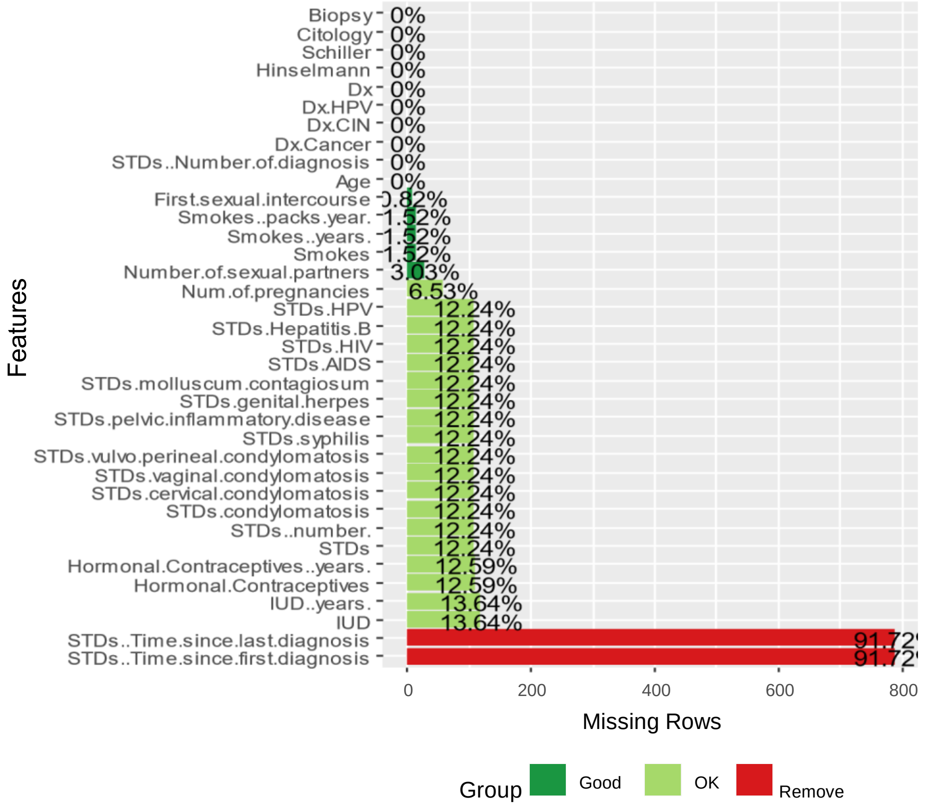
Introduction

My data set is taken from a hospital ward in Caracas, Venezuela. There are 858 patients, all the patients have been diagnosed with cervical cancer and hence are all female. There are 36 attributes of which 4 are target variables. The target variables are four screening strategies, namely: Hinselmann, Schiller, Cytology and Biopsy [4]. Cancer research is a well-developed scientific field with copious amounts of research, but risk classification is still an ongoing process that requires further research. It has been estimated that 493,000 develop the disease every year and of which 273,500 die from it [5]. In developing countries, it has been estimated to account up to 85% of each country's morbidity and mortality [5]. Cervical cancer is a tumour that is formed in the cervix, the tumour is an effect of rapid and uncontrolled growth of cells in the cervix [5]. In [5], several risk factors were considered and of which;

1. multiple sexual partners
2. Human immune deficiency virus positive
3. sexually transmitted infection history
4. early age at initiation of sexual intercourse

have been proven to have a statistically significant association with the onset of cervical cancer [5]. In my analysis, I will consider an additional 31 variables. Before analysing the data, column deletion will have to be performed as there are columns missing significant amounts of data; *Figure produced in R*.



After initial feature exploration in R, It is clear from above image that

‘’STDs:Time.since.last.diagnosis’’ and ‘’STDs:Time.since.first.diagnosis’’ have to be omitted as they have 91.7% of their data missing. For the remaining features, imputation will be made for missing values. I will replace missing values of continuous variables with its column mean and columnmedian for Integer/Boolean variables, this is done to limit bias. We have a case of ‘MAR’ (missing at random) in our data set as patients did not respond due to privacy concerns [4]. In [4], MAR is explained as; “Missingness on any variable in the analytic model may be explained solely using observed variables in the model. Unobserved variables do not explain missingness in any variable in the model” [4]. This holds in our data set as missingness in the second column omitted is explained by missingness in the first column omitted as both were left unanswered due to privacy issues. Also, imputation is ideal when non-responses does not exceed 20% for each Feature, provided the data set is sufficiently large.

