Springboard Capstone Project #2

Predicting Cervical Cancer

Saatvik Ramisetty

Introduction

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.

This project can help improve cancer risk detection and recognizing the symptoms that cause cervical cancer in women. Machine Learning plays an important in improving healthcare and understanding cancer and it's causes.

Data

Dataset and Paper available via these links.

The dataset comprises demographic information, habits, and historic medical records of 858 patients. Data covers a wide range of information including details like number of sexual partners, number of years the patient has been smoking for, number of pregnancies and if the patient has STDs.

Since the dataset is small and it's the case of imbalanced data, we must account for any errors or low scores.

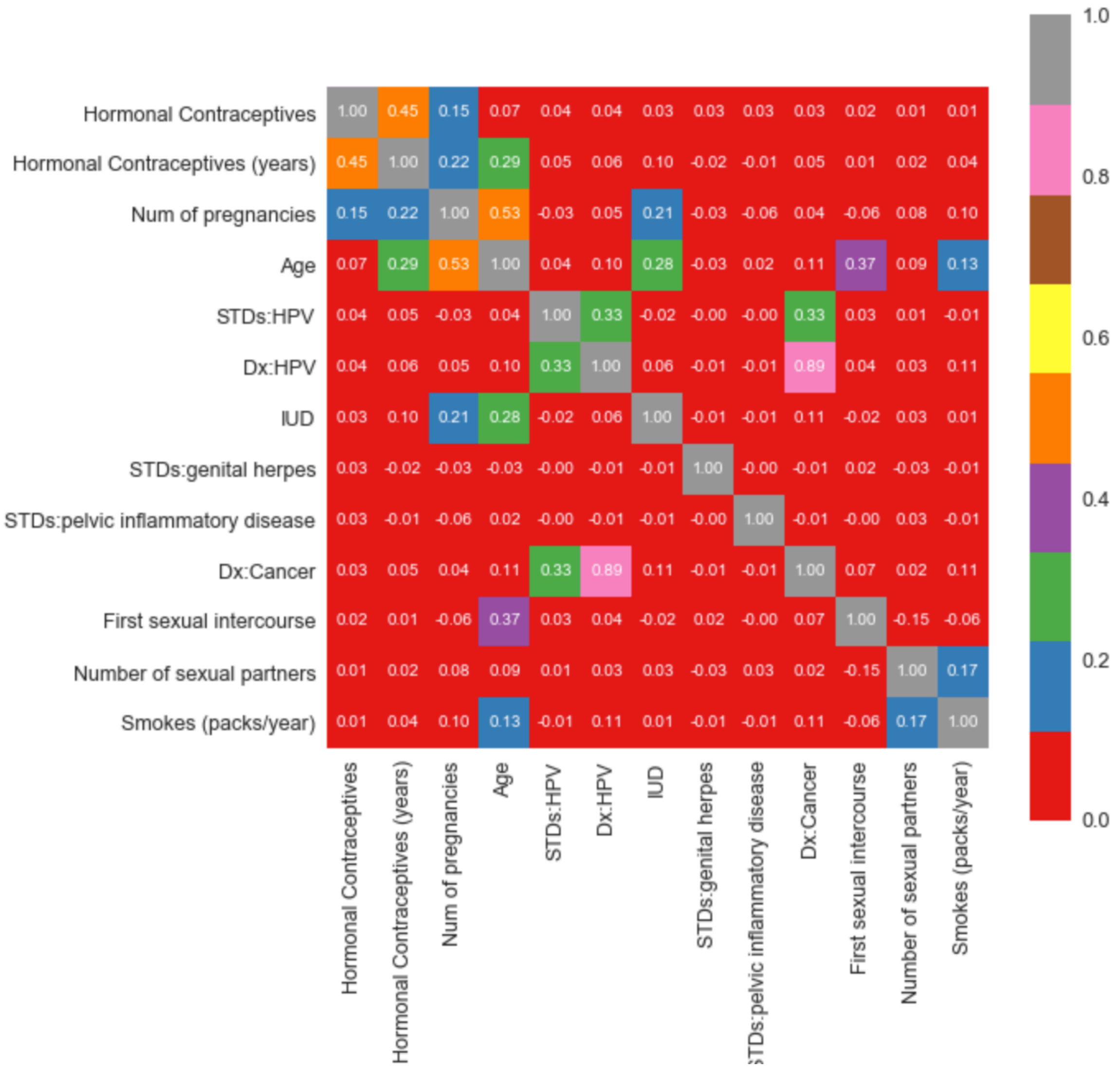
Practical Applications

As we gather more patient data, the model continues to be improved and can help decided if a patient requires screening or not. If so, which technique would be the most efficient technique for screening.

