Cervical Cancer Risk Analysis

By Yuwen Hsieh

Introduction:

What is cervical cancer?

Cervical cancer occurs at cervix of women. Cervix is the lower part of the uterus. Cervical cancer is caused by the HPV virus, which can be spread by sexual contact. Most of the time, women's body can fight off the virus by itself, but not all the time. The time when women's body cannot fight off the virus, it can lead to cancer. Women can have different kind of medical test such as HPV test to check for abnormality before cancer occurs. Cancer can be prevented by treating any problems before they become cancer.

Purpose for the analysis:

The purpose for the analysis is to identify the high-risk factors for cervical cancer with the help of machine learning. If we can identify the high-risk factors for the cancer, then women and medical field professions can make necessary precautions to try to prevent women from getting cervical cancer. The analysis also looks for preexisting patterns from the data with different visualizations.

Audience:

- Medical professionals
- Women
- Data Scientist/Analyst
- General public

Data:

Source:

UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. (Credit to Google)

Variable Descriptions:

Age: Age of patients (Numeric)

Number of sexual partners: The number of sex partner (Numeric)

First sexual intercourse: Age of first sexual intercourse (Numeric)

Num of pregnancies: Number of pregnancies (Numeric)

Smokes: Whether the smokes or not (Categorical)

Smokes (years): Number of years the patient smoked (Numeric)

Smokes (packs/year): A pack year is defined as twenty cigarettes smoked everyday for one year (Numeric)

Hormonal Contraceptives: Whether the patient used a type of birth control that uses hormones to prevent pregnancy (Categorical)

Hormonal Contraceptives (years): Number of years that the patient used hormonal contraceptives (Numeric)

IUD: whether the patient use the IUD (Intrauterine contraceptive device)(A device inserted into the uterus (womb) to prevent conception (pregnancy))(Categorical)

IUD (years): Number of the years that the patient used IUD (Numeric)

STDs: Whether the patient diagnosed with a STD (Sexually transmitted disease) (Categorical)

STDs (number): Number of STDs (Numeric)

STDs:condylomatosis: Whether the patient diagnosed with condylomatosis (Categorical)

STDs:vaginal condylomatosis: Whether the patient diagnosed with vaginal condylomatosis (Categorical)

STDs:vulvo-perineal condylomatosis: Whether the patient diagnosed with vulvo-perineal condylomatosis (Categorical)

STDs:syphilis: Whether the patient diagnosed with syphilis (Categorical)

STDs:pelvic inflammatory disease: Whether the patient diagnosed with pelvic inflammatory disease (Categorical)

STDs:genital herpes: Whether the patient diagnosed with genital herpes (Categorical)

STDs:molluscum contagiosum: Whether the patient diagnosed with molluscum contagiosum (Categorical)

STDs:HIV: Whether the patient diagnosed with HIV(Categorical)

STDs:Hepatitis B: Whether the patient diagnosed with Hepatitis B (Categorical)

STDs:HPV: Whether the patient diagnosed with HPV (Categorical)

STDs: Number of diagnosis: Number of the STD diagnoses of the patient (Numeric)

Dx:Cancer: Diagnosis of cancer (Categorical)

Dx:CIN: Diagnosis of CIN (Cervical intraepithelial neoplasia) (Categorical)

Dx:HPV: Diagnosis of HPV (human papilloma virus) (Categorical)

Dx: Diagnosed or not (Categorical)

Hinselmann: Hinselman test result (Categorical)

Schiller: Schiller test result (Categorical)

Citology: Citology test result (Categorical)

Biopsy: Biopsy test result (Categorical)

Note:

Hinselmann is a medical diagnostic test.

Schiller is a preliminary test for cancer.

Citology is the examination of cells from the body under a microscope.

Biopsy is a sample of tissue taken from the body in order to examine it more closely. A doctor should recommend a biopsy when an initial test suggests an area of tissue in the body isn't normal. Doctors may call an area of abnormal tissue a lesion, a tumor, or a mass.

Data Wrangling (Technical):

Missing data:

Some columns contain too much missing data, so they are dropped.

Missing data is handled by knn imputation with k = 5.

Imbalanced data:

Down sampling is performed by removing some rows containing missing data for the majority target class.

