


3rd year of Economist-Statistician Magistere's degree
M2 Case study number 2
Session 2021-2022

Cervical Cancer Risk Classification

Gibbs Sampler for Supervised Learning

Charlotte Bredy-Maux & Kayané Robach

Abstract

Cancer of the Cervix Uteri, commonly known as Cervical Cancer is one of the leading gynaecological malignancies in the world [Juneja et al., 2003] both in terms of incidence and mortality. To such an extent, the World Health Organization reported 604 000 new cases worldwide and about 342 000 new deaths from the said disease in its 2020 survey (versus 570 000 and 311 000 respectively in 2018). These numbers have been historically linked to Human Papillomavirus (HPV) infection throughout Odd-Ratio-based studies [Castellsagué, 2008]. Still, major discrepancies exist in the scientific literature around the epidemiological profiles and parthenogenesis of cervical cancers [Gillison et al., 2014]. The present study asks how a broader range of behavioral factors interact in risking cervical cancer, conditional to prior HPV infection. Results obtained on recent data suggest that exhaustive work on pre-processing and distributional analysis can highly improve predictions F1-scores from machine learning pipelines. The baseline F1-score stands at 0.18 on the original dataset with a decision tree. F1-score of 0.33 is attained using an extra trees on a spectral transformation of the original data; a F1-score of 0.2 is attained on processed data using Gibbs' algorithm sampler.

Key words: Cervical Cancer, HPV, Random Forest, Machine Learning, Classification, Neural Network, Decision Trees, SVD, Gibbs' Algorithm

Contents

1 Introduction

The histopathology of cervical cancer highlights the prevalence of Squamous Cell Carcinomas (SCC) within all reported cases i.e. 80 to 85% [Vinh-Hung et al., 2007].

As previously stated, central etiologic factor for the development of cervical cancer lies on persistent infection with high-risk oncogenic HPV types. Furthermore, other recognized risk factors for cervical cancer are identified by clinicians, including:

- increased potential exposure to sexually transmitted diseases
- histoy of vulvar or vaginal squamous intraeouthelial neoplasia (please figure charlotte)
- oral contraceptive use
- smoking status
- immunodeficiency

As such, cervical cancer has been historically described as a behaviorally-led disease by the literature, being particularly prevalent in India, low-developed African countries and Latin America. Recent studies front to such epidemiological view and define a novel interpretation of the cancer epistemological paradigm [Brücher and Jamall, 2014]. In particular, the Somatic Mutation Theory (SMT) explaining carcinogenesis seems to fall short at explaining increasing infection rates amongst individuals. More specifically, the clinical profile of patients still fits the middle-aged, multiparous, smoker portrait previously known. Notwithstanding, the disease's etiology becomes glowingly unknown to practitioners, in spite of the mass vaccination and prevention campaigns against HPV held since the 2000's [Franco et al., 2001], making it a 'sporadic' cancer. Hence, the present study ought to address the looming imprecision around the stated risk factors acting as stand-alones or interaction patterns builders. More specifically, the empirical strategy parts from the classical odd-ratio approach and focuses on machine learning classification as a built-in forecasting method on an unbalancedness target. Hence, the Material and Methods for such analysis will be presented in Section III. The Empirical Strategy will be thereafter described in Section IV, prior to the Machine Learning Modelling and Analysis Section. Finally, Conclusion and Limitations will be found in Section VI. The reader is kindly invited to refer to all sources and appendix material in Section VII and VIII respectively.

2 Material and Methods

The present work aims at enhancing cervical cancer risk classification using machine learning models on observed and synthesized data. We evaluate our classifiers based on the F1 score.

		Predicted label	
		Non cancer	Cancer
True label	Non cancer	**T**rue **N**egative	**F**alse **P**ositive
	Cancer	**F**alse **N**egative	**T**rue **P**ositive

Table 1: Confusion matrix

The F1 score is the harmonic mean of precision[1] (rate of correct positive predictions) and recall[2] (rate of correctly predicted positives). We can compute this score using the confusion matrix according to the following formula:

$$\text{F1 score } = \frac{\text{TP}}{\text{TP} + \frac{\text{FN} + \text{FP}}{2}}$$

The main challenge of this research work is to alleviate the unbalance of the data. Indeed very few patients have cervical cancer, hence the focus on positive predictions.

2.1 Research Questions

How to counteract data imbalance to enhance precision and recall ? How to use correlations in the explanatory variables to produce synthesized affected patients ?

