Cervical Cancer Predictive Modeling
Project Report

Springboard Data Science Career Track
Student: Sofy Weisenberg
Date: 02/28/20

Background

(from Kaggle) About 11,000 new cases of invasive cervical cancer are diagnosed each year in the U.S. However, the number of new cervical cancer cases has been declining steadily over the past decades. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide. In the United States, cervical cancer mortality rates plunged by 74% from 1955 - 1992 thanks to increased screening and early detection with the Pap test.

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, among underprivileged populations in particular who do not receive the recommended invasive Pap screenings. Creating a predictive model based on preliminary, non-invasive patient data could provide the healthcare system with an alternative method to target those at greatest risk for the disease for preventative care or additional targeted diagnostics.

Such a model would potentially address the gap in current screening and could give patients an idea of their risk status for the disease without the need for costly laboratory tests. For example, by filling out a questionnaire during a regular doctor visit, a risk score could be assigned to a patient and follow-up testing recommended for high risk individuals. This could be particularly effective for those patients who have refused to be screened in the past for personal or financial reasons.

These patients would otherwise go without any diagnostic or predictive information in their medical records, which could be potentially critical to correct diagnosis if early symptoms present themselves themselves in the future. Therefore the proposed model would give caregivers some relevant information for clinical decision making in the cases of such patients that could lead to early diagnosis and successful treatment of cervical cancer, potentially reducing persistent mortality rates from the disease.

Data Set

(from Kaggle) The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset 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 (missing values).

The following 36 attributes were collected for each patient, with the last Biopsy attribute being target diagnostic variable.

1. (int) Age
2. (int) Number of sexual partners
3. (int) First sexual intercourse (age)
4. (int) Num of pregnancies
5. (bool) Smokes
6. (bool) Smokes (years)
7. (bool) Smokes (packs/year)
8. (bool) Hormonal Contraceptives
9. (int) Hormonal Contraceptives (years)
10. (bool) IUD
11. (int) IUD (years)
12. (bool) STDs
13. (int) STDs (number)
14. (bool) STDs:condylomatosis
15. (bool) STDs:cervical condylomatosis
16. (bool) STDs:vaginal condylomatosis
17. (bool) STDs:vulvo-perineal condylomatosis
18. (bool) STDs:syphilis
19. (bool) STDs:pelvic inflammatory disease
20. (bool) STDs:genital herpes
21. (bool) STDs:molluscum contagiosum
22. (bool) STDs:AIDS
23. (bool) STDs:HIV
24. (bool) STDs:Hepatitis B
25. (bool) STDs:HPV
26. (int) STDs: Number of diagnosis
27. (int) STDs: Time since first diagnosis
28. (int) STDs: Time since last diagnosis
29. (bool) Dx:Cancer
30. (bool) Dx:CIN
31. (bool) Dx:HPV
32. (bool) Dx
33. (bool) Hinselmann
34. (bool) Schiller
35. (bool) Cytology

    ---

36. (bool) Biopsy: target variable

Limitations of the data set:

- 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, either due to genuinely not knowing or willfully trying to conceal their medical and/or sexual history

- 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.

Data Preparation

The dataset has 36 columns, with missing values marked as '?'. The last column 'Biopsy', hold the target variable. All attributes are either numerical or boolean (0/1 binary). In order to correctly identify missing values in the data, the '?' character must be replaced with the standard NaN. This will allow for more informative exploration of the extent of the missing data. This is accomplished using the *df.replace('?', np.nan)* function.

There are still some columns with mixed or 'object' types. These must be transformed to numeric values (int or float) in order to use them for statistical or machine learning purposes. This is accomplished using the *pd.to_numeric()* function.

Missing values:

In order to work with a complete dataset, the missing values will be replaced by their medians for each column respectively. This may result in some skewing, particularly for the binary attributes (for which the median is equal to the mode). The merits of this technique will be tested later on in evaluation of performance metrics for the model against a model in which missing data is not replaced (ommitted or kept).

