Audience Persona

Amit is the head of sales of a reputed pharmaceutical company. He is curious in knowing how sales can be influenced by data driven solutions. He has a basic understanding of machine learning technicalities and wants to know how it is used in practice. Amit is interested in knowing if a data science product or modeling approach is a feasible solution to his business.

GitHub URL for report - https://github.ubc.ca/mds-2021-22/DSCI_542_lab2_calexa03/blob/master/Contraceptive_predictor_report.pdf

GitHub URL for approach - https://github.com/UBC-MDS/contraceptive_method_predictor

Study of contraceptive usage prediction model based on Support vector classifier in Indonesia

Abstract

Contraceptives have been the current generation's solution to unplanned pregnancies, psychological problems related to childbirth and so on. During the recent years, there has been a significant increase in the sales of various types of contraceptives confirming on how it has become an inherent part of the society. Though contraceptives do have its own side effects like every other drug, the use of them has advanced the human rights of people to determine the number and spacing of their children [1].

Here we propose a machine learning algorithm to predict the use of contraceptive amongst married women given their demographic and socio-economic status such as education, number of children, working status, etc. We use the Support Vector classifier algorithm to perform the analysis. The evaluation metric used was Accuracy, the model achieved an accuracy of 74% and the area under the curve (AUC) of 78%. That is given a married woman our model can predict with 74% accuracy (74 out of 100 times accurately) if she uses contraceptive or not.

While this prediction will not be ethical to be deployed and put into production as it requires sensitive information regarding the person. It can be used internally to analyze the market by forecasting sales of contraceptives and the amount of investment the company can put in based on the current market which is very dynamic. It will also help understand patterns and the factors that contribute to sales in order to make a more data driven strategy to a business model.

Introduction

The contraceptive usage has not just increased but has identified its own market. The market is projected to reach 30.15 billion USD by 2027 [2]. As a part of the pharmaceutical industry, it is important to understand the factors that influence this highly profitable sector. The contraceptive industry is not just about the number of total customers but the number of repetitive customers who use contraceptive on a regular basis which has also increased over the years. [3]

Figure 1 shows the increase in Contraceptive usage and the frequency of usage within customers. We can observe the increasing trend in the 21st century. This gives us more reason to believe that the industry should not only focus on expanding the market but adopt the right kind of analytical practices that help identify such patterns which help exploit the existing market to its potential through data driven solutions.

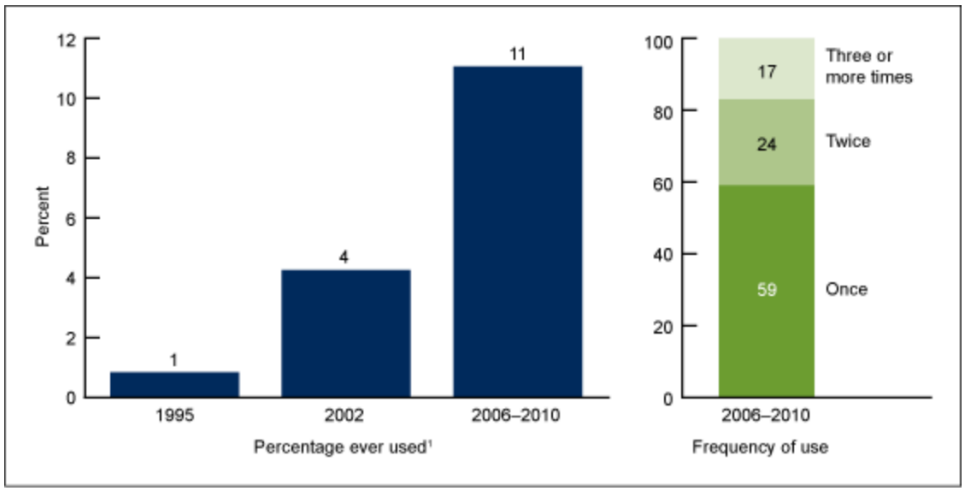

Figure 1 Percentage of sexually experienced women who have ever used emergency contraception has increased over time

There have been recent studies which show such intricate micro sectors within the market. One such example is given below where low income is related to contraceptive usage. [4] As observed in Figure 2, the market is segmented by highly social economic factors such as Gender, Region, poverty etc. Therefore, the data required to perform analysis is scarcely available and its mostly released through competitions which have been conducted through surveys.