2.2 Data Presentation: Sample and Sources

We get a dataset of 838 people with uterus from Kaggle [3] These patients are described with 34 explanatory variables that we might group into 5 categories:

- sexual behavior,
- smoking status,
- contraception,
- sexually transmitted diseases (STD),
- medical check-ups

The average patient is 26 years old, until the study she had 2.5 sexual partners, the first one around her 17 years old. She has 2 or 3 children, she smoked for 1.21 years 0.45 cigarette packs per year. She took hormonal contraception for 2.35 years, she had very little or no STD and she did very little medical check-ups.

98% of the data represents healthy patients.

First, we focus on 12 variables of interest: age, number of sexual partners, first sexual intercourse, number of pregnancies, years spent smoking, packs of smoked cigarettes per year, number of cigarettes packs smoked, smoking while on hormonal contraceptive, years spent with an hormonal contraception, years spent with an intrauterine device (IUD), number of sexually transmitted diseases (STD), number of STD diagnosed.

[1] $\frac{TP}{TP+FP}$
[2] $\frac{TP}{TP+FN}$
[3] https://www.kaggle.com/code/naveenram/ml-models-for-prediction-of-cervical-cancer/data

3 Empirical Strategy

3.1 Data Pre-Processing

As aforementioned, the data includes extensive descriptors of behavioral patterns, medical history and screening processes, albeit being often redundant. A first line of arbitrages made on the said dataset concerns the selection of the most relevant variables for each set of descriptors. In particular, and in conformity with the literature [Åhren et al., 2000], the strategy acknowledges the consensual or potential tobacco dose-effects driving risk at the individual level. Hence, the smoking status will be proxyed by the number of cigarette packs consumed per year and the number of years spent smoking. Likewise, oral contraception will be considered from the point of view of number of years spent under such and we will take into account an interaction term between smoking status and taking hormonal contraception. In particular, Intrauterine Device is refereed to as the copper non-hormonal version. All STD will be accounted for within a single dummy variable equal to 1 if the patient has been reportedly infected by any STD at the time of the survey. Finally, and most importantly, all analysis will be done conditionally to HPV infection and unconditionally to screening processes. Indeed, such medical examinations are often performed as a prognosis confirmatory procedure and not an agnostic exploratory one due to their high cost. Therefore, selection bias can occur and be translated into the data when predicting actual cancer if based on cytologies as well as Hinselman's and Schiller's examinations. Further data analysis and exploratory statistics will be presented within the following sections.

Exploring the data we noticed a kind of 'outlier' in the subsample of affected patients. Only two women smoked in this subsample, one said she was smoking 37 packs per year for 37 years. We suspect an error in this information and decided to set the number of packs smoked per year at 8.4 for this observation. This value corresponds to the average cigarette packs consumed per year in the healthy dataset.

3.2 Data Analysis

3.2.1 Descriptive Statistics

The first tools of analysis provided to the reader refer to the stratification of the dataset, in order to accurately appreciate its unbalanced nature. Indeed, less than 2% of patients are diagnosed as cancer-positive. Namely, position and dispersion statistics are computed within each subset of cancer-positive and cancer-negative samples.

All difference-in-mean and in-variance hypothesis will be tested using the Non-Parametric Mann-Whitney U-Test.

3.2.2 Inferential Statistics

As previously stated throughout the Descriptive Analysis subsection, and beyond the unbalancedness of the dataset, one might wonder whether intrinsic differences exist between the cancer-positive and cancer-negative subsamples' data generative processes. The rationale in performing inferential statistics at this stage of the empirical strategy is twofold. First, finding an absence of differences in distributions might indicate insufficient information in the set of behavioral variables at hands to draw sensible decision rules for risk prediction. This argument rejoins the most recent literature Brücher and Jamall [2014] on cervical cancer etiology and parthenogenesis i.e. the sporadic or environmental oncogenetic paradigms. Furthermore, studying differences create a clear benchmark

for the synthetic dataset to be compared with later on using the Gibbs' Algorithm (which will be presented in the following subsections).

From an agnostic point of view on the data, and given the multimodal distribution functions observed during univariate analysis, distributions between cancer-positive and cancer-negative subsamples are compared with a Non-Parametric Mann-Whitney U-Test. Usual T-test and F-test for difference in means and variances are used when the distributions are gaussian which is not the case here. Formally speaking, the test relies on similar rationales as the Wilcoxon Test and works on observation's ranks. In the population framework, for P_1 and P_2 two populations for which one ought to test differences in position and dispersion, $X = (X_1, X_2, ..., X_n)$ i.i.d random variables issued from P_1 and $Y = (Y_1, Y_2, ..., Y_m)$ its counterpart in P_2:

- Step 1: One computes the sample X and the sample Y and ranks each observation

$$X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m$$

  in increasing order to get for e.g. something of the following form

$$X_{[1]}, Y_{[1]}, X_{[2]}, X_{[3]}, Y_{[2]}, ..., X_{[n]}, Y_{[m]}$$

- Step 2: One then counts the number of times $X_i > Y_j \ \forall i \in [\![1, n]\!], \forall j \in [\![1, m]\!]$, we denote this count U_X.

