Name: Salifyanji J Namwila
Course: Math 050
Paper: Final Project.
Date: 11/19/2021

PROJECT TITLE
The Bayesian Perspective on Breast Cancer Diagnosis in Women.

Section I.
INTRODUCTION

Breast Cancer is the most common female cancer worldwide. All most 25% of all cancers with an estimated 1.67 million new cancer cases diagnosed in 2012. Breast cancer is the most common type of cancer among women and its incidence is increasing day by days. The lifetime risk of developing breast cancer in women is approximately 1/8 in USA, 1/ 12 in Europe, 1/40 in Asia (WHO 2008). As per the cancerindia.org, one woman dies of breast cancer every 8 minutes in India. Breast cancer is the most common cancer in women and account for 27% of all cancers. As per the previous study 1,44,937 new cases registered and deaths 70,218 in 2012. According to the World Health Organization breast cancer was responsible for 502,000 deaths in 2005 alone and 1,301,867 of new cases of breast cancer resisted. Health case authorities continuously doing efforts to overcome this merciless disease in which one of the efforts is screening. By screening the breast cancer can be detected in early stages and thus the treatment can be more effective. Many other methods also available such as mammography, ultrasound, CT, and MRI. Mammography is the most widely used screening method. However, even though all these methods present engineering success because they do not necessarily present us with insights into how representative the underlying models are of the data they train on, readily inform us of how much variation is explained by their underlying models, medical practitioners are prone to delivering flawed results ignorantly. This lack of insight about representativeness of the model, lack of knowledge about interactions between predictor parameters, and hence lack assurance of the accuracy of the models deems these various screening methods as not necessarily Scientific success, wherein scientific success is a step closer at sourcing predictor relations. At the cradle of better cancer health should lie constant insight into why results are so, not simply through memorization of learning patterns. Enter my project which, even though does not explain why the combination of certain predictions in cancer diagnosis results in positive cancer diagnosis, still gets a step closer at enlightening us about interactions between important parameters in the model. This is a significant step opening doors to the right kind of research in understanding the development of cancer.

Breast cancer is one of the major life-threatening diseases in the world today. But early diagnosis would ensure timely treatment and even recovery. Apart from medical procedures that help to identify cancer and the skit -learn library, techniques such as Bayesian inferences and stats modelling of general linear models can also be applied to predict the chance that a patient has breast cancer by analyzing their cell data.

In my proposed method of predicting breast cancer, I use a dataset that provided through Biopsy Images generated by Mammography method. I employ Bayesian Inferences and stats modelling techniques on this breast cancer data to predict the likelihood that a patient has the disease by analyzing their cell data. The point of my project is not to necessarily substitute existing prediction method but to supplement them in a way that not only delivers diagnosis but also triggers a deeper understanding of trends of important features used in the predictions of cancer and realistically presents the accuracy of the model to users. Bayesian Inferences from the PYMC3 posterior distributions of generalized linear regression models and stats models can be used together achieve these goals precisely. Bayesian Inferences can be used to visualize how various parameters correlate with each other. For instance, how does the texture of a tumor change as the fractal dimension of the tumor increases? How is the smoothness of the tumor likely to change as its texture increases? If the tumor is so smooth, how does that affect the chance of the patient being positive? Are there parameters that carry more importance than others? What would happen to the accuracy of the model if we removed information about Area, Radius, Concavity, Compactness and Perimeter of the tumor? Would the prediction be so different from the actual diagnosis? In this work, an attempt is made to answer all these questions and more derivatives by applying Bayesian Inference and informative statsmodels to a data set on which a Logistic regression from the Scikit-Learn Library, a variation of linear models, was applied by Kaiser Moo, an experienced data enthusiast in data science. The performance metrics of the stats model are analyzed against those of Kaiser's logistic regression model and disparities explained.

The rest of the paper is organized as follows: Section 2 presents the related work that covers Kaiser Moo's research work. Section 3 includes the architecture of the proposed work. Section 4 describes the pre-processing of data. Section 5 describes the methodology which is used for the proposed work. Section 6 discusses the results and Section 7 concludes the proposed work and Highlights Future work.