Outliers:

- Upon examination of the descriptive statistics, all continuous variables appear to have reasonable distributions, with only a handful of outliers (a few high values in Age, Number of sexual partners, Number of preganancies, and First sexual intercourse.

- Upon visual inspection of the histograms, the distributions appear reasonable with no significant presence of outliers. The only exception perhaps being Hormonal Contraceptives (years) which may require additional examination.

Therefore, all outlier values will be kept for further analysis.

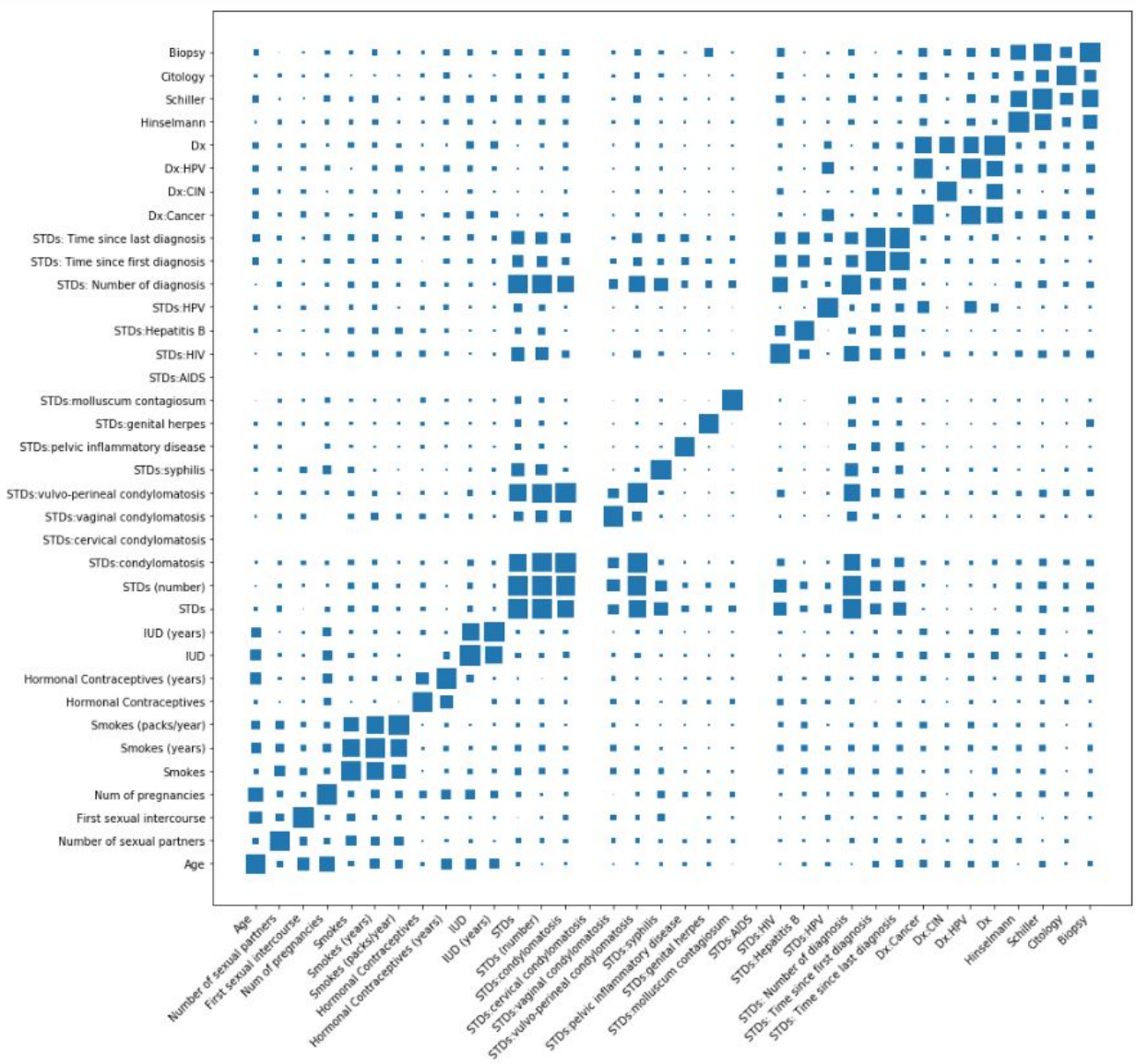
Additional considerations:

- 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.