- Step 3: Likewise one computes the number of times $Y_j > X_i \ \forall i \in [\![1, n]\!], \forall j \in [\![1, m]\!]$, we denote this count U_Y.

The test hypothesis indicates:
H_0: The distributions generating X in P_1 and Y in P_2 are equal
versus
H_1: The distributions generating X in P_1 and Y in P_2 differ
The test statistics equals $U = \min\{U_X, U_Y\}$ and, for large samples, U is approximately normally distributed $U \sim \mathcal{N}(m_U = \frac{n \cdot m}{2}, \sigma_U = \sqrt{\frac{n \cdot m \cdot (n+m+1)}{12}})$. At the α Type I error rate

$$\exists! \ a \in \mathbb{R} \text{ such that } \mathbb{P}(|u_{obs}| \leq a) = 1 - \frac{\alpha}{2}$$

Then the null hypothesis is rejected if and only if $u_{obs} \notin [-a, a]$.

Applying the present test theory in its bilateral version to the data indicates that different distributions generate the following random variables among cancer-positive and cancer-negative samples:

- Age (p-value = 0.00095)
- Age of the first sexual intercourse (p-value = 0.00235)
- Number of years spent under a copper intrauterine device (p-value = 0.00216)

The said differences can be better appreciated with the boxplots of the triplet of variables according to oncologic status as follow:

Disparity between healthy and affected samples thus depends on these variables but this is probably due to the unbalance in our dataset. Therefore, we need to synthesize data with a Gibbs

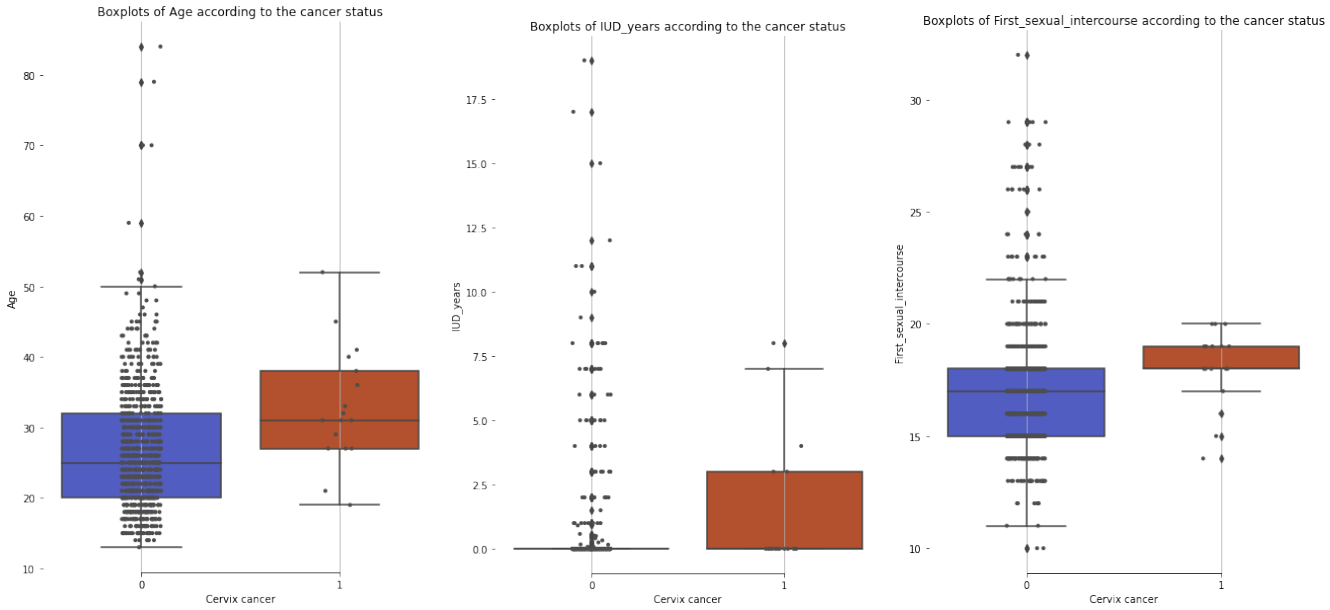

Figure 1: Boxplots for the Age (left), duration spent under IUD (middle), the first sexual intercourse (right) distributions according to oncological status

sampler, and more specifically cancer-positive profiles, to balance these differences in age, age of the first intercourse and the duration wearing a intrauterine device (IUD), but also to be able to predict cancer-positive profiles.