Section II
RELATED WORK

Kaiser Moo uses one of the common libraries for training machine learning model, the open source sklearn library to analyze ultrasound data of breast cancer tumors. Several characteristics of the tumor are given in the dataset, and the scientist creates a logistic regression model that predicts whether a patient has a positive breast cancer diagnosis based on the tumor characteristics (Kaiser M. 2020). Her data set has the following characteristics:

1. ID (patient ID)
2. Name
3. Radius (the distance from the center to the coreference of the tumor)
4. Texture (standard deviation of gray-scale values)
5. Perimeter (circumference of the tumor, approx.
6. Area
7. Smoothness (local variation in radius lengths)
8. Compactness
9. Concavity (severity of concave portions of the contour)
10. Symmetry
11. Fractal dimension
12. Age

Kaiser performs two experiments distinguished by the number of parameters used. In the first experiment, she performs exploratory data analysis (EDA), applies mean imputation for missing values, builds a classification model (logistic regression), and interprets the results using sklearn's classification report and confusion matrix. In the second experiment, she uses the same dataset, however, she computes the variance inflation factor (VIF) to locate and drop the highly correlated variables, re-runs the logistic regression model then compares the performance of both logistic regression models.

Upon evaluating the performance of her model in the first experiment using the testing set of size 171 women, Kaiser finds that out of the 105 women predicted to not have breast cancer, 7 were misclassified, making the model 93% accurate. And out of the 66 women predicted to have breast cancer, 10 were misclassified, making the model 85% accurate. The second model in, on the other hand is less accurate when tested on testing set of size 150. Out of 91 women predicted to not have breast cancer, 14 were incorrectly classified, making the model only 84.6% accurate. And out of the 59 women predicted to have breast cancer, 36 were incorrectly classified, making the model only 38.98% accurate.

Section III
LOGICAL FLOW THROUGH PROPOSED WORK

In this paper I use the same data as Kaiser, which Wisconsin Breast Cancer Dataset that is publicly available for download and use. I set up three different experiments, in each experiment pre-processing and presenting it in a canonical way by trimming undesired features to suite that experiment's goal.

In the first experiment, whose goal is analogous to Kaiser's first experiment, in the preprocessing, I drop non-numerical values, apply mean imputation for missing values, standardize all the continuous predictor variables to allow for a scaled interpretation of results, build a statsmodels classification model (logistic regression), and I finally interpret the results using `evaluate` function that I define and implement from scratch in my background Google Collaboratory Notebook to report the accuracy and confusion matrix. (Present accuracy comparisons here).

In the second experiment, whose goal too corresponds to Moo's second experiment which tries to predict breast cancer with only non- multi colinear predictors, that is, independent variables that are not highly correlated, I compute the variance inflation factor (VIF) to locate and drop the highly correlated variables, re-run the logistic regression model with only the net predictors left, and finally compare the performance of the resultant logistic regression model to the one in the first experiment and to the one in Kaiser's second experiment. (Present accuracy comparisons here).

Finally in the third experiment, whose goal is to develop a slightly more complex model by adding interaction terms of the most important predictors, I add interaction terms columns to the data in the second experiment and re-run the logistic regression model with only the new interaction terms included, and finally compare the performance of the resultant logistic regression model to the one in the second experiment. (Present accuracy comparisons here).

Section IV
EXPERIMENT AND METHODOLOGY

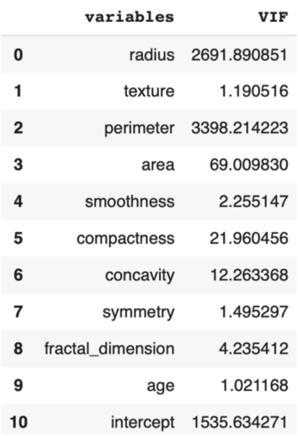
At the core of my project is to find out the relative accuracies of statsmodels Logit model and sklearn's Logistic regression model. The first experiment is specifically designed to answer this question. The second experiment is designed to answer my question about discarding predictors that make the model overly complex., while the third experiment aims to reveal the influence of interactions between the most important predictors in the model.