Data Transformation & Cleaning

Data transformation has perhaps been the most challenging part of the project. The dataset has many missing values which would generally be handled by dropped the missing rows/columns.

But due to the small size of the dataset, it's important to use feature correlation and impute median/mean for missing values.



Using heatmaps, we try to understand the correlation between hormonal contraceptives, age and number of pregnanices. We can use the the mean of the features to decide if the patient with missing values is replaced with “1” or “0”.

Similar transformation is done for IUD, STDs etc.

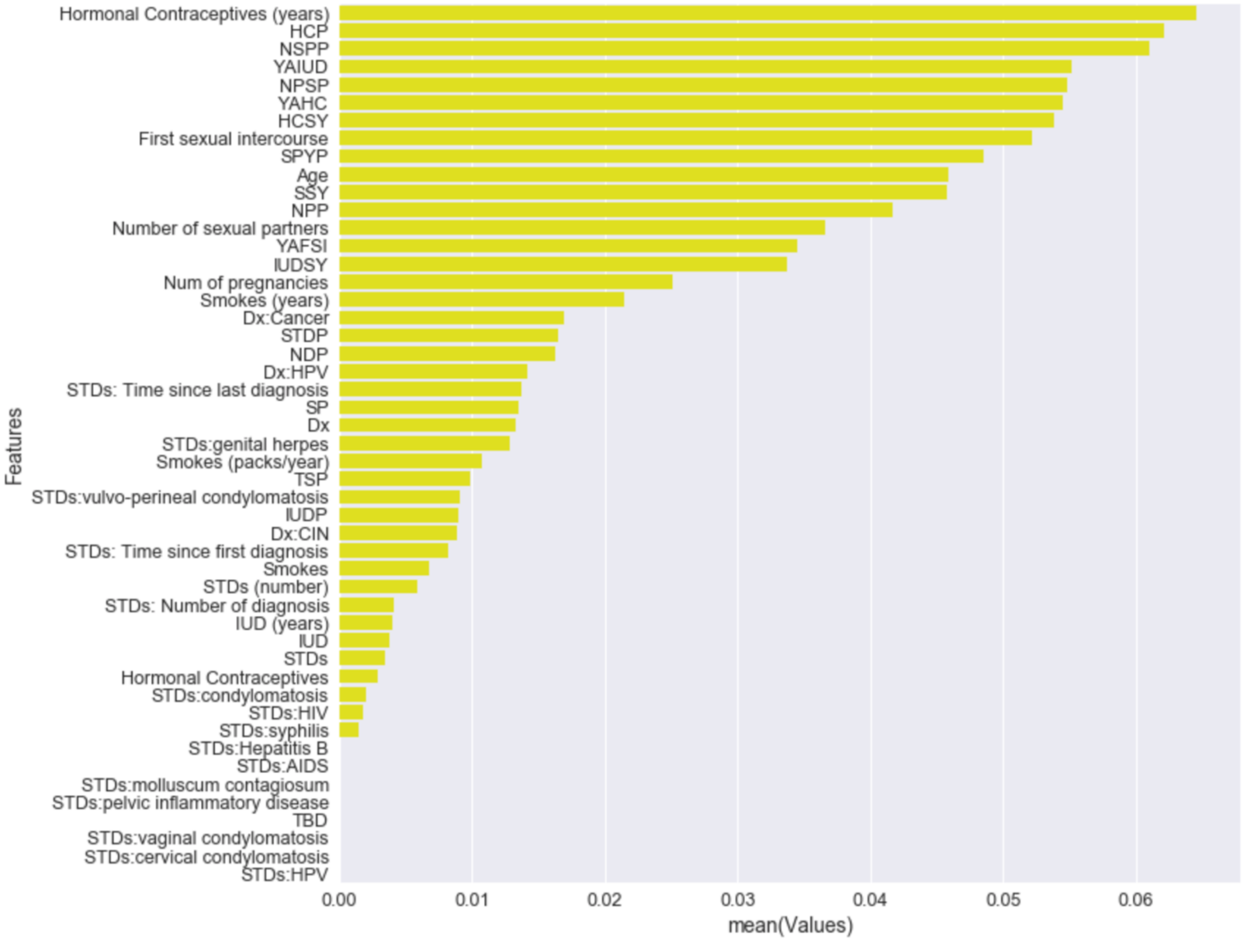
Feature Engineering

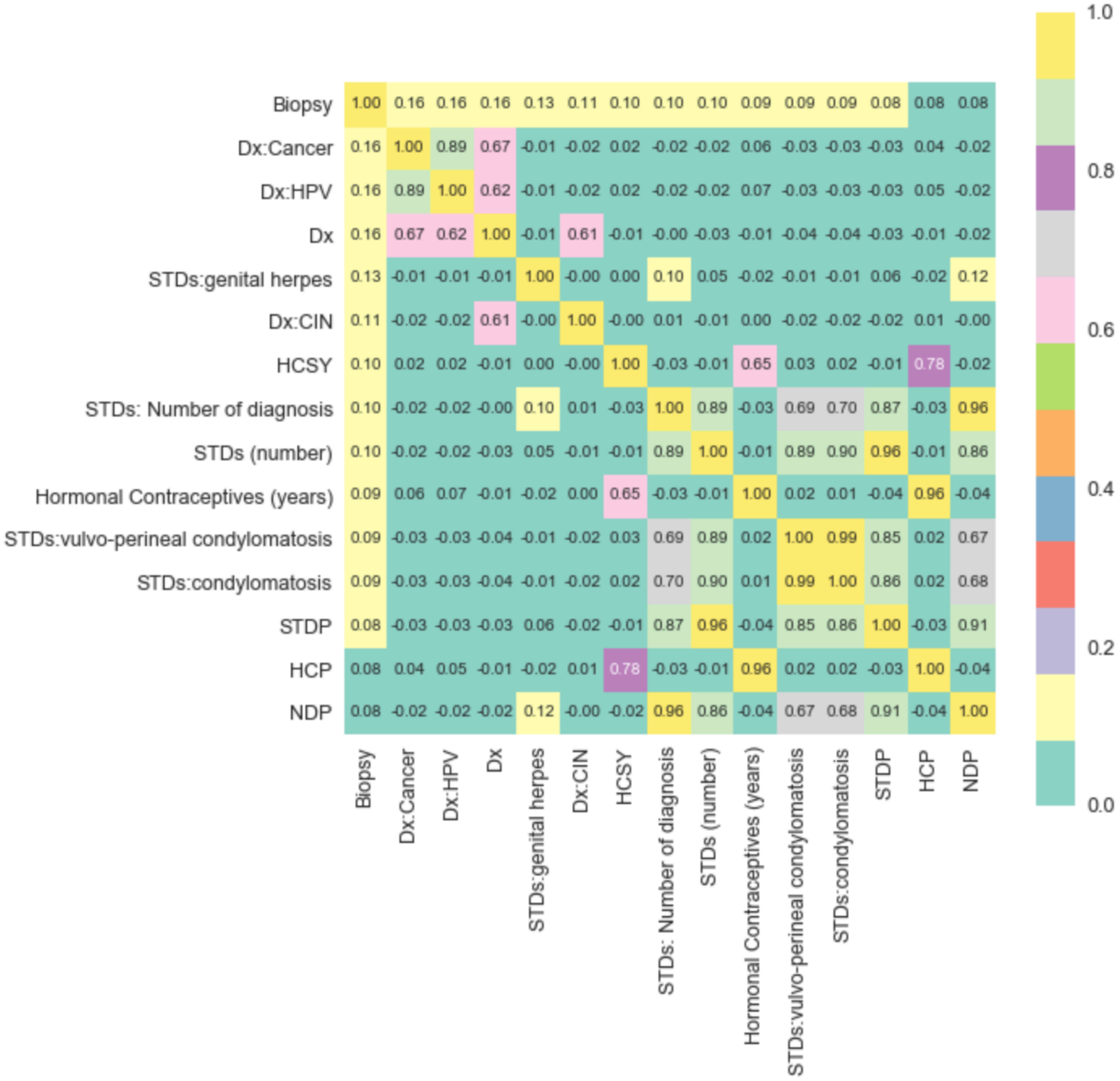
Feature engineering is performed to create new interesting features that could potentially care more weight in the analysis.