In the next subsections, the reader would be ale to observe the differences in distribution between observed and synthetic data. Before comparing both -observed and synthesized- distributions, we will focus on the Gibbs sampler algorithm in the following subsection.

3.2.3 Gibbs Algorithm

Patient k is represented by a vector of n covariates $(X_1, X_2, \ldots, X_n)$. Sampling affected patients is non trivial since we have no direct way to synthesize data; indeed we do not know the 'theoretical' joint distribution of our explanatory variables among patients affected from cervical cancer.

Gibbs sampler uses correlations among the covariates to sample from the multivariate distribution.

For instance, consider n explanatory variables $X_1, X_2, \ldots, X_n$ that are highly correlated which means we suppose that $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, X_n$ can help predict X_i.

$$\begin{aligned}\mathbb{P}(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) &= \frac{\mathbb{P}(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n)}{\mathbb{P}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)} \\ &\propto \mathbb{P}(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n)\end{aligned}$$

We want k samples from the joint distribution $\mathbb{P}(X_1, X_2, \ldots, X_n)$.

- The algorithm needs to start with an initial patient, initial vector $({X_1}^{(0)}, {X_2}^{(0)}, \ldots, {X_n}^{(0)})$.
- We want to produce the next sample $({X_1}^{(l+1)}, {X_2}^{(l+1)}, \ldots, {X_n}^{(l+1)})$ therefore we will draw each variable ${X_i}^{(l+1)}$ based on the components we produced so far, that is to say based on ${X_{j<i}}^{(l+1)}$ and ${X_{j>i}}^{(l)}$: ${X_i}^{(l+1)} \sim \mathbb{P}({X_i}^{(l+1)}|{X_1}^{(l+1)}, \ldots, {X_{i-1}}^{(l+1)}, {X_{i+1}}^{(l)}, \ldots, {X_n}^{(l)})$.

- We repeat the process k times.

To summarize, we draw each variable X_i (k times) from the joint distribution of $(X_1, X_2, \ldots, X_n)$ using the conditional distribution of $(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ (and to do so we use a catch since, at step $l+1$ when drawing the i^{th} covariate, we know the covariates for $j < i$ from step $l+1$ and the covariates for $j > i$ from step l).

We now need more details to understand the main step. To sample

$$X_i{}^{(l+1)} \sim \mathbb{P}(X_i{}^{(l+1)}|X_1{}^{(l+1)}, \ldots, X_{i-1}{}^{(l+1)}, X_{i+1}{}^{(l)}, \ldots, X_n{}^{(l)})$$

imagine that we know some prior about this conditional distribution, for e.g. we assume a gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Thus we estimate parameters μ and σ by maximum likelihood which allow us to draw an observation $(X_i)^{(l+1)} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. Without more information we will always assume a gaussian distribution since it works well.

3.3 Sampling

Correlation Analysis further investigates the relationship between the predictors and altogether serve as empirical rationales for Gibbs sampling and data balancing.