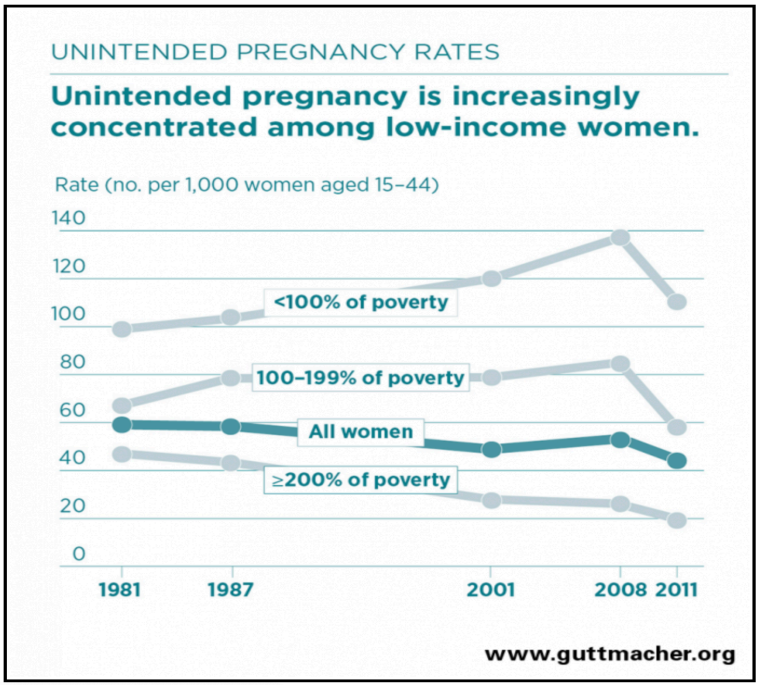

Figure 2 Relation between income and contraceptive usage

This is the reason why the industry should adapt to Machine Learning (ML) algorithms which we propose here so that you can identify such patterns for internal purposes without needing to conduct surveys at regular intervals. A ML model can help in predictive analysis and give out more nonlinear trends present in the data without the need for much financial costs.

Though there has been a vast work on contraceptive usage analysis in recent times [5] through competitions and publications. We have not observed anything related to the demographic and socio-economic characteristics of married woman in the study. In addition, there is not much on data from Indonesia in specific. In our approach, we have constructed a contraceptive usage predictor for married woman based on Age, Education, Husband's education, Number of children ever born, Religion, Working status, Husband occupation, Standard of living index, Media exposure. These features are highly sensitive but give an accurate analysis of the given market in Indonesia. The model is based on Support Vector classifier algorithm to classify target levels 1 – Use of contraceptive and 0 – No use of contraceptive.
The approach used was on a limited set of data and only a few features, but the model serves as a starter to show the wide benefits that machine learning can provide to develop strategies based on forecasting the use of contraceptives among women.

Methods

Data and exploratory analysis

The data is open sourced from UCI Machine Learning Repository (Dua and Graff 2017) [6]. It is a survey conducted in Indonesia of married woman who were either pregnant or weren't aware if they were at the time of this survey.

Column name	Description	Type	Values
Wife age	Wife's age	Numerical	any positive values
Wife education	Wife's education	Categorical	1=low, 2, 3, 4=high
Husband education	Husband's education	Categorical	1=low, 2, 3, 4=high
Number of children ever born	Number of children ever born	Numerical	any positive values
Wife religion	Wife's religion	Binary	0=Non-Islam, 1=Islam
Wife now working?	Is wife working or not	Binary	0=Yes, 1=No
Husband occupation	Husband's occupation	Categorical	1, 2, 3, 4
Standard-of-living index	Standard-of-living Index	Categorical	1=low, 2, 3, 4=high
Media Exposure	Media exposure	Binary	0=Good, 1=Not good
Contraceptive method used	Contraceptive method used (Class Attribute)	Categorical	1=No-use, 2=Long-term, 3=Short-term

Figure 3 Data description

The sample contains of 1031 individuals where each row in the data set represents the sample taken from women, including her demographic information and their socio-economic status represented as columns. The target was recorded as Contraceptive method used. The data is limited to only a few attributes as these are sensitive information and participants were reluctant beyond a point. The data description is as shown above in Figure 3

The following design approach was followed for an efficient model building process. It involved fetching data from the cloud to our local system. The data was then split to train and test with

training set containing 70% of the data, the split was done such a way that the train and test are split with the same ratio of target levels. The test set was used only for the final model predictions to ensure no data leak of test information to the model. The training set undergoes various stages before the model prediction such as

- Data cleaning/preprocessing – where we prepare the data to make it model ready,
- Exploratory data analysis (EDA) – this is where we explore the initial patterns showcased by the collected data through plots and queries,
- Model building – choosing the right machine learning algorithm and hyperparameters.
- Prediction – The finalized model is used for predicting on test data.

