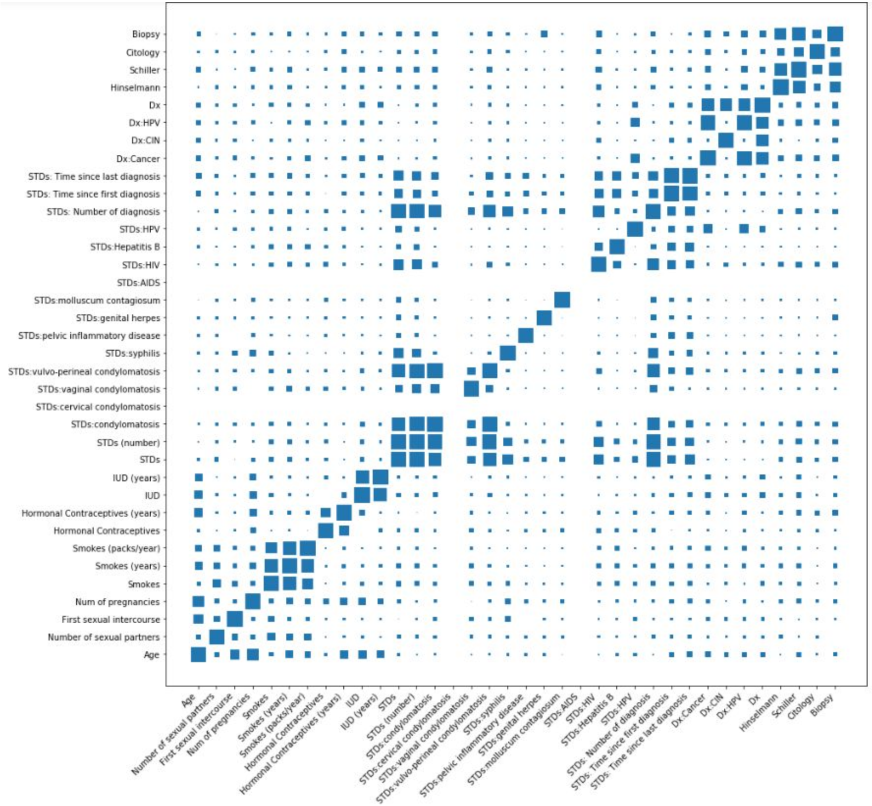
- Data was manually entered by hospital staff and contains an unknown number of **entry errors** (some are easily apparent by examining the data).
- Data was acquired by asking questions of patients
  - some patients refused to answer certain questions (**missing data**)
  - it can be assumed that some **patients answered incorrectly**
- The **patient population is biased**, with potentially higher than average rates of cancer positive outcomes.
  - small, single center study (could be improved with data from multiple hospitals)
  - patients are from generally low socioeconomic status in a developing country (high risk)
  - patients are those seeking gynecological attention, so are more likely to have had past interventions (pregnancy related, STD related, etc.)
- The dataset is **highly imbalanced** with only 6.4% positive Biopsy outcomes.