Exploratory Data Analysis

Looking for strong correlations in the data:

A heat map of the linear (Pearson) correlation coefficients between each of the different variables could give a quick indication of the relative strengths of the different correlations, where either high or low extremes indicate a strong correlation. In the plot below, this is accomplished with the size of the marker at each variable intersection indicating the (absolute) strength of the correlation.

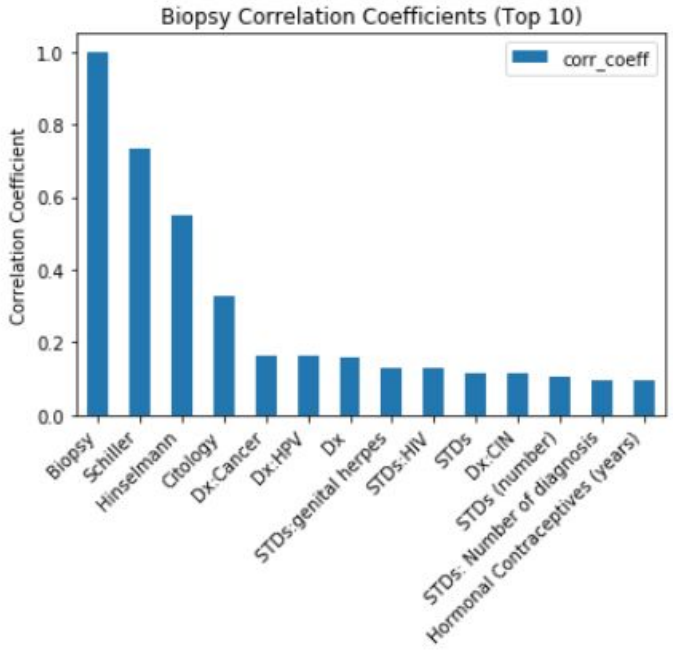


Note: 2 variables STDs:cervical condylomatosis and STDs:AIDs are all 0 values after imputation and therefore a correlation coefficient cannot be calculated for these variables.

From the plot above, a few general relationships may be observed:

- 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 (diagnosis) group of explanatory variables appear to have strong correlation to all target variables, and to a lesser extent, some of the STD group and hormonal contraceptive group

These observations can be further explored by looking at the top 10 correlation coefficients for each of the target variables (see below).



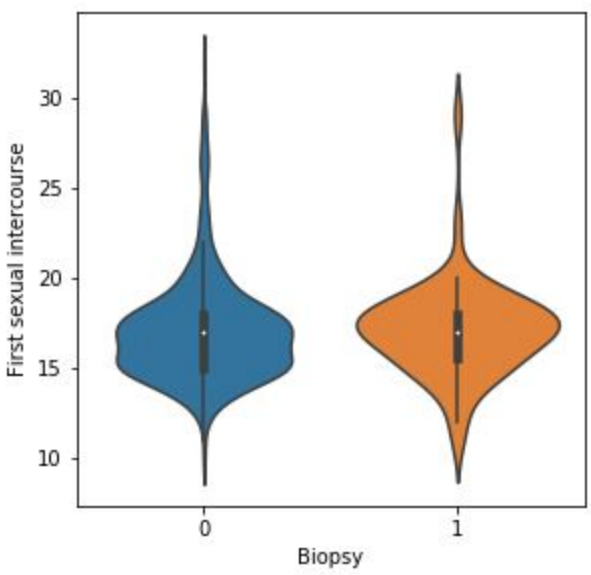
From the above breakdown of the top 10 correlation coefficients for each of the explanatory variables, the following observations may be made:

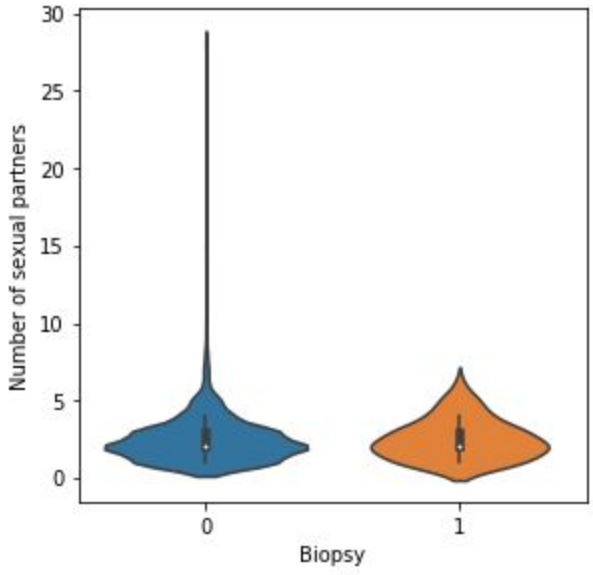
- the Dx:Cancer, Dx:HPV have the strongest correlation with all 4 of the target variables.
- total number of STDs and previous cervical diagnosis as well as years of hormonal contraceptive use also seem to have strong correlations across the board.

These observations will help guide subsequent statistical modeling and machine learning.

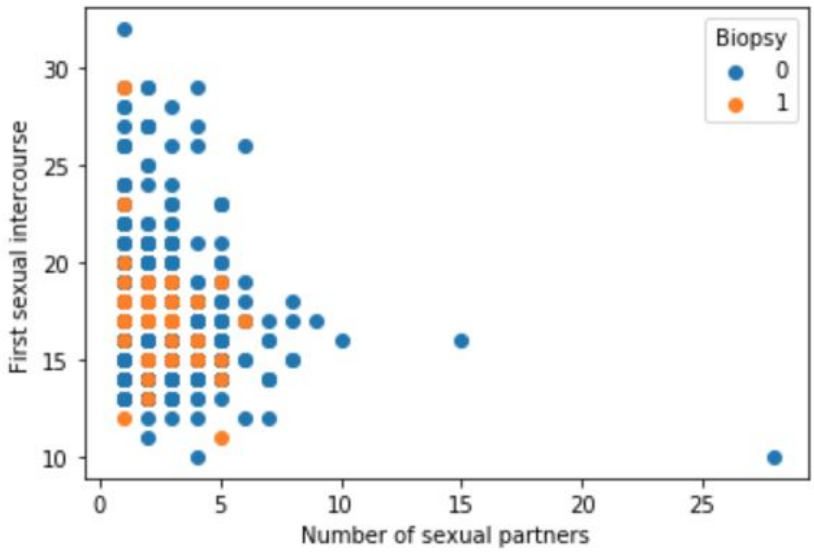
Exploring Sexual Behavior Variables:

Though not strong linear predictors on their own, 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. Even when considered together, there do not appear to be any strong visual correlations/clusters between the sexual behavior variables and the Biopsy variable (see below).



Statistical Inference

Statistical Hypothesis Testing:

The target variable is binary and therefore its probabilistic relationship to other of the

explanatory variables can be investigated (these may be either binary or continuous). One possible route of investigation splits the data into two subgroups and tests whether the proportion of positive target values is statistically significantly different in the two groups by using statistical hypothesis testing. For example, is there a statistically significant difference between the proportion of positive target values in the following subgroups for variables of potential interest:

- Hormonal contraceptives used < 5 years vs.> 5 years
- IUD used < 1 years vs. > 1 years
- History of STDs: (STDs (number) = 0) vs. (STDs (number) > 0)
- Smoker vs. Non-smoker

This can be accomplished by defining the following hypothesis test for subgroups A, B:

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).