Other:

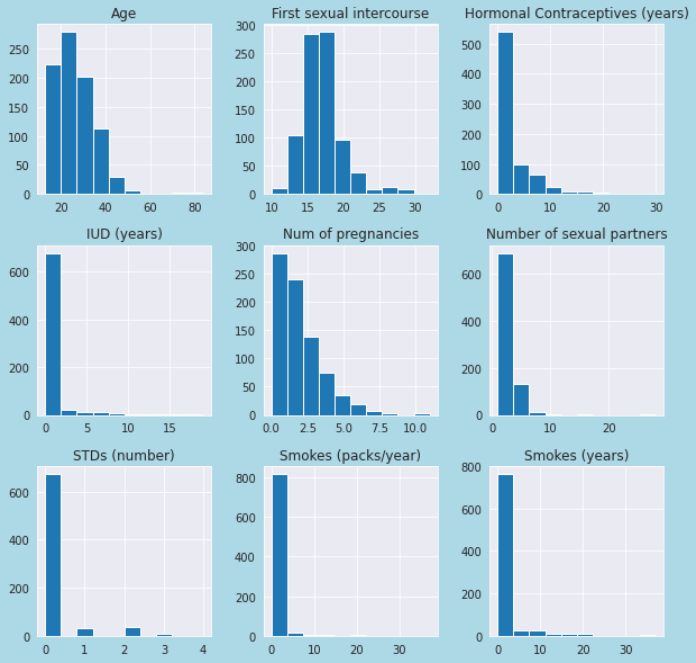
Some columns are dropped because they are all potential target variables, but this analysis only focus on 1 target variable.

Some columns are dropped because the information contained in those columns can be determined by other columns. For example, “Smokes” column is always 1 if “Smokes (years)” is not 0. These are redundant information and they can cause problems for machine learning models.

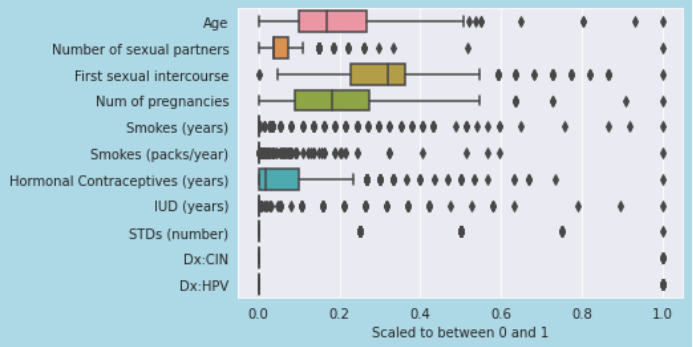
Stratified train test split is performed with testing data size proportion of $\frac{2}{9}$.

Exploratory Data Analysis:

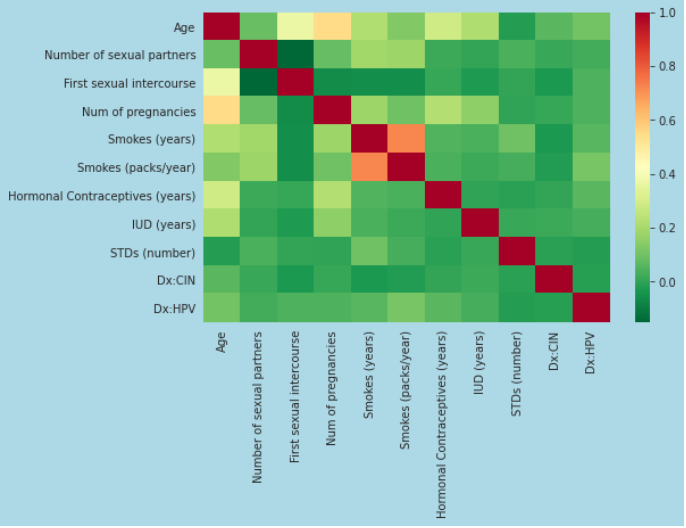
Visualization for modeling use only:



All columns seem to be right skewed. In addition, HC, IUD, number of sexual partners, STDs, smokes columns seem to be skewed a lot with most values being 0.

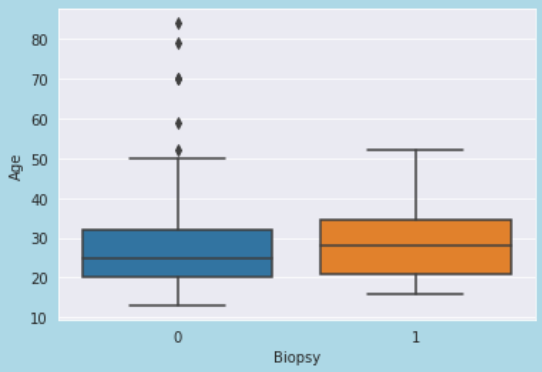


The data contains a lot of outliers.