Exploratory Data Analysis

Data Preparation

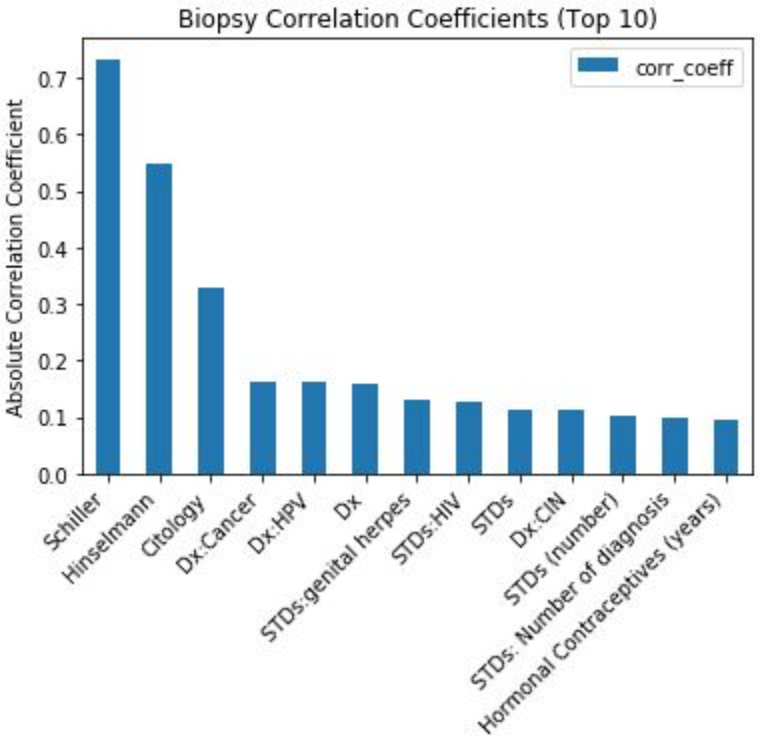
- **Missing values** were imputed using the attribute median.
    - *Note: for 0/1 attributes, this may result in increasing the bias toward the existing majority.*
- The attributes **“STDs (years since first diagnosis)”** and **“STDs (years since last diagnosis)”** were excluded entirely as more than 90% of values were missing (no recorded diagnoses for these patients). Imputation over such a large section of the data would not be an effective representation of the data.
- The attribute **“Smokes (packs)”** had a very wide range of values and occasionally had the same value as “Smokes (years)”, indicating that likely some of this data was incorrectly recorded or not standardized in the questioning (e.g. packs per year vs. packs per day).
    - However, without insight into resolving these data collection issues, the data was kept as is.

Exploring Variable Relationships



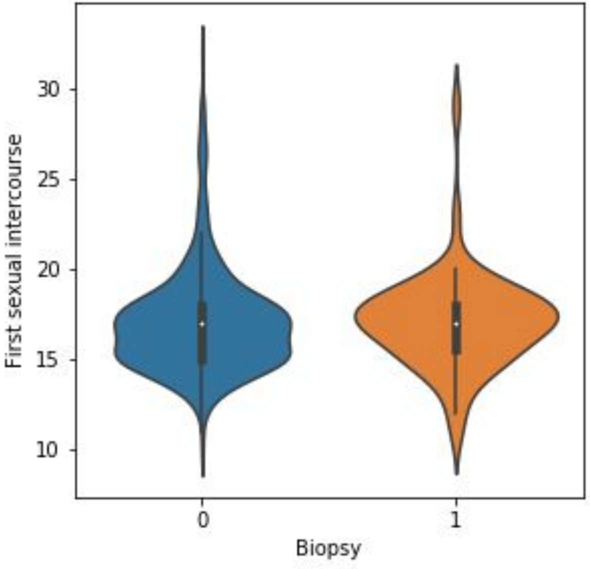
- Related variables have high correlations to one another (for example, the target variables to one another, the smoking history variables, the STD variables, etc.)
- The Dx (previous cervical diagnosis) group of explanatory variables appear to have strong correlation to the target variable, and to a lesser extent, some of the STD group and hormonal contraceptive group

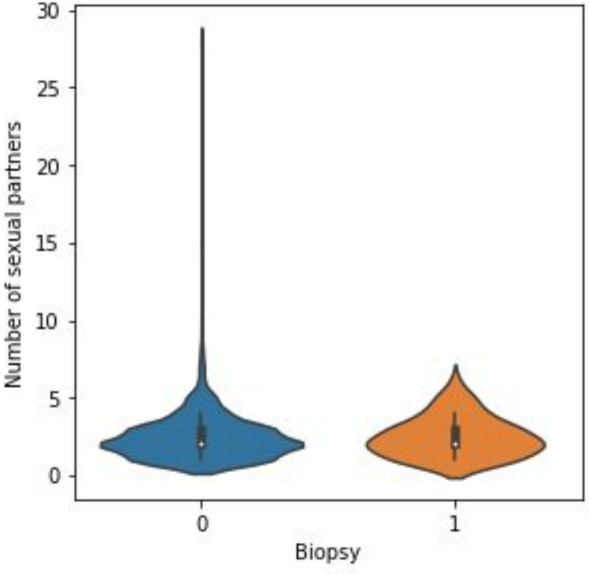
Exploring Variable Relationships (cont’d)



- The correlation to the 3 other screening tests is more trivial, so an additional 10 most correlated variables are shown.
- Linear correlation coefficients of Biopsy to other variables are quite low, hinting at the need for a more complex model.