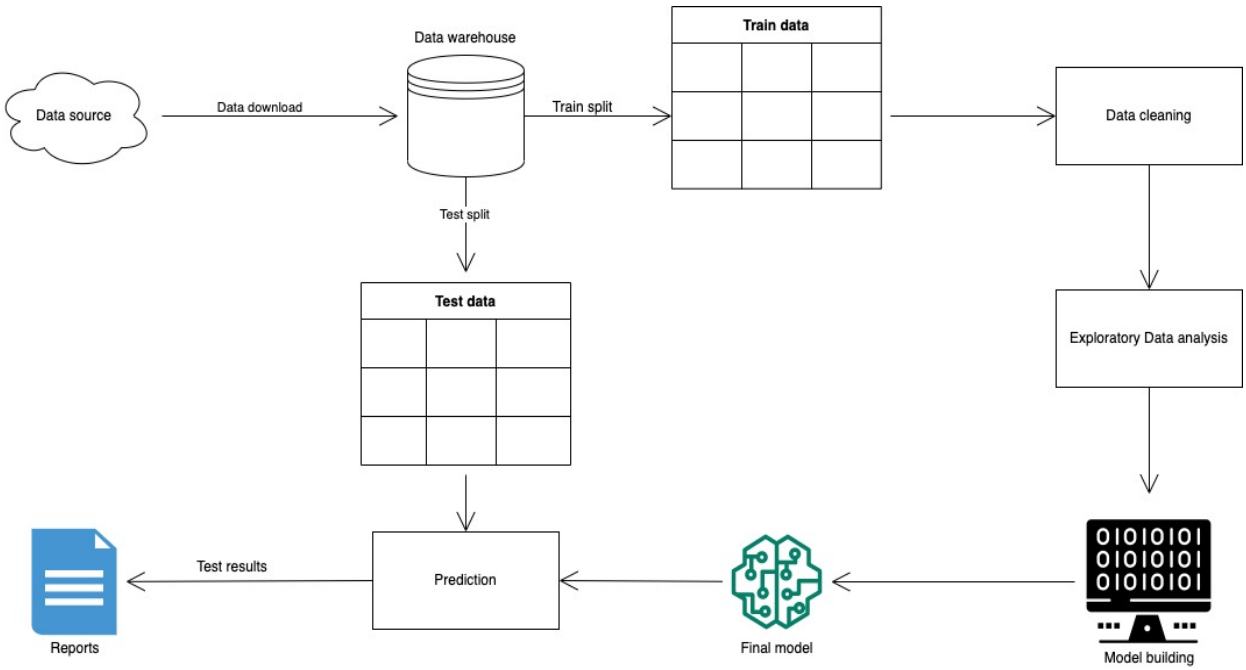

Figure 4 Design approach

The target had three sub categories/classes representing the contraceptive usage level i.e. 1 (“No-use”), 2 (“Long-term use”), followed by 3 (“Short-term use”) .

The data had no significant class imbalance for us to handle. But in order to not dilute from our problem statement we would combine levels 2 and 3 to “Use” vs 1 as “No use” in further steps.

On visualizing our data (Figure 5,) we found that Media exposure and Wife working status could play an important role in predicting the use of contraceptive as it had a greater number of records associated to target 2 and 3 in our data. Which isn’t actually true as we show in the further steps after model building. Another reason to know that not all patterns which are visible to naked eye are the true reasons for the outcomes, further pressing on the need for machine learning.

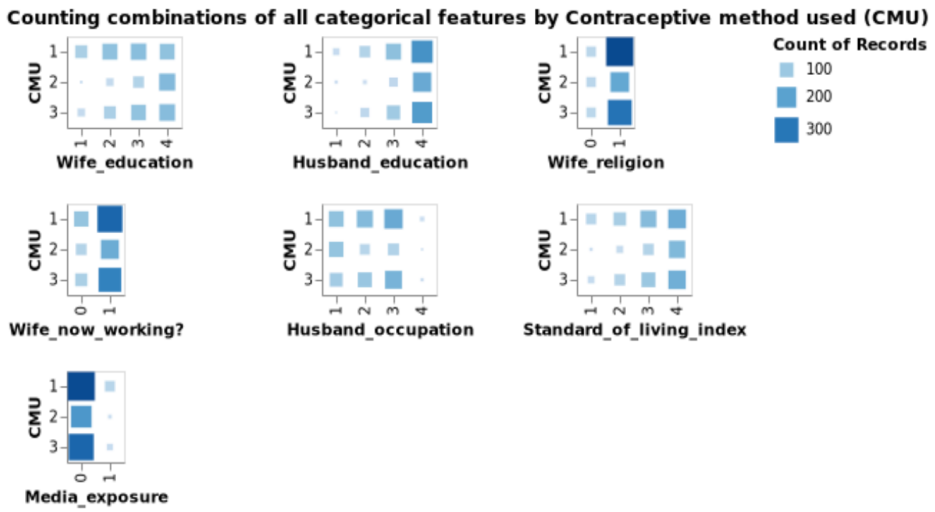

Figure 5 How some features look important on the outset.