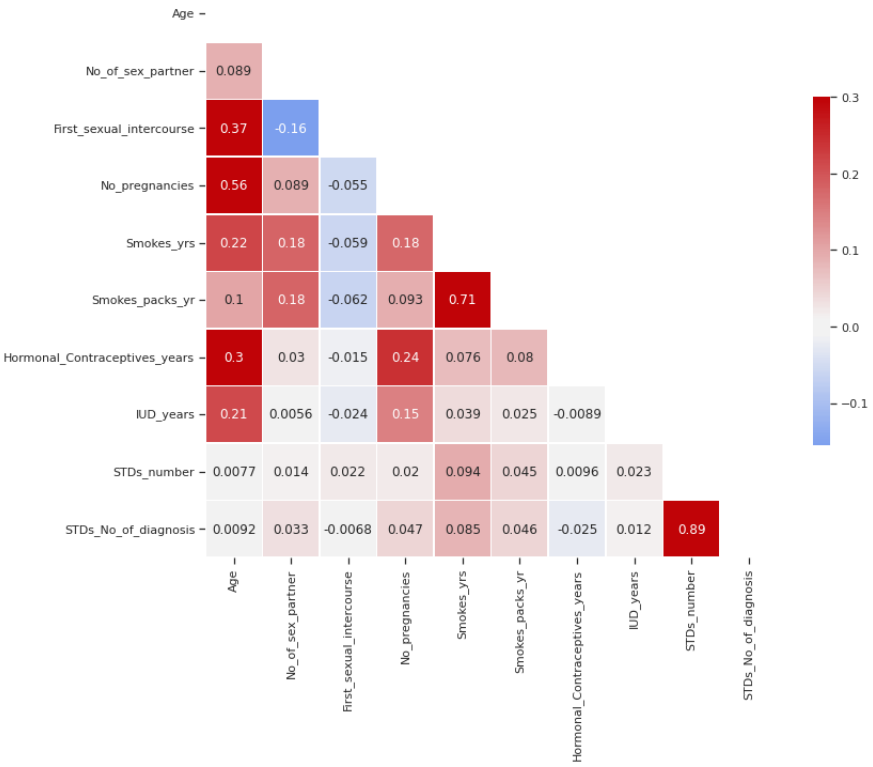

Figure 2: Correlation matrix for all variables of interest

To such extent, one can observe the highest pairwise linear correlation coefficients for the Age and Sexual Behaviours predictors (i.e. age of the first sexual intercourse, hormonal contraceptive or intrauterine copper device and the number of pregnancies). Furthermore, the latter is also positively correlated to smoking status, as one can expect. The rather short range in correlation coefficient's gradient binds the rest of the analysis to an additional degree of freedom: only pairs

of variables associated to a coefficients greater than 0.2 in absolute value will be considered to synthesise data with the Gibbs Algorithm.

The final set of features includes the variables:

- Age
- Age of the First Sexual Intercourse
- Number of Pregnancies
- Number of years spent smoking
- Number of cigarettes packs smoked per year
- Number of year spent using hormonal contraception
- Number of year spent wearing a copper IUD

We do not take into account sexually transmitted diseases since 95% of positive-cancer patients have been infected by the HPV (Human Papillomavirus), which is the only STD presents in the affected subsample of patients.

We are only allowed to synthesize data with the above descriptors, therefore theses covariates will be the ones considered in our future machine learning models.

Results from the sampling process are displayed below (figure 3). To easily compare the generated distributions to the original ones (for positive-cancer profiles), we sample 100 synthesized data to be compared with the 17 original data.

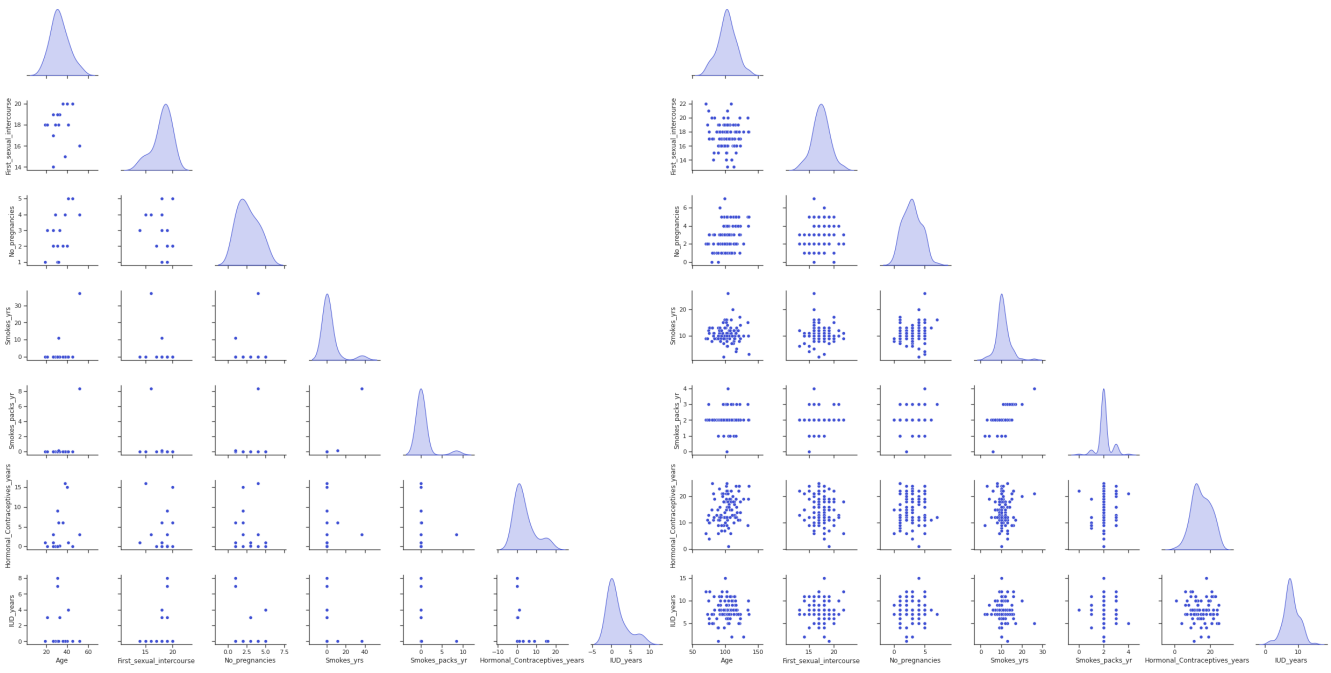

Figure 3: Real positive-cancer profiles (right) and synthesized positive cancer profiles (left)