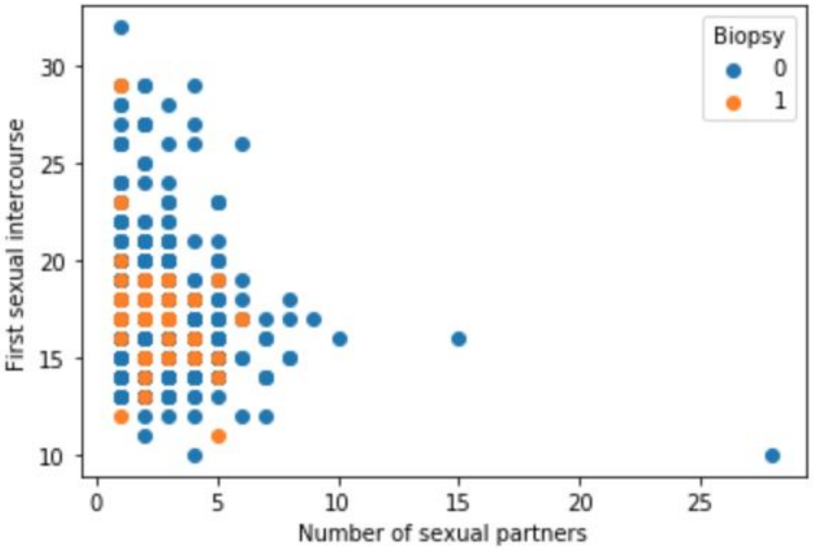
Exploring Sexual Behavior Variables





- Though not strong linear predictors, it may be interesting to look at the distributions of the sexual behavior variables to see if there is any indication of a trend.
- Neither appear to be heavily skewed and take on a full range of values for both positive and negative outcomes.

Exploring Sexual Behavior Variables (cont’d)



- Even when considered together, there do not appear to be any strong visual correlations/clusters between the sexual behavior variables and the Biopsy variable.

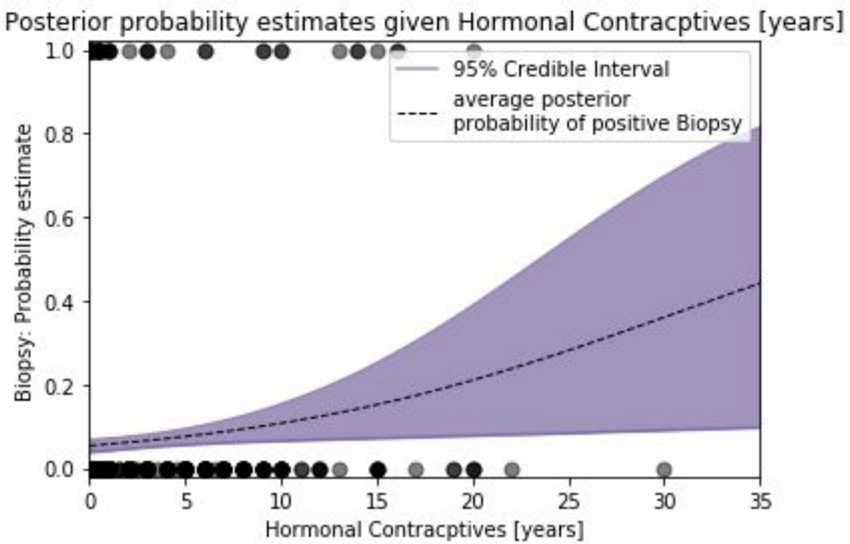
Statistical Inference

Exploring additional variables: Frequentist Inference

- Defining the following hypothesis test:
  - Ho: p(Biopsy=1|B) - p(Biopsy=1|A) = 0
  - Ha: p(Biopsy=1|B) - p(Biopsy=1|A) > 0
  - Level of significance to be tested alpha = 0.05 (95% confidence).

