Predicting Cases of Cervical Cancer in the 'Hospital Universitario de Caracas'

Ben Herndon-Miller and Nicole Jaiyesimi

11/25/2018

Introduction

Data Description

The data was collected at Hospital Universitario (which is located in Caracas Venezuela) in 2017. There are 36 attributes in the dataset, all of which are either integers or booleans (e.g. age, number of sexual partners, whether or not the individual smokes, whether or not the individual has AIDS). There are 858 instances of the 36 attributes, but there are many missing values; we will conduct a coverage analysis of the data in our final project.

Furthermore, the description of the dataset does not include detailed information on a variety of variables. First and foremost, there are 4 variables that could be considered the outcome variable of interest for a diagnosis of cancer: *Hinselmann*, *Schiller*, *Citology*, or *Biopsy*. These all represent different screening techniques for identifying and diagnosing cervical cancer. Since we need to determine a single outcome variable to predict, we proceed using the binary variable *Biopsy* as it is the gold standard for diagnosing cervical cancer. We will discard the other potential outcome variables from our dataset before we proceed.

First, let us examine the number of missing values for each variable before we proceed.

Coverage Analysis and Missing Data Methodology

Variable	Variable_Type	Missing_Values	Percent_Missing
Age	int	0	0.0000000
Number.of.sexual.partners	int	26	0.0303030
First.sexual.intercourse	int	7	0.0081585
Num.of.pregnancies	int	56	0.0652681
Smokes	bool	13	0.0151515
Smokes..years.	int	13	0.0151515
Smokes..packs.year.	int	13	0.0151515
Hormonal.Contraceptives	bool	108	0.1258741
Hormonal.Contraceptives..years.	int	108	0.1258741
IUD	bool	117	0.1363636
IUD..years.	int	117	0.1363636
STDs	bool	105	0.1223776
STDs..number.	int	105	0.1223776
STDs.condylomatosis	bool	105	0.1223776
STDs.cervical.condylomatosis	bool	105	0.1223776
STDs.vaginal.condylomatosis	bool	105	0.1223776
STDs.vulvo.perineal.condylomatosis	bool	105	0.1223776
STDs.syphilis	bool	105	0.1223776
STDs.pelvic.inflammatory.disease	bool	105	0.1223776
STDs.genital.herpes	bool	105	0.1223776
STDs.molluscum.contagiosum	bool	105	0.1223776
STDs.AIDS	bool	105	0.1223776

Variable	Variable_Type	Missing_Values	Percent_Missing
STDs.HIV	bool	105	0.1223776
STDs.Hepatitis.B	bool	105	0.1223776
STDs.HPV	bool	105	0.1223776
STDs..Number.of.diagnosis	int	0	0.0000000
STDs..Time.since.first.diagnosis	int	787	0.9172494
STDs..Time.since.last.diagnosis	int	787	0.9172494
Dx.Cancer	bool	0	0.0000000
Dx.CIN	bool	0	0.0000000
Dx.HPV	bool	0	0.0000000
Dx	bool	0	0.0000000
Biopsy	bool	0	0.0000000

We can see from this coverage analysis that the majority of our variables have missing data points. However, most features are not missing a high-proportion of values, except for Time Since First/Last Diagnosis of STD. These are all missing for the patients who stated that they had not had any STD's, so it would be inaccurate to simply replace them with 0 as that would bias the model towards more patients having a small time period since their STD. Replacing them with the mean or median value would also not make sense since the true value would be infinity since they have never had a diagnosis. For the sake of simplicity, we will move forward by removing these two variables from our analysis.

However, we still must deal with the missing values for the other variables. For the integer variables of *Age of first sexual intercourse*, *Number of sexual partners*, and *Number of pregnancies*, we can replace the missing values with the median values of their respective variables. In contrast, for the variables pertaining to smoking, STD's, and contraceptives, we cannot simply replace the missing values with the median value as they are based on boolean values. Thus, we will then remove the rows with missing values for the missing variables and report

After cleaning the data with the above methodology we are left with a dataset of 726 observations and 31 features (30 predictors). We can now move forward with exploratory analysis and predictive modelling.

Exploratory Data Analysis

Predictive Modelling