Cervical Cancer Risk Factors

Kristie Kooken

August 12, 2023

Introduction

Cervical cancer is a type of cancer that starts in the cervix. The cervix connects the vagina (birth canal) to the upper part of the uterus (womb). Anyone with a cervix is at risk for cervical cancer (www.mayoclinic.org). In the United States, there are about 11,500 cases diagnosed, with 4,000 women dying from this type of cancer each year. Stated another way, for every 100,000 women, seven new Cervical cancer cases are reported, and two women will die of this cancer (www.cdc.gov). Two characteristics of this cancer are that it most often occurs in women over 30 years old and is often associated with a long-lasting infection with certain types of human papillomavirus (HPV). By screening and HPV (human papillomavirus) vaccination, a whopping 93% of cervical cancer cases could be prevented (www.cdc.gov).

The goal of this project is to develop a model to predict cervical cancer and human papillomavirus (HPV) using a dataset which has health information and screening diagnositic tests. Because the outcome is a binary category, a classification model will be used. Two logistic regression models will be developed for each target of cervical cancer and HPV. Results of this model could be used to develop a groundbreaking intervention that would reduce the burden of this cancer. This intervention would focus on statistically meaningful characteristics of predictors of cervical cancer. This is important because lives can be saved by an affordable intervention which would deliver the best medical practices without the burden of treatment for cervical cancer and allow the patient to better manage their health and wellbeing. Such a low-cost intervention would be highly desirable to all types of health insurance carriers as it can lower costs associated with cancer care and treatment for these companies. Other modeling techniques that may be explored are tree-based methods (random forest) which can be a suitable alternative for assessing risk factors (www.datasciencecentral.com).

Data Source

The data supporting this model development was downloaded from the UC Irvine Machine Learning Repository and has cervical cancer risk factors (Fernandes, K., et. al., 2017). The data focuses on the prediction of indicators/diagnosis of cervical cancer. Features on this dataset include demographics, habits, and historical medical records. This data, collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela, has 36 columns and 858 rows of data. In the initial dataset, several patients decided not to answer some questions due to privacy concerns; thus, there are missing values for these patients. The dataset is useful for this project because of robust demographics, health history, and habits around sexuality. Because of the strong relationship between cervical cancer and HPV, it is critical to understand the individual's sexuality and sexual practices. Without data of this nature, researchers are left in the dark without understanding the social component of the causes of this cancer and being unable to access information on habits that could put an individual at risk.

Methods & Modeling

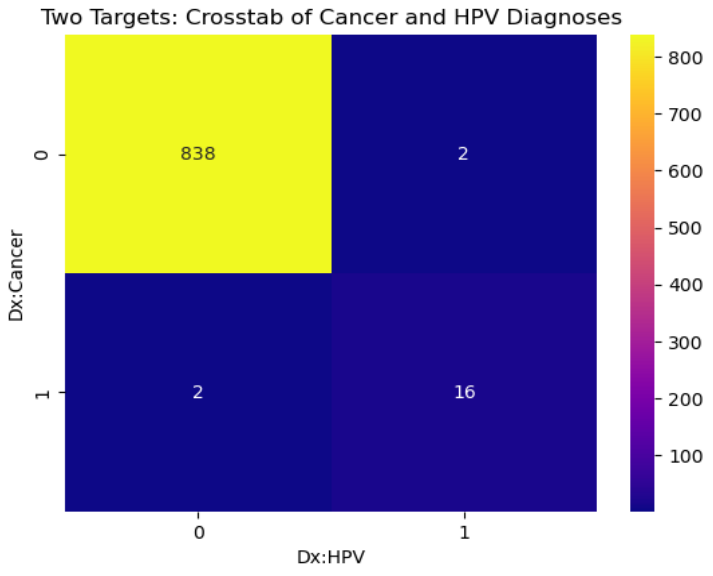
The data was scrubbed to replace missing values with individual column medians for continuous variables and individual column modes for binary categorical variables. Histograms for continuous variables and bar charts for binary categorical variables were generated and inspected to examine the distribution and identify outliers. No outliers were found for any variable. It was noted that all continuous variables were skewed to the left (e.g., Age in years); since this was not unexpected, no action was taken.