Here are some of the features that were created

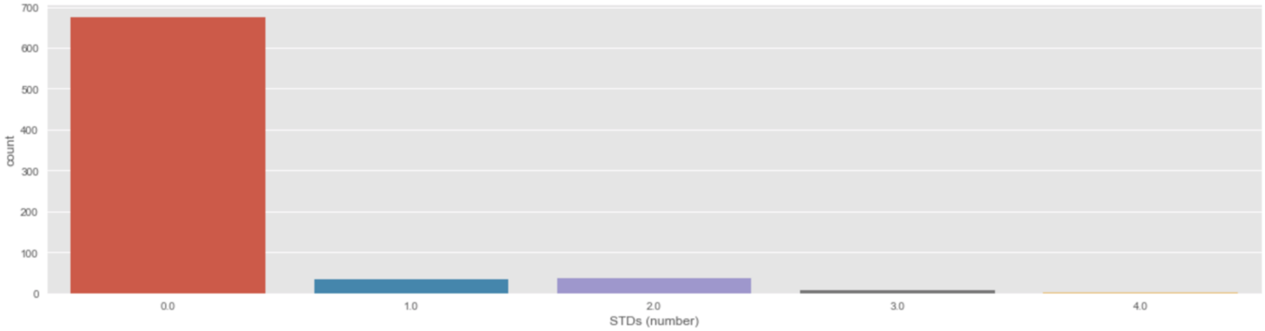
- YAFSI : No. of years since the patient had first sexual intercourse
- SSY : No. of years patient did not smoke
- SPYP : % of Partners after first intercourse over the years
- SP: Smoking percentage over age.
- HCP : Hormonal Contracepticès percentage over age
- STDP: STDs percentage over age
- IUDP: IUD percantage over age
- TSP : Total No. of cigarettes of patient smoked
- NPP : Number of pregnancies percantage over age
- NSPP: Number of sexual partners percentage over age
- NDP : Number of STDs diagnosis percentage over age
- TBD : Time betweem diagnosis
- YAHC : No. of years patient didn't take Hormonal Contraceptives
- YAIUD: No. of years patient didn't take IUD
- NPSP : Average pregnancy over one sexual partner
- IUDSY: No. of years patient takes IUD after first sexual intercourse percentage
- HCSY : No. of years patient take Hormonal Contraceptives after first sexual intercourse percentage

Feature Importance and Feature Correlation

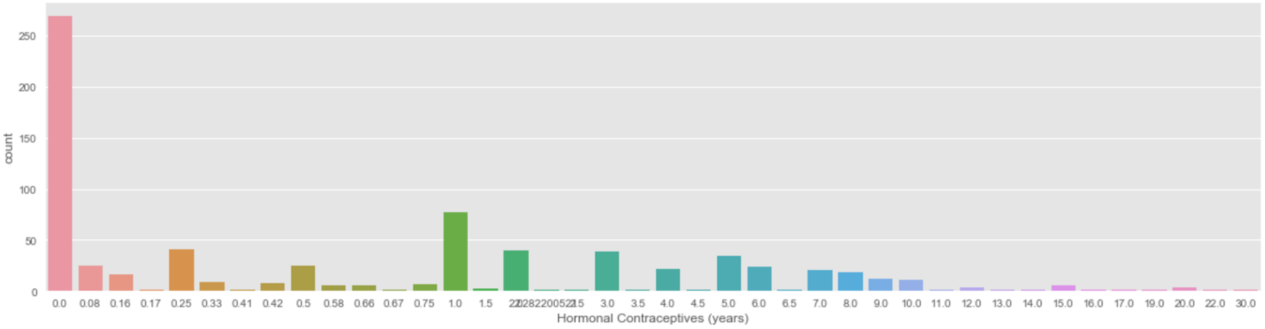

Hormonal Contraceptives (years)
HCP
NSPP
YAIUD
NPSP
YAHC
HCSY
First sexual intercourse
SPYP
Age
SSY
NPP
Number of sexual partners
YAFSI
IUDSY
Num of pregnancies
Smokes (years)
Dx:Cancer
STDP
NDP
Dx:HPV
STDs: Time since last diagnosis
SP
Dx
STDs:genital herpes
Smokes (packs/year)
TSP
STDs:vulvo-perineal condylomatosis
IUDP
Dx:CIN
STDs: Time since first diagnosis
Smokes
STDs (number)
STDs: Number of diagnosis
IUD (years)
IUD
STDs
Hormonal Contraceptives
STDs:condylomatosis
STDs:HIV
STDs:syphilis
STDs:Hepatitis B
STDs:AIDS
STDs:molluscum contagiosum
STDs:pelvic inflammatory disease
TBD
STDs:vaginal condylomatosis
STDs:cervical condylomatosis
STDs:HPV
Features
0.00
0.01
0.02
0.03
0.04
0.05
0.06
mean(Values)