4 Model Fitting

4.1 Model Evaluation Framework

According to the challenges cited hereinabove as to predict cancer-positive profiles, we need a metric that will take care of both, correct positive predictions and correctly predicted positives. As previously stated we therefore evaluate our models with the F1-score which focuses on positive predictions.

4.2 Predictions

The model fitting framework follows a random train-test split partitioning of the data prior to naive i.e. untuned model fitting. As a preventive over-fitting controlling measure, the training sample contains both observed and synthetic data whereas the testing sample only contains observed data. We compare our results to the baseline, that is to say, to predictions obtained on observed data only (unbalance dataset).

In order to built the bests models we use a recursive elimination feature method (from sklearn) to find the most relevant descriptors for prediction.

An exhaustive range of classifiers are trained[4], validated and tested for selection and the following results serve as examples for classifiers that lead to positive F1 score, namely: arborescent models for the original dataset, and a Perceptron, a LDA and a ridge classifier for the observed and synthesized data.

	observed	**observed + synthesized**
Extra Trees	age, first sexual intercourse F1 score 0.15	F1 score 0
Decision Tree	number of pregnancies, age, first sexual intercourse, years smoking, years with hormonal contraceptive F1 score 0.18	F1 score 0
LDA	F1 score 0	number of pregnancies, age, years smoking, years with IUD F1 score 0.20
Perceptron	F1 score 0	first sexual intercourse, number of pregnancies F1 score 0.12
Ridge Classifier	F1 score 0	number of pregnancies, years smoking, years with IUD F1 score 0.20

Table 2: Results for observed data and observed + synthesized data

4.3 Predictions using spectral transformations

Spurious associations might find in the data by our machine learning models. Therefore it is natural to wonder when and how one can hope to infer information about the potential unmeasured covariates from the observed data and whether it is possible to use that information to remove some of the confounding-induced bias in the estimates. Ćevid et al. [2018] show that under some

[4]linear discriminant analysis (LDA), extra trees classifier, decision tree classifier, perceptron, stochastic gradient descent classifier, ridge classifier, nearest centroid, support vector classifier

assumptions about the structure of the data and the confounders, spectral transformations allow to decrease the impact of confounders and to obtain better estimates of the studied effects, sometimes with increased predictive performance.

We describe the transformation used for our challenge hereafter. Consider the setting of a high-dimensional linear regression, and let the singular value decomposition of the design matrix $X \in \mathbb{R}^{n \times p}$ be $X = UDV^\top$, with $U \in \mathbb{R}^{n \times n}$ a matrix whose columns are the left singular vectors of X, $D \in \mathbb{R}^{n \times n}$ a diagonal matrix whose diagonal entries are the singular values of X, and $V \in \mathbb{R}^{n \times p}$. Then, the spectral transformations considered is the one proposed by Ćevid et al. [2018], which caps the singular values of X to their median while keeping the same matrices U and V. Since U is semi-unitary, this can be done by taking a transformation F with:

$$F = UQU^\top \quad \text{where} \quad Q = \begin{pmatrix} \tilde{d}_1/d_1 & 0 & \dots & 0 \\ 0 & \tilde{d}_2/d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{d}_n/d_n \end{pmatrix}$$

Here, $d_1, d_2, ..., d_n$ are the singular values of the design matrix X, i.e. the diagonal entries of D, and $\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_n$ are new singular values that are chosen such that $\forall i \in \{1, ..., n\} : \tilde{d}_i = \min(d_{\lfloor n/2 \rfloor}, d_i)$ so that the spectral transformation is capping the singular values to the median of the original spectrum.

Then, we can write the transformed design matrix $\tilde{X}$ as:

$$\tilde{X} = FX = UQU^\top UDV^\top = U \begin{pmatrix} \tilde{d}_1 & 0 & \dots & 0 \\ 0 & \tilde{d}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{d}_n \end{pmatrix} V^\top.$$

Then, with F defined as above, the resulting estimating procedure suggested is simply to estimate parameters by using a classifier on the transformed design matrix $\tilde{X} = FX$ and on the outcome (which is not transformed here since it is a dummy variable).

In this section, and as a final experiment, we build on the approach presented above and try to apply it to our classification problem.