Descriptive statistics were generated on all values to inspect ranges as well as to confirm missing data was imputed as described.

Spearman's correlation matrix was run on the continuous variables. Spearman's correlation was used for this as all continuous variables are skewed to the left and Spearman's is a nonparametric alternative to Pearson's correlation (www.statisticsbyjim.com). Results from this analysis showed the continuous STD variables ('STDs (number)', 'STDs: Number of diagnosis', 'STDs: Time since first diagnosis', 'STDs: Time since last diagnosis') are all highly correlated with each other (for 'STDs (number)' and each respective variable: r=0.94, r=0.94 & r = 0.94). Based on this finding, only 'STDs (number)' was kept. Likewise, the pair of variables, 'Smokes (years)', 'Smokes (packs/year)', had a very strong correlation (r=1.0) and 'Smokes (packs/year)' was dropped from the dataset.

A total of six variables were dropped from the dataset: four variables based on the correlation analysis and two variables that lacked documentation.

The relationship between the target variable of each proposed model was explored to assess the feasibility of having a single statistical model instead of two separate models. A crosstab heatmap was generated for Dx:Cancer and Dx:HPV showing the close relationship between these two variables (shown below).



In order to test the strength between these two variables, a chi-square statistic was run. Chi-square was used because it measures the degree of association between two categorical variables. The result of the chi-square test was significant ($p<.0001$) indicating a lack of independence between the two variables. This finding confirms the best choice for modeling is to have two models (one for each target) instead of developing a single model to predict cervical cancer and using a diagnosis of HPV as a feature. Likewise, using both variables in a single model with cervical cancer as the outcome and HPV as a feature could also lead to data leakage since the two variable have nearly identical distributions and values.

For each target variable, cervical cancer and human papillomavirus (HPV), there are very few positive diagnoses; thus, the data is imbalanced. Because of this imbalance, all data was kept to prevent removing positive cancer diagnoses.

Because many machine learning technique use distance-based calculations to understand patterns in the data, scaling was performed on the continuous variables. The continuous variables have different ranges and because of this, the MinMax method was used to normalize the data as it preserves the original distribution of the variables and works well for variables with different ranges. The binary categorical variables are coded as 1/0 and thus did not require additional processing.

Initial baseline modeling was performed on the scaled data for two reasons – 1) to understand model performance before any feature selection technique and 2) to develop a dummy classification model to gauge if the imbalance of the target variable was influencing model accuracy. For the dummy classifier, the option of “stratified” was used to ensure the model was not biased towards the majority class.

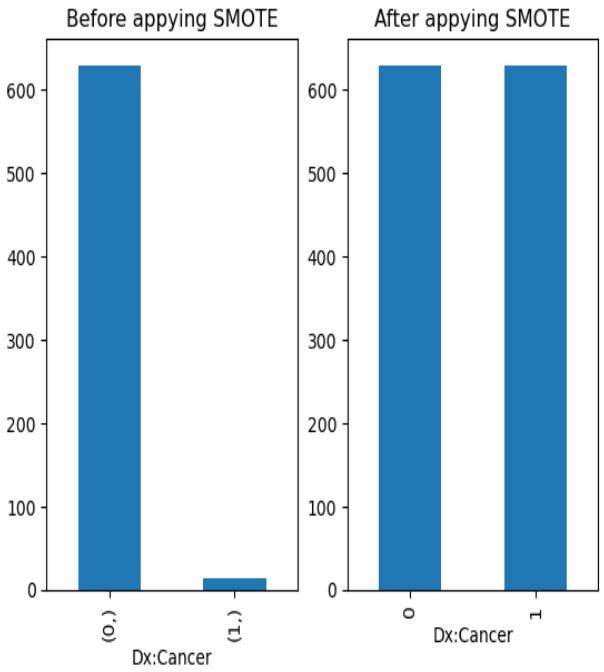
In order to reduce features and to address any potential for overfitting, a LASSO regression was conducted. LASSO stands for Least Absolute Shrinkage and Selection