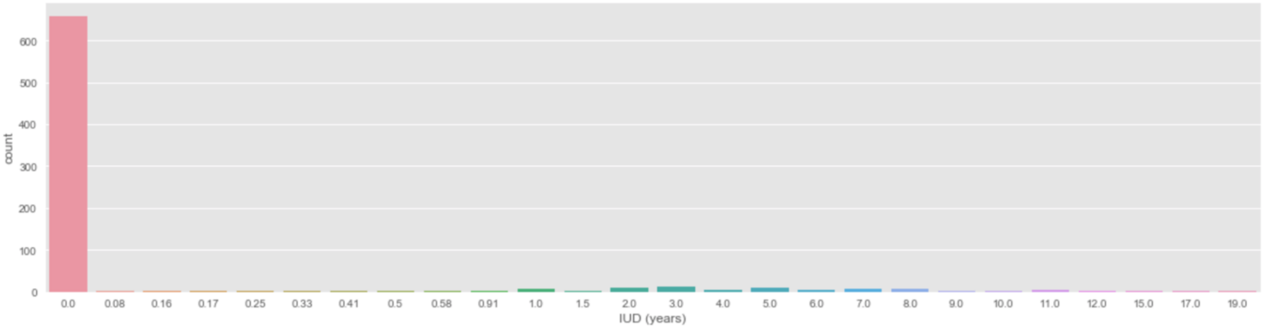


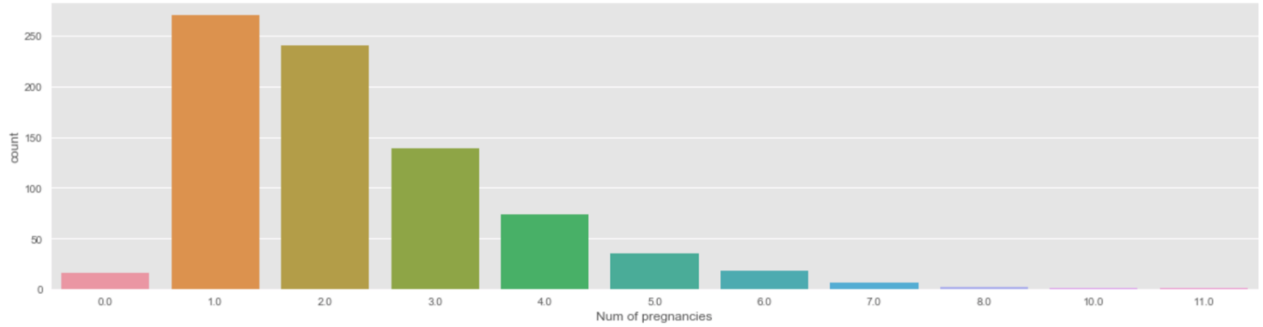
Exploratory Data Analysis

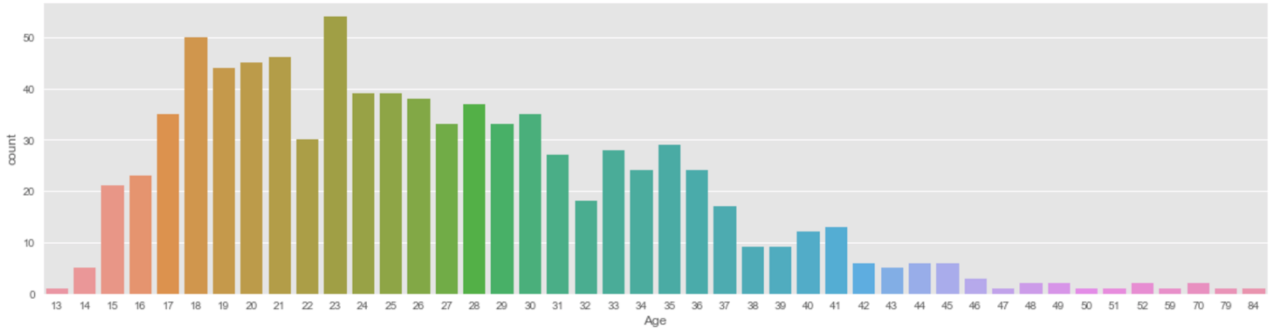


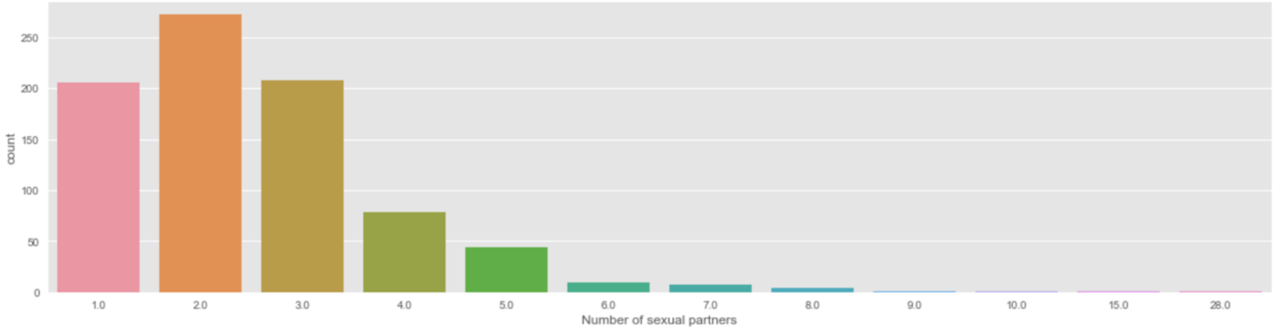
Most patients don't have any STDs, hence making it challenging to correlate STDs to any screening technique.