Section V.
PRE-PROCESSING DATA

Data pre-processing is a data mining technique that used for filter data in a usable format. Because the real-world dataset is available in a variety of formats, it's not available as per my project's requirements, hence the necessity to fitter the dataset into a canonical (understandable, and workable) format. Data pre-processing is a proven method of resolving such format disparity issues. In my project I used VIF and standardization methods to pre-process the dataset.

A. Variance Inflation Factor (VIF).

The variance inflation factor is a measure for the increase of the variance of the parameter estimates if an additional variable, given by exog_idx is added to the linear regression (Statsmodels).. It is a measure for multicollinearity of the design matrix, exog (Hayes A. 2021). One recommendation is that if VIF is greater than 5, then the explanatory variable given by exog_idx is highly collinear with the other explanatory variables, and the parameter estimates will have large standard errors because of this. (Statsmodels).



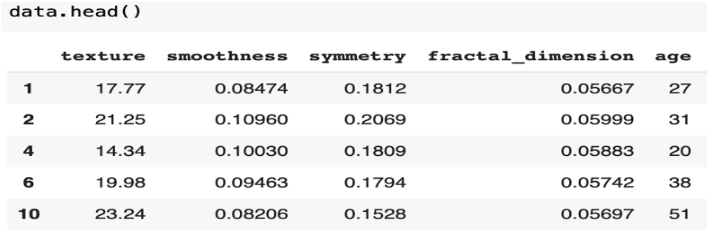
	variables	VIF
0	radius	2691.890851
1	texture	1.190516
2	perimeter	3398.214223
3	area	69.009830
4	smoothness	2.255147
5	compactness	21.960456
6	concavity	12.263368
7	symmetry	1.495297
8	fractal_dimension	4.235412
9	age	1.021168
10	intercept	1535.634271

Figure 1: Variance Inflation Factor of Predictor Variables.

After computing the VIF of predictor variables and removing all with VIF above 5, only 5 unique predictors are left.

B. Standardization.

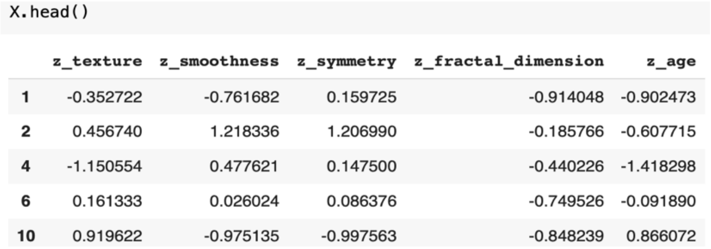
Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1.



```
data.head()
```

	texture	smoothness	symmetry	fractal_dimension	age
1	17.77	0.08474	0.1812	0.05667	27
2	21.25	0.10960	0.2069	0.05999	31
4	14.34	0.10030	0.1809	0.05883	20
6	19.98	0.09463	0.1794	0.05742	38
10	23.24	0.08206	0.1528	0.05697	51

Figure 2: Head of Dataset before standardizing continuous predictor variables.



```
X.head()
```

	z_texture	z_smoothness	z_symmetry	z_fractal_dimension	z_age
1	-0.352722	-0.761682	0.159725	-0.914048	-0.902473
2	0.456740	1.218336	1.206990	-0.185766	-0.607715
4	-1.150554	0.477621	0.147500	-0.440226	-1.418298
6	0.161333	0.026024	0.086376	-0.749526	-0.091890
10	0.919622	-0.975135	-0.997563	-0.848239	0.866072

Figure 3: Head of Dataset after standardizing continuous predictor variables.