Operator and is a type of regression where data is ‘shrunk’ towards a central point such as the mean. LASSO identifies a feature where the coefficient can be reduced to zero and then this feature can be removed from subsequent analyses (www.blog.trainindata.com). This technique can reduce variance if there are many insignificant features. For this data, seven features remained after 21 were removed. These are the remaining features: ‘Hormonal Contraceptives’, ‘IUD’, ‘STDs:HPV’, ‘Schiller’, ‘Biopsy’, ‘Smokes’ and ‘STDs:condylomatosis’. For this analysis, interactions between features were not explored to determine if any were significant. However, further exploration is needed in this area since significant interactions can imply a more complex model than a linear model. If there are significant interactions, it would be beneficial to consider using tree-based methods such as random forest or gradient boosting as these methods are more capable to handling interactions (www.geeksforgeeks.org).

After employing feature selection, the data was reassessed for baseline modeling. The same option of ‘stratified’ was used for the dummy classifier to ensure the model is not biased towards the majority class.

Because model performance did not improve after feature selection, and imbalance in the minority class was likely responsible for the poor performance, SMOTE was used to balance the data. This technique was chosen over weighted logistic regression because the event of cervical cancer or HPV is rarer and SMOTE is a better technique for severe imbalance (www.towardsdatascience.com). A drawback of SMOTE is that it can lead to overfitting. In order to address this, SMOTE was only performed on the training data in order to prevent data leakage as well as potentially address overfitting. Future work for this model would include reviewing model performance between SMOTE train data and testing data to ensure they have a comparable improvement from the initial training data. If the performance

metrics between the two are not comparable, it could be a sign of overfitting. The histograms of before and after applying SMOTE on Dx:Cancer are shown below.



After creating balanced data, the final logistic regression model for predicting cervical cancer was run. For this model, hyperparameter tuning for logistic regression was performed in order to select the best hyperparameter to achieve the optimal performance on the test dataset. The settings for the tuning were chosen based on the following: 1) different solvers control differences in performance or convergence and are used to optimize the model's parameters during the training process, all solvers listed can work with L2, 2) the penalty of L2 (ridge regression) is selected, this is the default setting, and 3) different values for c_value were explored to assess the best choice for regularization strength where lower c values mean stronger regulation and larger c values mean weaker regulation.

After reviewing the results of the hyperparameter tuning, the ‘best’ hyperparameters are used for the final logistic regression model for the target Dx:Cancer.

The below table combines presents the classification report results for each of the three sets of models run:

Model	Accuracy	Class	Precision	Recall	F1	Support
Dummy Classifier	0.97	0	0.98	0.99	0.98	210
	-	1	0.00	0.00	0.00	5
Baseline Logistic Regression	0.98	0	0.98	1.00	0.99	210
	-	1	0.00	0.00	0.00	5
Dummy Classifier*	0.97	0	0.98	0.99	0.98	210
	-	1	0.00	0.00	0.00	5
Baseline Logistic Regression*	0.98	0	0.98	1.00	0.99	210
	-	1	0.00	0.00	0.00	5
Final Logistic Regression*^	0.87	0	0.98	0.88	0.93	210
		1	0.07	0.40	0.12	5