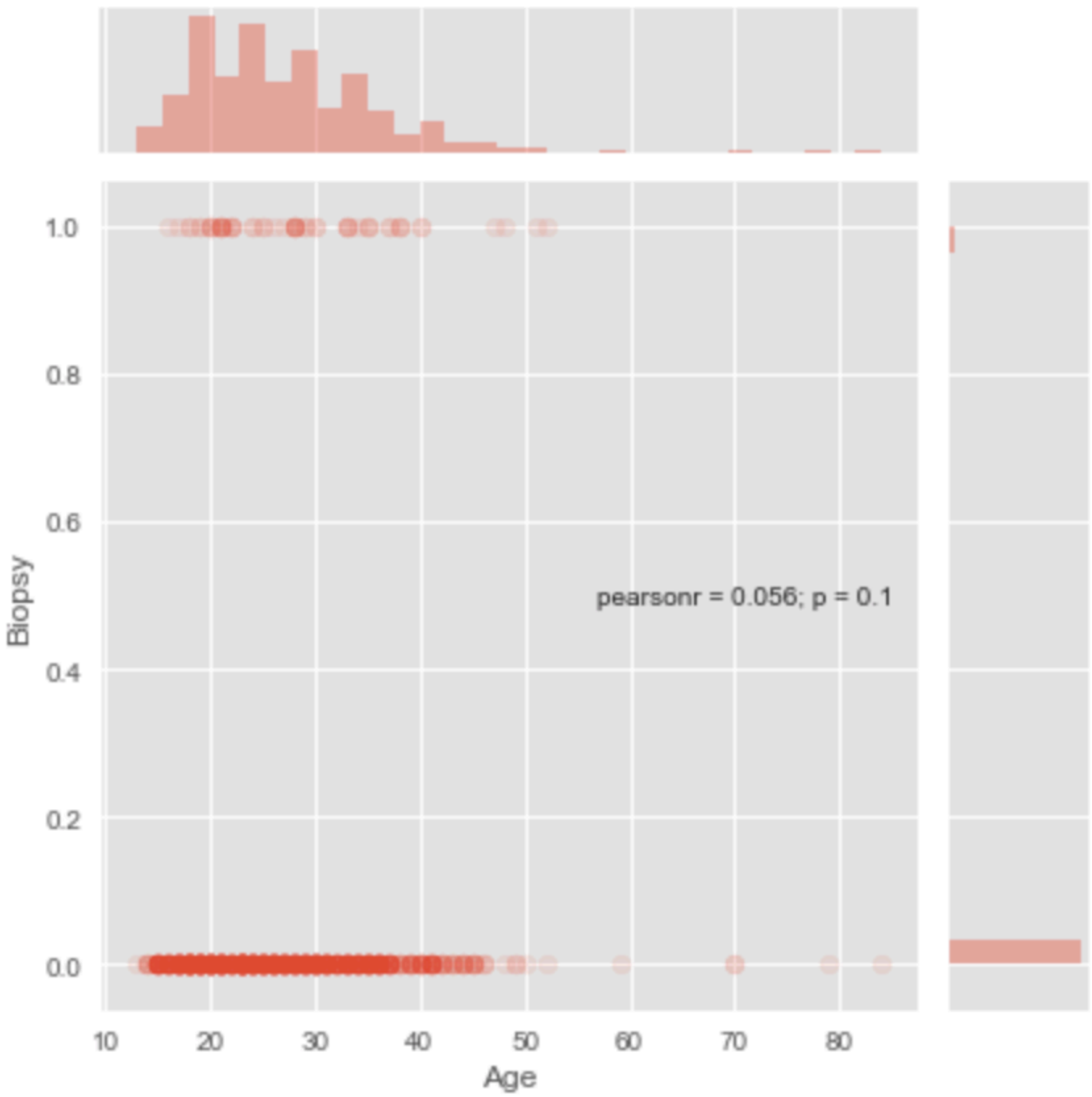




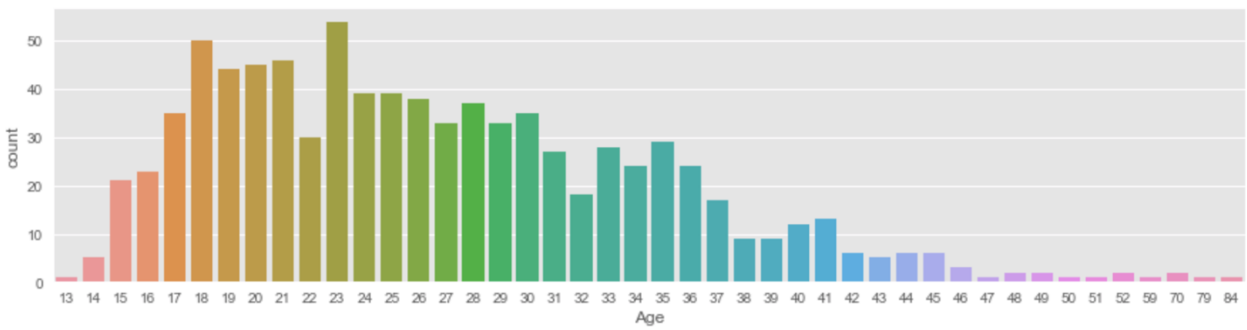



count
0
10
20
30
40
50
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
59
70
79
84
Age