The following assumptions for a two-proportion z-test were checked and met for all tests:

- The sampling method for each population is simple random sampling.
- The samples are independent.
- Each sample includes at least 10 successes and 10 failures.
- Each population is at least 20 times as big as its sample.

Results:
There was found to be a statistically significant difference in the proportion of positive Biopsy between the subgroups in all but the last hypothesis test (see table below).

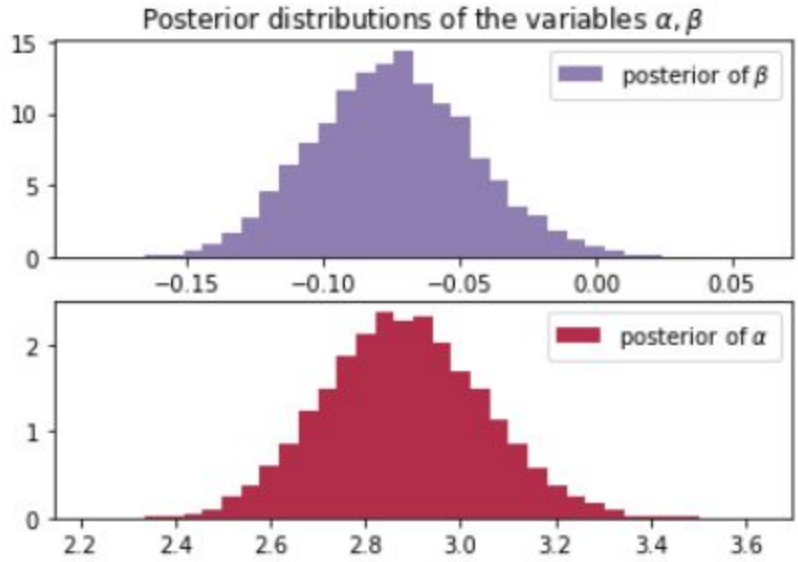
A	**B**	**p(Biopsy=1|A) [# samples]**	**p(Biopsy=1|B) [# samples]**	**p-value**
Hormonal contraceptives used < 5 years	Hormonal contraceptives used > 5 years	0.059 [710]	0.114 [114]	0.015
IUD used < 1 years	IUD used > 1 years	0.058 [786]	0.125 [64]	0.018
No history of STDs (STDs (number) = 0)	History of STDs (STDs (number) > 0)	0.055 [779]	0.152 [79]	0.0004

Smoker	Non-smoker	0.061 [735]	0.081 [123]	0.200

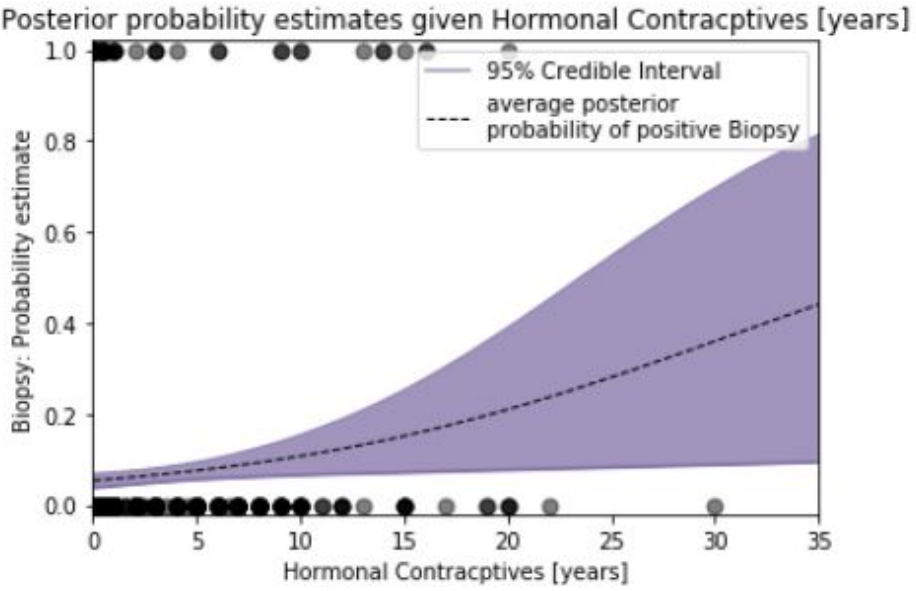
Bayesian Inference for a Continuous Variable:

Additionally, it may be asked: can a statistical inference model be applied to any of the explanatory variables? For example, given number of years of hormonal contraceptive use, what is the target variable probability of being positive?