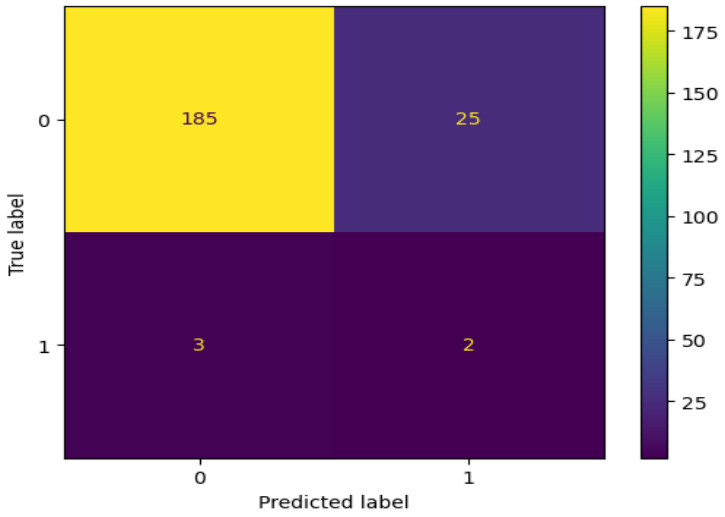
* With Feature Selection using LASSO
^ SMOTE applied to target and model tuning

Results from this table show that, prior to addressing the imbalance, all initial models were unable to correctly predict any instances of cervical cancer. Having zero values for Precision, Recall and F1means that no correct predictions were made. These zero values indicate there is an issue with imbalance or features may not have enough information to distinguish instances correctly or an issue with data quality (www.towardsdatascience.com). For this project, these poor performance indicators are presumably due to the severe imbalance of the target.

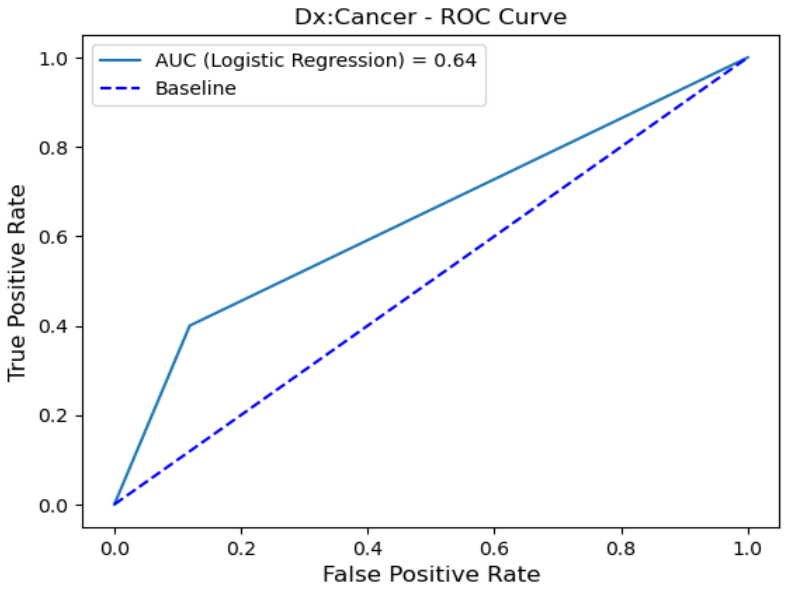
Accuracy for any health model is extremely critical. To check the performance of each model, accuracy scores and recall scores are presented and for the final model, accuracy scores, confusion matrixes, recall scores, and ROC plots are presented.

The last steps for the final model were to create a confusion matrix plot as well as ROC plot with the ROC area under the curve (AUC) score displayed. The confusion matrix is a visualization of the true negatives, false positives, false negatives, and true positives which are the correct and incorrect number of predictions for each class. The ROC plot was picked as a visualization because it shows how well the logistic regression model discriminates between the positive and negative classes at different classification thresholds. The AUC

score represents the area under the ROC curve; higher AUC means the model was better at correctly classifying instances. The ROC plot also includes a 0.5 classification which is the decision threshold (meaning it shows 50% positive and 50% negative which is about the same as random guessing and indicates that the classification model is not discriminating between the two classes) (www.towardsdatascience.com). The confusion matrix and ROC plots are displayed below.