A	**B**	**$p(\text{Biopsy}=1\mid A)$ [# samples]**	**$p(\text{Biopsy}=1\mid B)$ [# samples]**	**p-value**
Hormonal contraceptives used < 5 years	Hormonal contraceptives used > 5 years	0.059 [710]	0.114 [114]	0.015
IUD used < 1 years	IUD used > 1 years	0.058 [786]	0.125 [64]	0.018
No history of STDs (STDs (number) = 0)	History of STDs (STDs (number) > 0)	0.055 [779]	0.152 [79]	0.0004
Smoker	Non-smoker	0.061 [735]	0.081 [123]	0.200

Exploring additional variables: Bayesian Inference



- Example of a Bayesian probabilistic model for predicting the probability of positive biopsy given the number of years of contraceptive use.

- Prior was a logistic function and the observed data was entered as a Bernoulli variable [0/1].

- The credible interval increases drastically as the years of hormonal contraceptive use increase, but there does appear to be an overall observable trend in the data.

Machine Learning Models

Modeling Overview

- Supervised, multivariate binary classification problem
- Data split into stratified training set (75%) and test set (25%)
- Data normalized to [0,1] range before model training
- Models built using Scikit-learn and Keras/TensorFlow
  - Logistic Regression
  - Support Vector Classifier (SVC)
  - Random Forest Classifier
  - XGBoost (with and without missing values)
  - k Nearest Neighbors Classifier (k-NN)
  - Artificial Neural Network
- Hyperparameter tuning performed using GridSearchCV with 5-fold cross-validation

Evaluation Metrics: Precision-Recall

$$Precision = \frac{TP}{TP + FP}$$

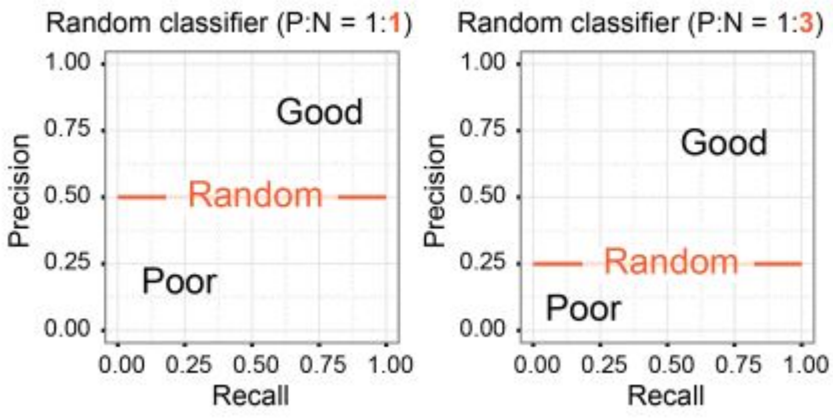
$$Recall = \frac{TP}{TP + FN}$$

		Predicted Biopsy	
		0	*1*
Actual Biopsy	*0*	***TN***	***FP***
	1	***FN***	***TP***

- Reviewing both precision and recall is useful in classification problems where there is an imbalance in the observations between the two classes.
- The goal is to minimize False Negatives (achieve high recall) and minimize False Positives (achieve high precision). But there is more emphasis on recall (False Negatives are more costly).

Evaluation Metrics: Average precision, PR Curves

- Average precision (AP) score summarizes a precision-recall plot as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.
    - For a balanced dataset, a random classifier would have an AP = 0.50.
    - For an imbalanced dataset, a random classifier would have an AP = proportion of positive class. In our case, this is AP = 0.064.