I will be using a Random Forest Regressor Algorithm for my analysis, the target variables are recorded as a binary outcome and I will have to focus on only one dependent variable, namely; Biopsy as it is the most widely used screening strategy.

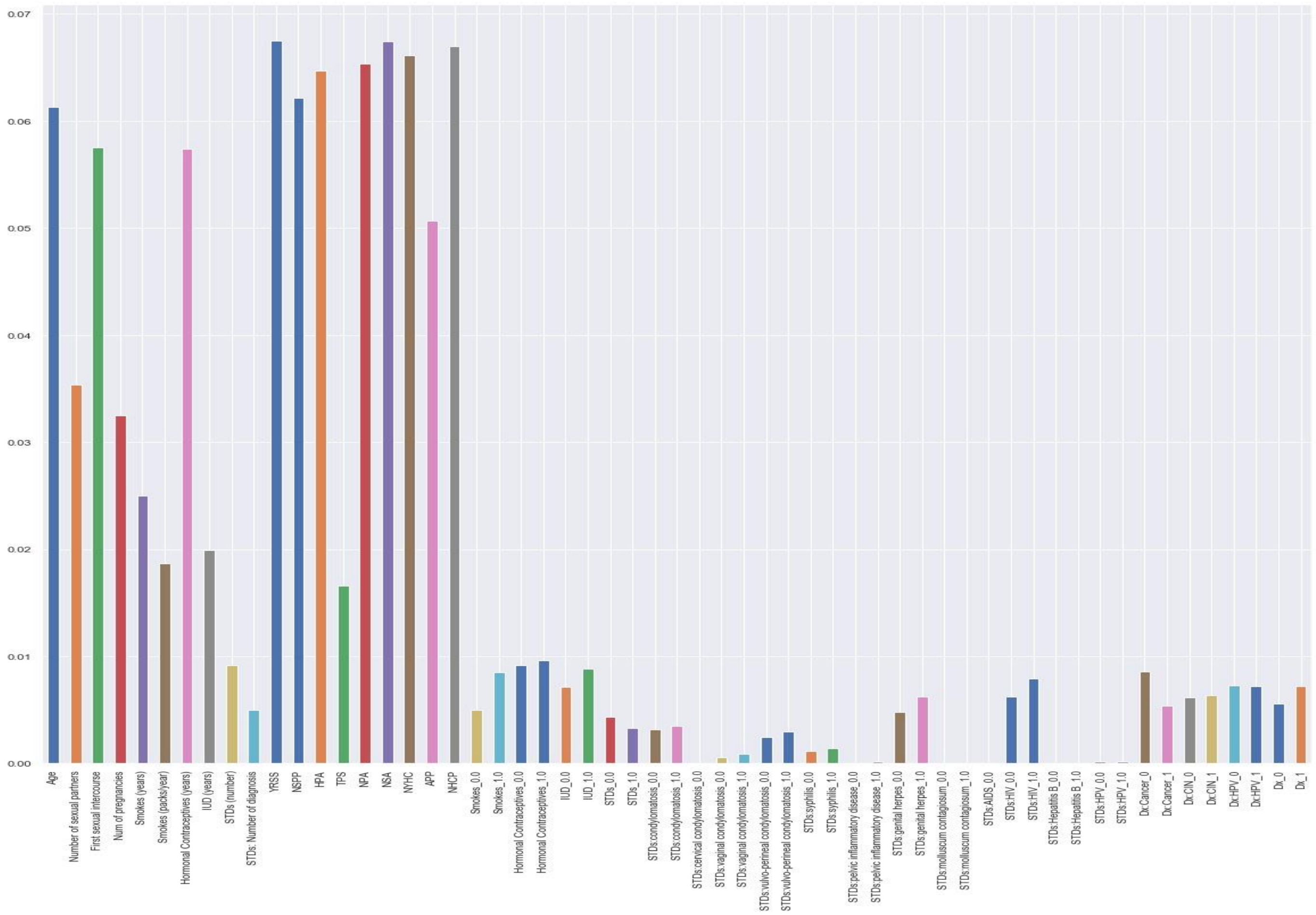
Conclusion

This dataset had a significant portion of its entries missing, hence data pre-processing was crucial for a Random forest Regressor to be performed. The missing values was deemed to be ‘Missing At Random’ according to [4]. The count of positive readings of the dependent variable was problematic as there were only 55 positive detection of cervical cancer by Biopsy method and 803 negative cases, also on many instances positive readings for Biopsy coincided with missing values for other attributes which in my opinion reduced goodness of fit of model. This dataset had 4 dependent variables that were non-congruent regarding cervical cancer detection, for a better analysis of risk factors it would be ideal to have one accurate screening method as this would allow more focus on analysing risk factors.