The confusion matrix shows 185 as true negatives, 25 false positives, 3 false negative and 2 true positives.



The ROC plot shows a line above the 0.5 threshold, however the AUC score of 0.64 indicates a moderate to low ability to discriminate between the positive and negative classes. Likewise, the ROC plot shows that there can be room for improvement in the model's performance based on the its closeness to the 0.5 threshold.

After running this final model for the prediction of cervical cancer, the same sequence of analyses was run for the second target, Dx:HPV. It is important to note that all the results from this second modeling effort are the same as for Dx:Cancer. This is because the targets match completely for 89% of the subjects (16 subjects out of 18 total for each target are on the same row of data). The remaining 4 subjects (2 in Dx:Cancer and 2 in Dx:HPV) have different ages but have the same responses on other characteristics. Likewise, these subjects did not have any data imputed on any variable except for those that were dropped after the correlation analysis during data cleaning. Because of the reasons explained above, the details and results of those analyses run for the second target, Dx:HPV, are not included.

Results

Model results from this analysis indicate that the initial baseline logistic regression model had a higher accuracy score (98%) than the dummy classification model (97%). Likewise, after feature selection, these accuracy results remained the same for each model. Upon further investigation, the confusion matrix results for the baseline models showed 210 as true negatives, 0 as false positives, 5 as false negatives and 0 as true positives. These results demonstrate that the imbalanced data does not produce an informative model regardless of how high the accuracy is. Having a 0 value for true positives for cancer diagnosis is not accurate and highly misleading.

After using the oversampling method to solve the imbalance in the target, the final logistic regression model was rerun with tuning. The accuracy score from this model was

87%, a lower value than seen with the initial model. Though the accuracy score is lower, after investigation of the confusion matrix results, the balanced model is much more accurate in terms of what we already know. We know there are cancer diagnoses, and that the positivity rate is not 0 thus we know that the imbalance in the data drives the high accuracy. Since a cancer diagnosis is a very serious life changing event, the model needs to accurately identify true positives, false positives, and false negatives cases. Below is the comparison between the final model to the model with imbalance:

- True negatives: 185 (initial baseline regression was 210 so this has reduced with balanced data)
- False positives: 25 (initial baseline regression was 0 so this has increased with balanced data)
- False negatives: 3 (initial baseline regression was 5 so this has reduced with balanced data)
- True positives: 2 (initial baseline regression was 0 so this has increased with balanced data)

Likewise, comparing the results for recall show the following:

- Recall of no cancer diagnosis went down from 1.0 to 0.88: meaning there are more positive cancer diagnoses in the final model that the initial model did not identify.
- Recall of yes cancer diagnosis went up from 0.0 to 0.4: meaning the final model identified more positive cancer diagnoses.
- Precision for the final model also increased, though more modestly, 0.07 for the final model compared to 0.00 in the imbalanced model.

Conclusions

Being able to accurately predict a cancer diagnosis would be a great advantage for all humanity. In order to do this effectively, any model will need to have a high degree of accuracy, balanced data as well as provide robust information on true positives, false negatives, and false positives. The final model for this project was an improvement over the imbalanced model. Though the final model had an acceptable accuracy (87%), the accuracy is not high enough to predict cervical cancer in a real-world setting and to become the basis for developing an intervention. However, the information gained from this effort does demonstrate that further modeling is needed, such as classification tree methods, to determine if a better model could be developed with the current data.

For this project, the relationship between the diagnosis of cervical cancer and the diagnosis of HPV is very strong; thus, presenting results from two models would be redundant. However, the relationship between cervical cancer and HPV needs to be explored to determine the differentiating characteristics between the two. Because of the strength of this relationship, further modeling efforts could include having multiple targets (e.g., multi-target regression) or using the predictions resulting from one model as input features for the second model – stacking machine learning (www.machinelearningmastery.com).