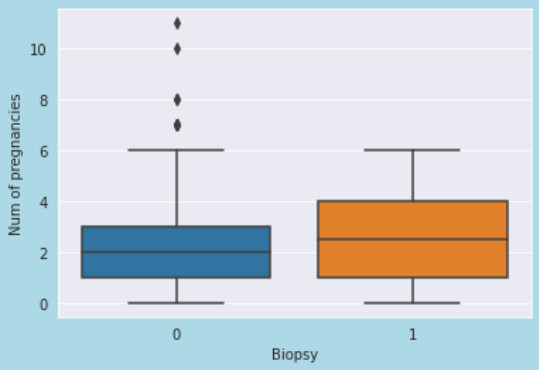


The only features that are alarmingly correlated to each other are the smokes columns, which is no surprise. These columns will be combined into 1.

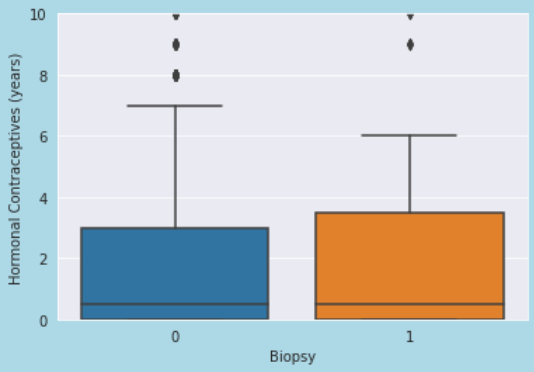
Visualizations with insights:



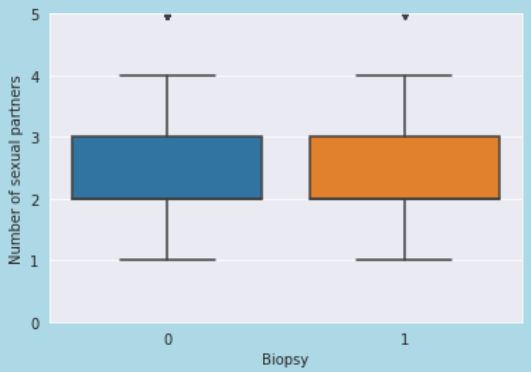
People diagnosed with cancer seems to be older than those not diagnosed with cancer on average.



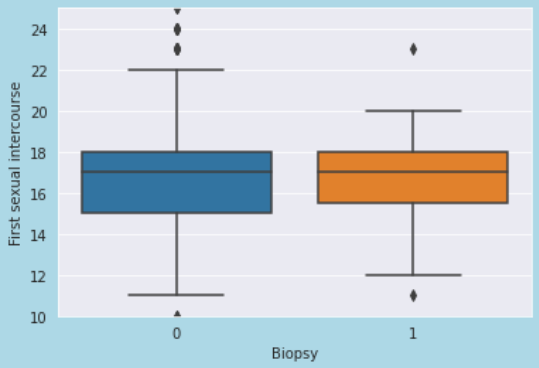
People diagnosed with cancer seems to be having higher number of pregnancies than those not diagnosed with cancer on average.



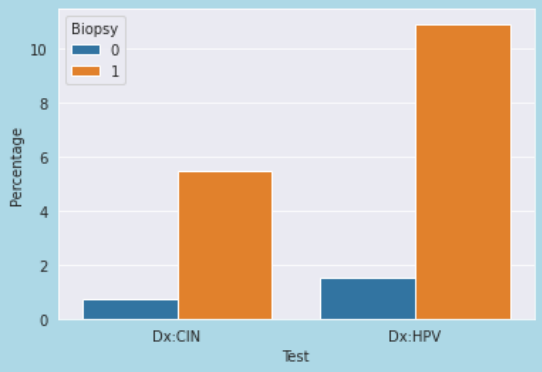
It doesn't seem to be having much difference of the average HC years between people diagnosed with cancer and those are not.



It doesn't seem to be having much difference of the average number of sexual partners between people diagnosed with cancer and those are not.



It doesn't seem to be having much difference of the average first sexual intercourse between people diagnosed with cancer and those are not.