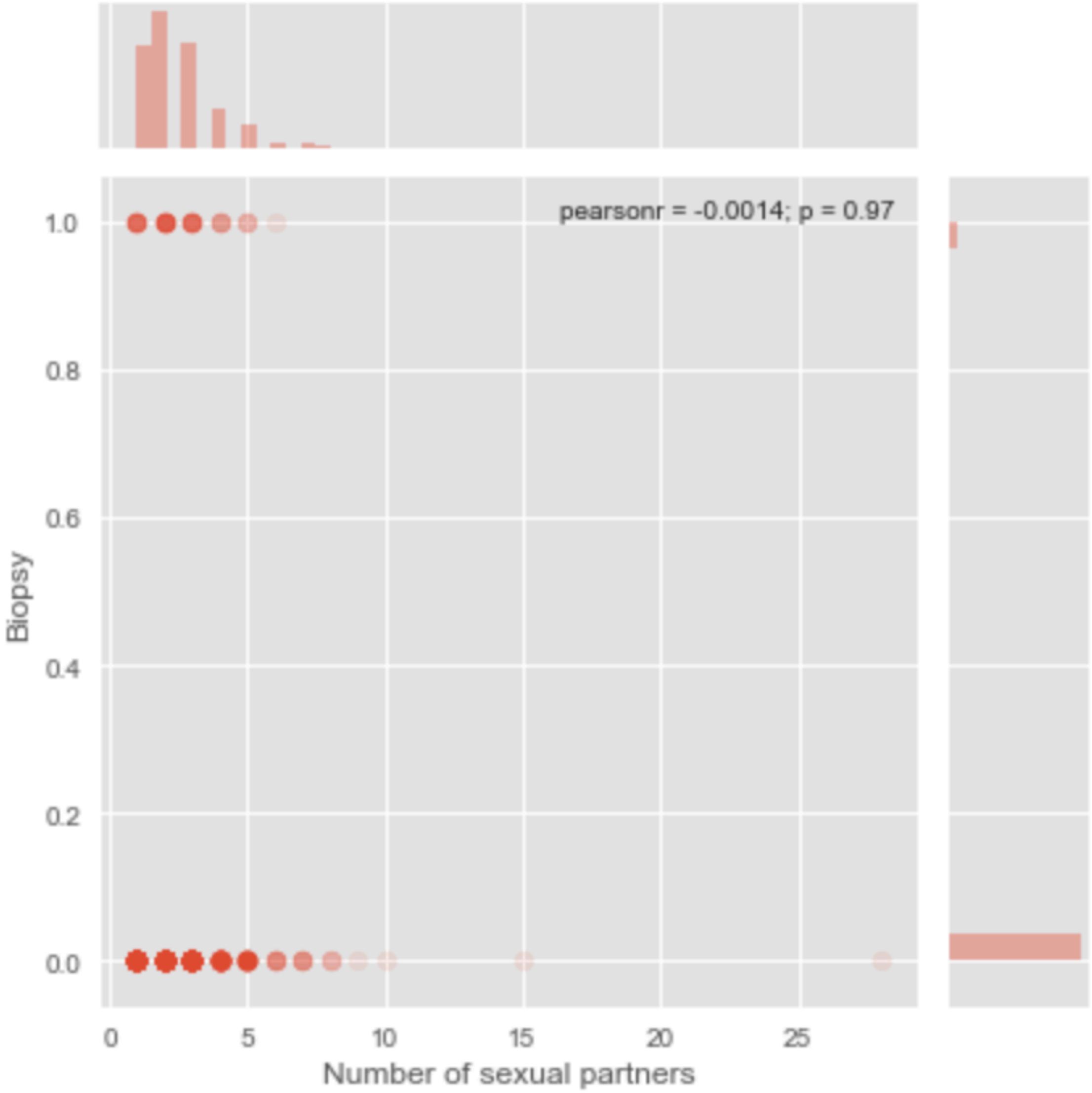

count
0
50
100
150
200
250
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
15.0
28.0
Number of sexual partners

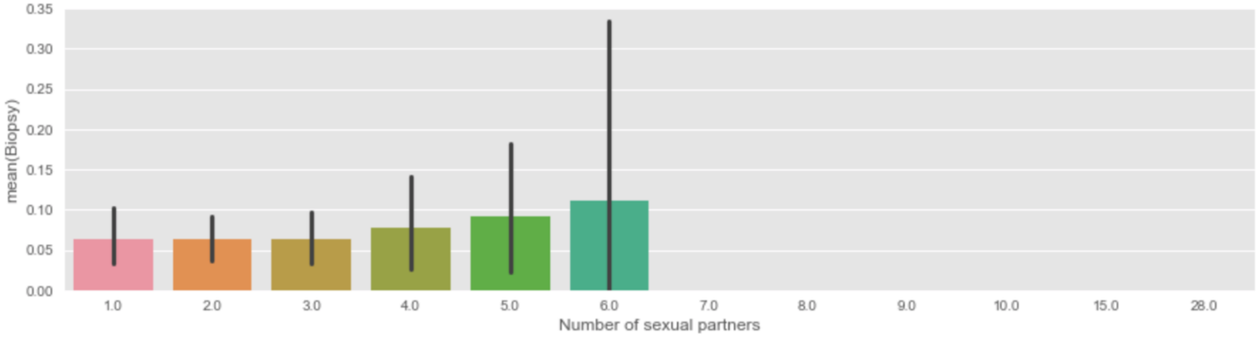


Age has a good positive correlation with most of the patients in the age group of 20-35 requiring biopsy



Against our intuition, number of sexual partners has a low correlation with Biospy





Machine Learning, Model Selection

Training a model on imbalanced datasets is always super challenging and the best way to handle it is oversampling. SMOTE is used on the training dataset to improve our scoring metrics and fit a useful model.

We further use GridSearchCV to hypertune the parameters. This improves efficiency and helps us compare and select the best model.

We use F1 Score and ROC AUC Score to evaluate the models.

Results

	F1 Score	ROC_AUC
Random Forest	0.210	0.639
Adaboost	0.076	0.515
KNearest Neighbour	0.384	0.614
Decision Tree	0.384	0.689
Logistic Regression	***0.378***	***0.759***

Conclusion

Imbalanced datasets tend to give poor results but building custom scoring fuctions and using class weights can improve the model.

Also, dropping features and comparing different models might be a great improvement.

From the implemented models, Logistic Regression seems to yield the best results with an F1 score of 0.378 and ROC_AUC of 0.759.

In the future,

- A cost function can be built with the right thresholding.

- More patient records to avoid oversampling.