After a careful inspection of correlation maps, we could see that ‘Age’ was correlated with ‘number of pregnancies’. There are many cases of correlations between features, but they are noninformative as their correlation is essential and intuitive., such as; ‘STDs(number)’ correlated with ‘STDs’ and ‘IUD’ correlated with ‘IUD(years)’ and so on. The dataset was sampled in a way that had no concern for multicollinearity of features. As an example, there are 12 different types of STD’s named and they are all correlated with ‘STDs’, ‘STDs(number)’ and ‘STDs number of diagnosis’ to varying degrees. Having features that are not correlated would substantially improve the process of identifying cervical cancer risk factors. We also saw that Hormonal Contraceptives and Age, IUD and Age influences Biopsy screening method, the reason for this effect is noteworthy and could mean either that they are valid risk factors or that they only increase detection of cervical cancer which implies that screening method is inadequate.

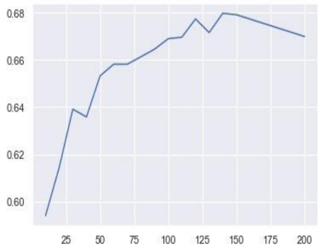
In [5], Human papilloma virus (HPV) is given as the reason for cervical cancer but according to our investigation, it seems not to be detected at all by biopsy screening method, suggesting it has no effect on cervical cancer or that the screening method is very inaccurate, given the extensive research conducted on HPV and cervical cancer it seems more plausible that our screening method has defects. After first model fit, we obtained an R^2 = 0.00317, this R-value is very low, we will begin to optimise a C-stat which is a measure of goodness of fit for binary outcomes in a logistic regression model. It will give us the probability a randomly selected patient who has experienced cervical cancer risks has a higher risk score than a patient who has not experienced the risks.

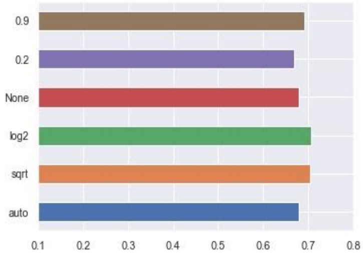
Initial C-stat calculation gives; 0.66452, C-stat ranges from 0.5 to 1, a value above 0.7 is considered a good model fit. We will try optimizing the model to obtain a C-stat > 0.7. I have introduced dummy variables for all the categorical variables, I have also introduced 9 columns created from existing attributes and have plotted below a feature importance graph.

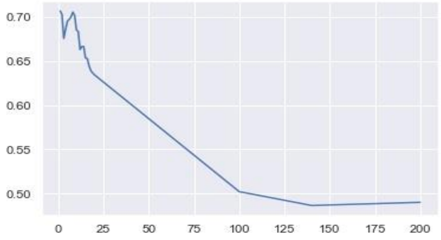


The 6 highest bars above are all the newly created columns, this puts the remaining risk factors into perspective, that is their weak correlation with biopsy result. These new columns are somewhat correlated with a positive reading on a Biopsy screening method, it is a very weak correlation. After introducing dummy variables for all categorical variables and running RF regressor algorithm, I obtain C-stat = 0.688. This is a 3.4% improvement but still below threshold. To optimise our existing model, I must find;

1. Optimal Number of trees for RF
2. Optimal maximum number of variables RF can test in each node
3. Optimal minimum number of samples required to be at each node







The 3 graphs are in order of list above, we can clearly see that 140 trees, log2 and 1 sample is our optimum for our final model. The final C-stat is; 0.707 which is just over the threshold of; 0.7.

After my investigation and analysis of the dataset, the main problem I encountered is that of confounding variables, i.e. variables that influence both the dependent variable and independent variables. My analysis has also shown me many instances of multicollinearity between variables. The ratio of positive to negative Biopsy readings was hugely exaggerated, that is very little positive readings. The inclusion of known and proven risk factors in this dataset was also a hindrance. A dataset with an accurate detection method of cervical cancer, that has no confounding/multicollinearity of variables would be best suited for risk analysis.