Section VI
EXPERIMENT AND METHODOLOGY

When the model from the first experiment is trained and tested using the unseen testing set, results show that the statsmodels logistic regression is 97% accurate in predicting positive breast cancer, with precision and recall scores of 100% and 90.0% respectively. This is sufficiently above Moo's results in an analogous experiment wherein the model was 93% accurate, with precision and recall scores of 90.7% and 90% respectively. The results of this experiment are interesting because they indicate that statsmodels logit models are not necessarily inferior to sklearn's models. They could even make more accurate predictions as well as have better precision and recall scores. But what makes statsmodels logit regression model particularly invaluable is not only its (maybe) occasional better performance or equivalence to other phenomenal libraries like sklearn, but also its power to directly inform users of the goodness of fit of the model through the interpretation of Pseudo R^2. For example, the model has a pseudo-R^2 value of 0.7947, which indicates excellency of the model fit. This "goodness of fit" is also reflected by the comparison of the Log Likelihood (LL) of the estimated model and the Log Likelihood of the null model, the model containing no parameters but the intercept (Giselmar A et al, 2016). We see that there is an improvement from the null model with a LL of -266.94 to LL value to a LL value of -54.802 of the estimated model. Furthermore, the small P-value of 6.359e-85 only reinforces the statistical significance of the model.

For our purposes this is a great model already. However, it has nine predictor variables. A good model is one that captures the general trend in the training data while minimizing the number of parameters to reduce overfitting. The second experiment aims at reducing the optimizing the model by truncating features that are highly correlated and do not affect the model's accuracy greatly (Habshah M et al., 2013)

When the highly correlated features are discarded from the model, thereby leaving only five independent predictors, and the model is fit to a logistic regression model, results are quite astounding; the pseudo-R^2 of the model plunges from 0.7947 to only 0.4033, which reasonable because the Log Likelihood of the estimated model then only improves by just 91 from -225.43 LL of the null model. Even more, the accuracy of the model drops to 81.3%, with new precision and recall scores of 81.8% and 64.28% respectively. Even though the pseudo-R^2 value is 0.4033, which is a good fit as per standard convention (Giselmar A et al, 2016), for a task as critical as diagnosing breast cancer, the risk presented by the poor recall score is way too high. And so, an attempt is made to make the model more complex by adding predictors with a low Variance Inflation Factor. To achieve this, the most important predictors are identified using the Logit "divide-by-four rule" of betas. This step gives the upper bound of the predictive range between tumors that differ only in that beta's predictor if all other tumor characteristics remain the same. Upon calculating all the values, the characteristics smoothness, fractal dimension and texture have the highest upper bound predictive ranges of approximately 47.59%, -38.51%, and 33.98% respectively, implying that these tumor characteristics influence the model outcomes the most. Said differently, the likelihood of positive cancer diagnosis between patients that have tumors that differ only smoothness has the highest upper bound predictive range, followed by patients with tumors that differ only in fractal dimensions, followed by patients with tumors that differ only in texture. Symmetry and age, on the other hand, have the least upper bound predictive ranges so, it would not be so useful to evaluate interactions involving them.

With important predictors identified, three new interaction terms `z_smooth*z_text`, `z_smooth*z_fract_dim` and `z_text*fract_dim` are added to the model after ensuring that their VIF values are well below the recommended value of 5. In the process of identifying important predictors, I discover that age had the least upper bond predictive range of only -4.35%.

As per good practice, analyses of the now eight predictors are made using Bayesian Inferences with select priors. The purpose of this step is to understand how predictors associate with one another. Captivating findings are made. Plots of scatterplots of each of the predictor's beta coefficients against each of the other betas show that the coefficients are uniformly distributed. The exceptions are only in betas for texture vs smoothness, smoothness vs fractal dimensions and a little in texture vs fractal dimension; texture vs smoothness scatterplot shows a positive correlation. Said more clearly, if we had larger standard deviations of tumor texture, the standard deviation of smoothness s more likely to be larger as well. On the contrary, smoothness and fractal dimension seem to exhibit a negative association. The scatterplot shows that it is likely that there is either a larger effect of smoothness data set and a smaller effect of fractal dimension data set or vise-versa. Like the smoothness vs fractal dimension scatterplot, the scatterplot for texture vs fractal dimension, shows a negative but weak correlation – almost uniform- between the two variables.