To do so, we apply the transformation F matrix composed of both the training and testing sets as described above. For clarity, we display the profiles of the singular values of the design matrix before and after applying the spectral transformation in Figure 4 below.

4.4 Remarks and Limitations

Two points should be brought to the reader's attention regarding the classification of the data. Firstly, any algorithm is as accurate as the data it is fed with. And according to Occam's Razor Principle, underwhelming performances should first be attributed to improper feature engineering rather than poor machine modelling. To such an extent, the reader can notice that the baseline F1 score for observed data only stands at 0.15 and 0.18 for arborescent models (extra trees and decision tree respectively). Well balanced behavioral data better explains cervical cancer ; F1 score for both observed and synthesized data stands at 0.20 for LDA and ridge classifier.

While the spectral transformations allows a significant increase in the F1 score for observed data: the score increases to 0.33 for extra trees classifier, it leads worse results on the well balanced dataset (F1 score around 0.1 for LDA, perceptron and ridge classifier).

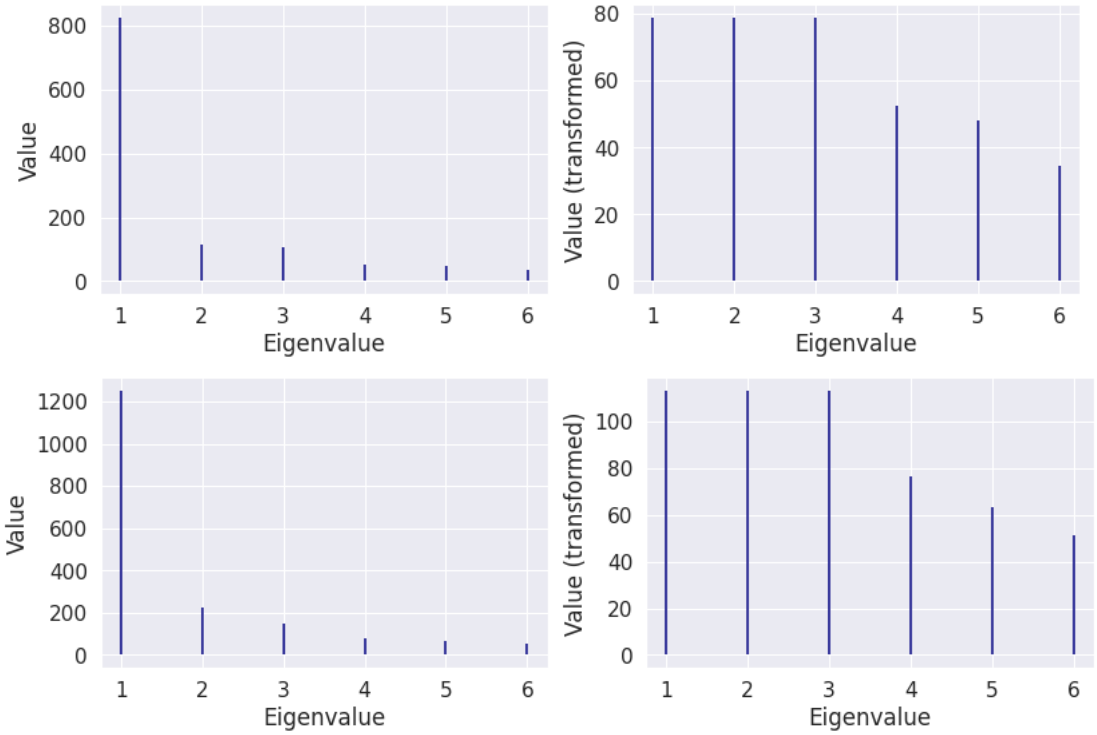

Figure 4: Eigen spectrum of observed features before SVD and after SVD (top) and of observed + synthesized features before SVD and after SVD (bottom)

	transformed observed	**transformed observed + synthesized**
Extra Trees	number of pregnancies, years with hormonal contraceptive, years with IUD F1 score 0.33	F1 score 0
Decision Tree	number of pregnancies, cigarette packs per year, years with IUD F1 score 0.18	F1 score 0
LDA	F1 score 0	number of pregnancies, age, years smoking, years with hormonal contraceptive, years with IUD F1 score 0.11
Perceptron	F1 score 0	number of pregnancies, age, first sexual intercourse, years smoking, years with hormonal contraceptive F1 score 0.10
Ridge Classifier	F1 score 0	number of pregnancies, age, years smoking, years with hormonal contraceptive, years with IUD F1 score 0.11

Table 3: Results for observed transformed data and observed + synthesized transformed data

Thus, the cervix prediction is best addressed with a highly random forest, also known as extra trees classifier, on observed data. In particular, the fact that extra trees perform slightly better than a decision tree (and way more than random forest not presented here because leading to null F1 score) indicates that:

- Particular focus should be put on over-fitting. It is not that the decision boundaries are so hard to be found that the model requires extensive depth i.e. a signal from a single Naive Decision Tree, but rather that the $F(X) = y$ relationship tying the features to the target is still latent to the model. Hence, multiplying the classifiers and averaging their results helps covering more of the domain of definition of the distribution F, and its associated loss function. In sum, the meta-model reduces the results' dependence on the sample that is drawn.