Using pymc3, such a Bayesian inference model has been constructed. The prior used was a logistic function with parameters alpha and beta and the observed data was entered as a Bernoulli variable [0/1]. Then a Markov chain Monte Carlo (MCMC) step method was applied to get a distribution of alpha and beta. (see below)



And a posterior probability estimate for the Biopsy variable was generated, along with a credible interval of 95%. (see below)



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 (upward slope). This was just one example to illustrate the potential of such an inference method for use in predictive clinical modeling.

Machine Learning Models

This dataset allows for exploration of a supervised, multivariate binary classification (machine learning) problem.

Preprocessing:

- For many of the machine learning algorithms to be implemented here, scaling the data is an important preprocessing step. Many of the variables, including the target y, are already in the [0,1] range. However, several require normalization to this range. This helps improve classification decisions that are based on the Euclidean distance between observations.
- The dataframe also needs to be split into the feature and target data and some data must be set aside for model validation. This is done by setting aside a test set (25%) using a stratified approach such that the target variable proportion of postives/negatives is reflected equally among the training and test sets.

Model selection:

Several machine learning methods will be implemented and compared, including both traditional ML methods and DeepLearning (neural network) methods. All models are 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

Model fitting:

Hyperparameter tuning is accomplished via grid search coupled with 5-fold cross-validation. Several hyperparameters are considered, and the parameters' performance is evaluated using the F1 score, which takes into account both precision and recall. Though the F1 score assumes equal weight for precision and recall (when in reality recall/sensitivity is of higher priority), since no other weighting scheme is known a priori, it is an acceptable default. As a note: other classifier scoring metrics (e.g. log loss, Brier score loss, ROC AUC) were also tried and gave very similar hyperparameter optimizations.

Selecting evaluation criteria:

Reviewing both precision and recall is useful in classification problems where there is an imbalance in the observations between the two classes (see here). Specifically, there are many examples of no event (class 0) and only a few examples of an event (class 1). This is indeed the case for this dataset.

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

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

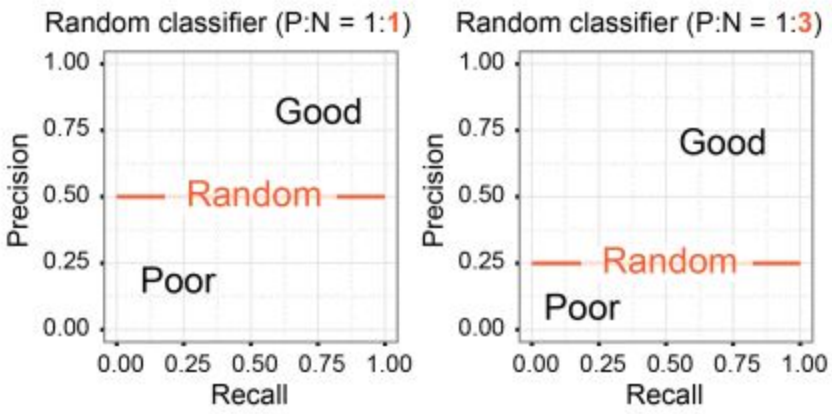
		Predicted Biopsy	
		0	1
Actual Biopsy	0	**TN**	FP
	1	FN	**TP**

The reason for this is that typically the large number of class 0 examples means we are less interested in the skill of the model at predicting class 0 correctly, e.g. high true negatives. Key to the calculation of precision and recall is that the calculations do not make use of the true negatives. It is only concerned with the correct prediction of the minority class, class 1.