With these discoveries made among predictors, a new model with interactions is trained in the third experiment.
The resulting model has a new Pseudo-R^2 value of 0.4819. Not so much of an improvement from the model in the second experiment. Yet, when evaluation is made, the model now has a better recall score of 82.1%, thereby cutting the number of False Negatives to 10 from 20 from the model in the second experiment. Even though this model is sufficiently less accurate than one in the first experiment (82% accurate and 73% precision score), it does say a lot about the

informativeness, effectiveness even, of interaction terms. Addition of interaction terms could be a creative way of engineering our data in circumstances where we have a shortage of data features. And given the nature of our task, if we were to make a choice between which model between the one from the second experiment or this experiment (third), we would pick this experiment as it favors the minimization of the more costly inaccurate diagnosis i.e., False Negatives.

Section VII
RESULTS AND DISCUSSION

The design of the various experiments answered all the questions the project aimed to answer. The Bayesian Inferences and statsmodels logistic models are not any less inaccurate than Moo's sklearn's logistic regression model. The former models are in fact handy as they readily inform the user of the goodness of fit of the model, thereby more or less implicitly warning the user on the likelihood of our model to be wrong. It would sure make a difference if all underlying models in the various breast cancer methods had this `feature` in them. So many patients would then know whether to take the news of diagnosis with or without a grain.

If patients with tumors only differ in age and all else is the same, that does not affect the diagnosis outcomes so much according to the dataset.

And if the standard deviation of the texture is large, the standard deviation of the smoothness is likely large too. However, if the smoothness is large, it is likely that the fractal dimension is small or vise versa.

A tumor's smoothness is the most important predictor in the diagnosis of cancer, followed by its fractal dimension, and then texture. All other non-multicollinear independent predictors do not weight as much in the diagnosis of cancer.

Section VIII
CONCLUSION AND FUTURE WORK

Overall, as researchers strive to alleviate the number of deaths incurred from breast cancer through early accurate diagnosis, there are so many tradeoffs they make when it comes to finding the right model to diagnose breast cancer. Perhaps, tagging information about goodness of these models to users in a handy way would make a difference in the way breast cancer is currently diagnosed. More so in all cases, I propose that if we make this tough decision and we will have to, that we uphold models that favor the least number of false negative diagnoses than false positives. For, it is better to make 1000 trips for 3 years, or forever, to a hospital and later realize you were in fact not positive than to stay home thinking you had a benign cyst, only to discover you not only have cancer but it is now stage IV and you'll have to die in the next few months, leaving your loved ones, kids, and family in grave pain because breast cancer is that deadly.
In this paper I propose the active incorporation of Bayesian Inferences and statsmodels in the accurate diagnosis of breast cancer. My future work, I will propose the deep learning method convolutional neural network, a way more accurate model mostly used for classification of images dataset that employs Bayesian Inferences to report informative findings. Bayesian Inferences are handy in delivering informative results, so let us build models that leverage the power of Bayes theorem to save patients from breast cancer through early and accurate diagnosis!

BIBLIOGRAPHY

1. Hayes A. (2021). Multicollinearity What is Multicollinearity Ref: https://www.investopedia.com/terms/m/multicollinearity.asp

2. Habshah M et al., (2013) Collinearity diagnostics of binary logistic regression model. Ref: https://www.tandfonline.com/doi/abs/10.1080/09720502.2010.10700699?journalCode=tjim20

3. Statsmodels. Outliers Influence Variance Inflation Factor. Ref: https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html

4. Giselmar A et al., (2016). Log-Likelihood- based Pseuod-R^2 in logistic Regression: *Driving Sample-sensitive Benchmarks.* Ref: https://journals.sagepub.com/doi/pdf/10.1177/0049124116638107

RELATED WORK
Article:
Kaiser M. (2020). Predicting Breast Cancer Using Logistic Regression. *Learning how to perform Explanatory Data Analysis, apply mean imputation, build a classification algorithm, and interpret the results.* Ref: https://medium.com/swlh/predicting-breast-cancer-using-logistic-regression-3cbb796ab931

AUTHOR’S GitHub REPOSITORY (contains dataset too):
Salifyanji J Namwila (2021). The Bayesian Perspective on Breast Cancer Diagnosis. GitHub Repository: https://github.com/Sheinstein/-Bayesian-Perspective-on-Breast-Cancer-Diagnosis

AUTHOR’S NOTEBOOK:
https://colab.research.google.com/drive/1hu6243boRjsLC4i54Dh1DEKeIOSHDcDl?usp=sharing