- The trees should not only be aggregated in a meta estimator but the best split found for all features and for all trees should be made randomly, before optimal selection amongst the features. This last point truly puts the emphasis on the exploratory nature of the learning process.

Hence, the evaluation framework encompasses all previous conclusions drawn so far about the quality of the data hypothesis put on the target determination, as well as the challenges it imposes on the search for optimal decision rules.

5 Conclusion

We investigated two main work tracks for this project in order to enhance predictions compared to the baseline obtained with a decision tree applied on the original dataset with the following selected variables (with a recursive feature elimination): number of pregnancies, cigarette packs per year, years with IUD. The first track studied, a Gibbs'algorithm to generate affected patients to encompass the unbalance in our original dataset. The process allows a slight increase in the F1 score compared to the baseline. Indeed LDA method and ridge classifier applied on observed and synthesized data produced predictions with a F1 score of 0.2 while arborescent methods on the original dataset produced F1 score of 0.15 and 0.18 (extra trees and decision tree respectively). The second track, using spectral transformations to minimise the confounding-induced bias in the estimate obtained from machine learning classifiers. The use of SVD decomposition of the features rather than their original form yields the extra trees classifier on original dataset as the best model. Using observed data only and considering the number of pregnancies, the years with hormonal contraceptive and the years with an IUD is thus sufficient to get a F1 score of 0.33, which is way more than other models we built for this project. However applying the spectral transformation on observed and synthesized data leads to worse predictions compared to the baseline.

Additional feature engineering and in particular, singular value decomposition highlights a complex graph representation of the feature map, with cross-determination and co-dependencies. Indeed, it seems intuitive to interpret age, the number of sexual partners, the number of pregnancies, and the contraceptive choices as the result of a common socio-ecological and anthropogenic process. Likewise, the smoking status and the contraceptive choice might also be inter-twinned within the same framework. Reflecting those frameworks as an holistic representation of the data, rather than the data itself helps better extracting the information encompassed and sensibly predict the outcome.

Finally, and to emphasize on the compounded effect of the Gibbs' Sampling Algorithm prior to Singular Value Decomposition for maximized information retrieval, the reader is invited to compare the difference in magnitude of the eigenvalues on observed and on hybrid data respectively (see figure 4).

6 References

References

Brücher, B. L. and Jamall, I. S. (2014). Epistemology of the origin of cancer: a new paradigm. *BMC cancer*, 14(1):1–15.

Castellsagué, X. (2008). Natural history and epidemiology of hpv infection and cervical cancer. *Gynecologic Oncology*, 110(3, Supplement 2):S4–S7. I International Symposium on Cervical Cancer:.

Ćevid, D., Bühlmann, P., and Meinshausen, N. (2018). Spectral deconfounding via perturbed sparse linear models. *arXiv preprint arXiv:1811.05352*.

Franco, E. L., Duarte-Franco, E., and Ferenczy, A. (2001). Cervical cancer: epidemiology, prevention and the role of human papillomavirus infection. *CMAJ*, 164(7):1017–1025.

Gillison, M. L., Castellsagué, X., Chaturvedi, A., Goodman, M. T., Snijders, P., Tommasino, M., Arbyn, M., and Franceschi, S. (2014). Eurogin roadmap: Comparative epidemiology of hpv infection and associated cancers of the head and neck and cervix. *International Journal of Cancer*, 134(3):497–507.

Juneja, A., Sehgal, A., Mitra, A., and Pandey, A. (2003). A survey on risk factors associated with cervical cancer. *Indian Journal of Cancer*, 40(1):15–22.

Vinh-Hung, V., Bourgain, C., Vlastos, G., Cserni, G., De Ridder, M., Storme, G., and Vlastos, A. T. (2007). Prognostic value of histopathology and trends in cervical cancer: a seer population study.

Åhren, A.-M., Johansson, R., Bergman, F., Wadell, G., Ångström, T., and Dillner, J. (2000). Smoking, diet, pregnancy and oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection. *British Journal of Cancer*.