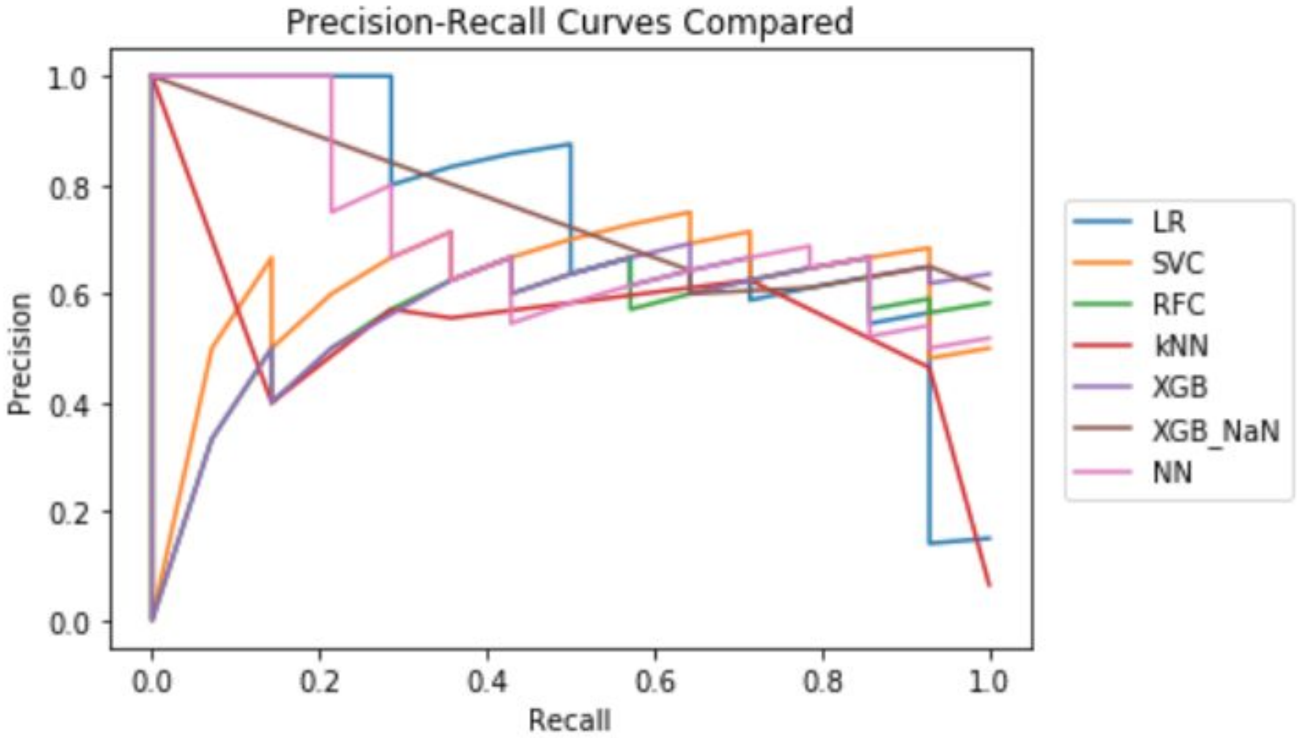
Additional Evaluation Metrics: Log Loss, F2 Score

In addition to the AP score, the following performance metrics will also be evaluated for each model:

- Logarithmic loss (log loss) measures the performance of a classification model where the prediction input is a probability value between 0 and 1.
- The F2 score is a variant of the traditional F-measure or balanced F-score (F1 score), which is the harmonic mean of precision and recall. The F2 score places stronger emphasis on the recall.
  - *for F1, □ = 1*
  - *for F2, □ = 2*

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Summarizing Model Performance



Summarizing Model Performance

Classifier	AP	F2-Score	Log Loss
Logistic Regression	0.49	0.70	0.63
SVC	0.61	0.86	0.10
Random Forest	0.35	0.52	0.10
XGBoost	0.52	0.75	0.11
XGBoost (with NaN)	0.61	0.86	0.36
k Nearest Neighbors	0.24	0.38	0.26
Neural Network	0.38	0.58	0.11

Best Model Performance

Test Set = 215 patients

<u>Support Vector Classifier (SVC)</u>

<u>XGBoost (with missing values)</u>

		Predicted Biopsy	
		0	*1*
Actual Biopsy	*0*	***TN***	***FP***
	1	***FN***	***TP***

		Predicted Biopsy	
		0	*1*
Actual Biopsy	*0*	***195***	***6***
	1	***1***	***13***

Conclusions

Main Takeaways

- Data analysis and statistical inference give clues into potentially interesting correlations between cervical cancer and patient history.
- A better-than-random classifier can be built to give clinical insight into cervical cancer risk for patients with known medical and sexual history predictors.
- Clinical use of such inferences and/or predictive models would require much more extensive validation and would benefit greatly from a larger, more diverse dataset.

Thank you!

For comments or questions, please contact me:
sofy.weisenberg@gmail.com

All python code can be found here:
https://github.com/s-weisenberg/Springboard/tree/master/Capstone_Project_1