A number of ethical concerns were noted for this project. While running these analyses, many of the initial features were not in the final model. Due to the personal nature of HPV and cervical cancer, a physician (or questionnaire) asking questions about the subject's sexual history is expected. However, after modeling this data, very few of these items (features) were in the final model. Ensuring that researchers are collecting relevant data is an essential ethical consideration. For example, many factors about sexual history for this data had missing responses because subjects had privacy concerns. However, the one question that is

known to be associated with HPV was not a collected item (people contract HPV by having unprotected sex (www.cdc.gov), specifically not using a condom). It would be more ethical and cost-effective to collect items that show a relationship to HPV and determine if more items/questions are needed for a cancer diagnosis. This way, the two diagnoses could be measured more independently, and this could potentially help to develop a more robust combined model.

Another risk is that results presented are inconclusive as these analyses have a limited scope for the planned models versus a dedicated effort to predicting cervical cancer or HPV. There could be unforeseen issues or limitations with the data such that even robust results would need to be part of a larger modeling framework to determine outcomes and key factors of cervical cancer and HPV. Likewise, further work for this effort will need to have a subject matter expert for cervical cancer and HPV in order to ensure the best quality data sources and appropriate medical interpretation.

Lastly, though there are few demographics in this dataset, it is especially important that the subjects who participated do not have their privacy violated. In the United States, hospital location would likely not be known according to HIPAA Privacy Rules. This provides federal protections for personal health information held by covered entities and gives patients an array of rights with respect to that information (www.hhs.gov). Data privacy is extremely critical and will only continue to be more so in the future. Without trust or the ability to be anonymous, people will not give important health information that could be critical to solving different health concerns, diseases, and cancers. This would be very unfortunate; medical providers as well as data science professionals need to ensure privacy is always kept.

References

Alternative to Logistic Regression (Accessed 2023, August 1). Data Science Central. https://www.datasciencecentral.com/alternatives-to-logistic-regression-2/

Beyond Accuracy: Precision and Recall (Accessed 2023, August 9). Towards Data Science. https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Cervical Cancer is Preventable (Accessed 2023, June18). Centers for Disease Control and Prevention. https://www.cdc.gov/vitalsigns/cervical-cancer/index.html

Cancer Statistics At a Glance (Accessed 2023, June18). Centers for Disease Control and Prevention. https://gis.cdc.gov/Cancer/USCS/#/AtAGlance/

Cervical cancer (Accessed 2023, June 18). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501

Feature selection with Lasso in Python (Accessed 2023, August 1). Training in Data. https://www.blog.trainindata.com/lasso-feature-selection-with-python/

Fernandes, Kelwin, Cardoso, Jaime, and Fernandes, Jessica. (2017). *Cervical cancer (Risk Factors)*. UCI Machine Learning Repository. https://doi.org/10.24432/C5Z310.

Glen, S. (2019). *Alternatives to Logistic Regression.* Data Science Central. https://www.datasciencecentral.com/alternatives-to-logistic-regression/#:~:text=For%20example%2C%20tree-based%20methods%20are%20a%20good%20alternative,work%20well%20for%20propensity%20score%20estimation%20and%20Categorization%2FClassification.

Genital HPV Infection – Basic Fact Sheet (Accessed 2023, August 1). Centers for Disease Control and Prevention. https://www.cdc.gov/std/hpv/stdfact-hpv.htm

Hyperparameters for Classification Machine Learning (Accessed 2023, August 1). Machine Learning Mastery. https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/

ML | Logistic Regression v/s Decision Tree Classification (Accessed 2023, August 10). Geeks for Geeks. https://www.geeksforgeeks.org/ml-logistic-regression-v-s-decision-tree-classification/

SMOTE (Accessed 2023, August 1). Towards Data Science. https://towardsdatascience.com/smote-fdce2f605729

Spearman's Correlation Explained (Accessed 2023, August 1). https://statisticsbyjim.com/basics/spearmans-correlation/

Stacking Ensemble Machine Learning With Python (Accessed 2023, August 1). Machine Learning Mastery. https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/