Both CIN & HPV tests shows that people diagnosed with cancer has much higher percentage of being tested positive than those not diagnosed with cancer.

Data Preprocessing (Technical):

Different type of data preprocessing is done on the training data to test for training machine learning model.

- Data binning are performed on numerical data.
- Some features contain mostly 0 with several outliers, and the outlier are replaced with 1.
- Smokes features are combined into 1 by multiplying.
- Combination of under sampling and over sampling are applied to the training data with SMOTE algorithm.
- Min Max Scaling is applied to the data.
- Log transformation is applied to reduce skewed data

Modeling (Technical):

Models tested:

- Random Forest
- Ada Boosted Decision Tree
- Gradient Boosted Decision Tree
- Support Vector Machine
- K Nearest Neighbors
- Logistic Regression
- Ridge Regression Classifier

The F1 score is used for testing model performance.

1. All models are trained on scaled training data. Random grid search hyper parameter tuning method is used to select the best performing model.
2. Different kind of preprocessed training data, and non-preprocessed data are trained on the best performing model with the optimal hyper parameters obtained from random grid search to see if it makes any different on the model's performance.
3. Dropping least important features is also tried to train the model to check if model performs better than using all features.

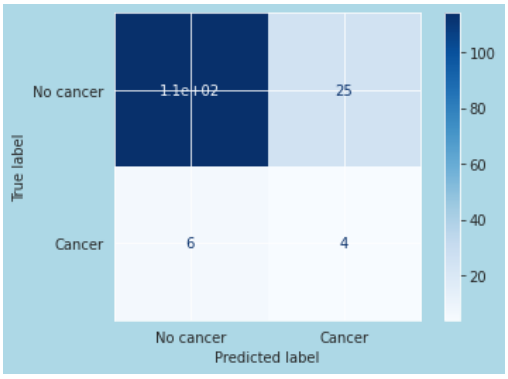
Final model selection:

The best model is found to be Random Forest. No preprocessing technique is found to be improving the model performance. Bayesian optimization technique is use for hyper parameter tuning again to see if better parameters can be found. It turns out no better parameters is found.

Final model performance:

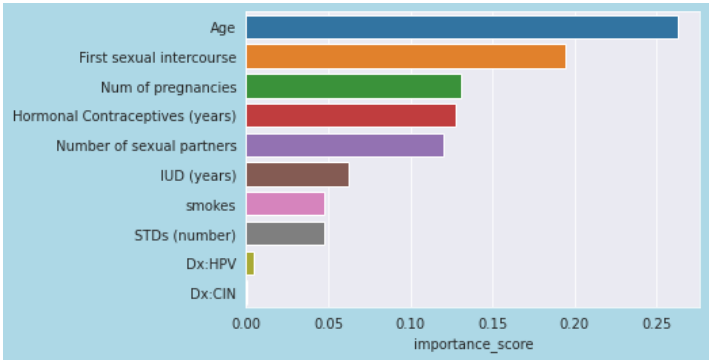
First looking at the testing data set, it shows that only around 7% of the data is having "Biopsy" column of 1. "Biopsy" is the target of the model.

Confusion Matrix:



The recall score for the model is found to be 0.4, which is not bad considering the testing data only contains about 7% of positive "Biopsy". The overall accuracy score is found to be 0.79.

Feature importance:



Age and First sexual intercourse seem to be the 2 most important risk factors for cervical cancer. Number of pregnancies, HC (years), and Number of sexual partners seem to be the medium risk factors for cervical cancer.

Conclusion:

Based on the visualizations from exploratory data analysis, we can see age, number of pregnancies, and the 2 medical test HPV and CIN are possibly the top risk factors for cervical cancer. As for the machine learning model, it is telling us Age and First sexual intercourse are the top 2 risk factors. Furthermore,

the model tells HC (years), number of sexual partners, and number of pregnancies are the medium risk factors.

The inconsistency of the insights is likely due to the highly imbalanced data. However, both the EDA and the model agrees on age, and number of pregnancies are possibly the risk factors of cervical cancer. It seems that older women and women with high number of pregnancies are at higher risk of cervical cancer than other women.

Future improvements:

- The model seems to be suffering from imbalance data a lot. Perhaps we need more data for the model to have a better performance and provide more convincing results than the current one.
- Missing data could also lead to the poor performance of model. If we know exactly why the missing data exists, then we could do a better job at filling in missing data.
- If possible, we can collect more meaningful features to improve the model.