Therefore, the precision-recall performance as well as the log loss, another common classifier performance metric, will be examined for the final logistic regression model. These metrics can later be used to compare various ML models.

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.



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

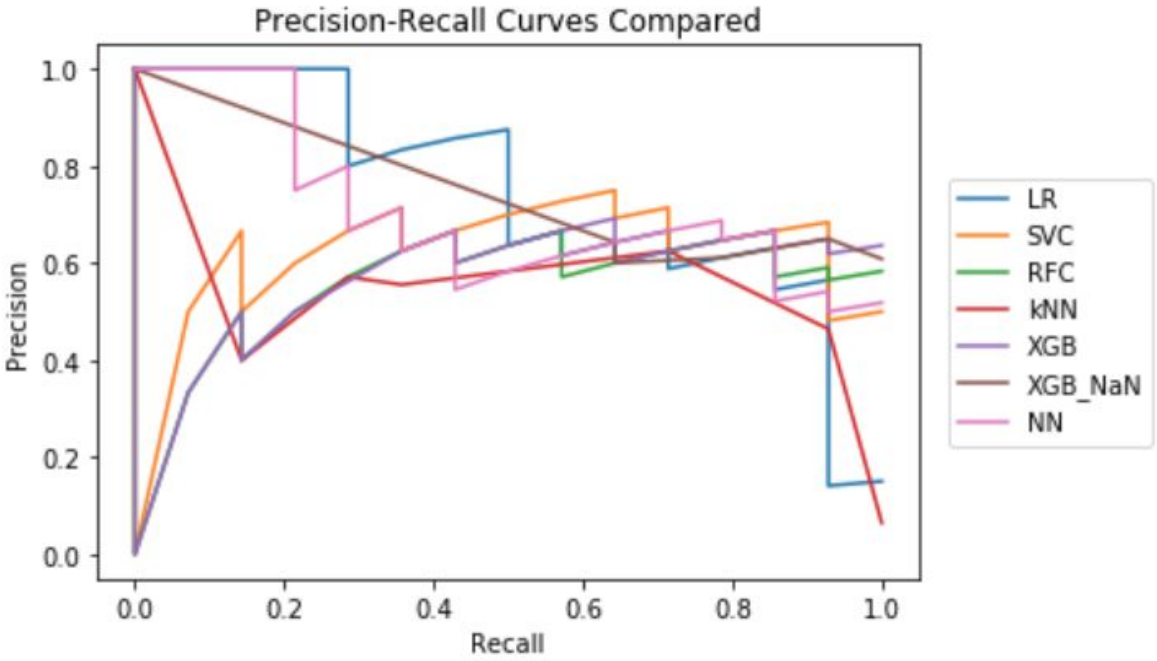
- Logarithmic loss (log loss) measures the performance of a classification model where the prediction input is a probability value between 0 and 1.
- The F2 score is a variant of the traditional F-measure or balanced F-score (F1 score), which is the harmonic mean of precision and recall. The F2 score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

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

- for F1, □ = 1
- for F2, □ = 2

Summarizing model performance:

The following are the PR curves for each of the models. They are all uniquely shaped due to differences in the classification mechanism. It is difficult to intuitively compare these curves. ROC curves are more intuitive to understand in a visual comparison, but are not particularly well suited for imbalanced classification problems such as this one.



The following is a comparison of the summarizing metrics for each of the classifier models:

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

The best performers are the SVC and XGBoost (with missing values) models. A slightly better performing DeepLearning model (AP = 0.6875, log loss = 0.1514) is presented in a PeerJ Computer Science paper which shows the benefits of combining dimensionality reduction and an autoencoder neural network architecture for this dataset.

Both of the top performing models have the same confusion matrix result on the test set of 215 patients (see below)

Support Vector Classifier (SVC)

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

XGBoost (with missing values)

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

Conclusions

- 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.