Data preprocessing

As mentioned earlier we combine levels of the target into 2 levels where 0 = No-use and 1 = Use. The distribution of our target then changes to 45% and 55% observations of “No-use” and “Use” respectively.

The following table Fig 6 shows different features in the data and how they were encoded to convert the given data to machine readable form for our model. The data as such had no missing values for treatment. The outliers were left untreated as they would be handled by our machine learning algorithm. Most packages used for the project are provided by scikit-learn.

Data Type	Variable	Transformation performed	Technique used
Numerical	Wife’s age, Number of children ever born	Scaling	Standard Scaling
Ordinal	Wife’s education, Husband Education,	Encoding	Ordinal Encoding
Ordinal	Husband’s Occupation,Standard of living Index	Encoding	Ordinal Encoding
Binary	Wife’s religion, Wife working Media Exposure	None	Pass through

Figure 6 Data encoding details

Data Analysis

The final model built on was using the support vector classification algorithm which assumes that the data distribution is independent of each other (i.e., rows are not related to each other) and are identically distributed. The evaluation metric chosen was accuracy as there was no class imbalance in our problem. Accuracy is given by the number of correct predictions out of the total number of samples. The baseline metric was given by a Dummy Classifier with an accuracy of

55% which is the score our model is trying to beat. Below this score would be considered not model worthy.
Even thou we have considered accuracy as the evaluation metric; we will keep a close observation on other metrics such as recall (which is the true positive rate of the positive class in our case "Use" of contraceptive) and the difference between the training and validation accuracy scores to check if our model is generalizing well.
As a process of model selection, multiple machine learning algorithms were tried. To name a few - Decision Tree, K- nearest neighbors, Logistic regression, etc. But Support vector classifier (SVC) performed the best in terms of accuracy for our data without any hyper parameter optimization. To further improve model scores Hyper - parameter optimization (i.e., Model parameter tuning) was performed on SVC algorithm by Random Search Cross validation method (RandomSearchCV) which improved model scores by 8% giving us our final model.
The final model was used to predict the labels of the test data unknown to the model up until now. The model achieved an accuracy of 74% which means given 100 samples of married women with their socio-demographic attributes, our model can accurately classify 74% women if she used contraceptive or not.

In terms of scalability, the analysis does not require much computational power. The program was executed in under 5 minutes on a 16gb RAM and Intel i5 processor machine. For easier use of the software, we have used Docker to export our software across all operating system and the set-up is quite intuitive.

Results

The final model was evaluated on accuracy and recall of the positive class which we considered to be predicting the usage of contraceptives. The prediction of the final model was done on the test data. The classification report is shown in Fig 7. The recall value of **0.90 (90%)** indicates a good true positive rate (TPR) for the "Usage" class while the **0.53 (53%)** indicates the TPR of the "No use" class.

X	precision	recall	f1.score	support
contra_no	0.7822581	0.5271739	0.6298701	184
contra_yes	0.7264151	0.8953488	0.8020833	258
accuracy	NA	NA	0.7420814	NA
macro avg	0.7543366	0.7112614	0.7159767	442
weighted avg	0.7496619	0.7420814	0.7303928	442

Figure 7 Classification report

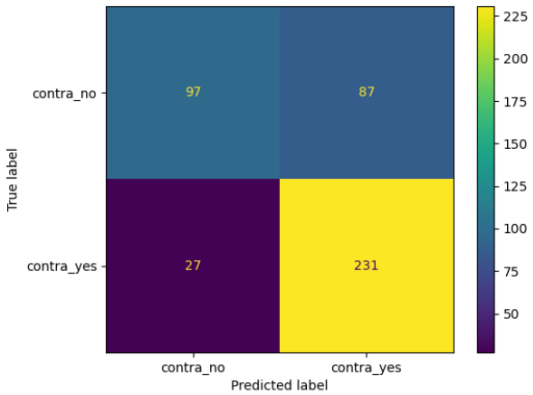

Figure 8 Confusion Matrix

The confusion matrix is shown in Figure 8. On observing the matrix, the model is predicting well on the total number of True positives i.e. 231 which are the woman that the model predicted correctly to be using contraceptive method out of 258. And True Negatives i.e. 97 which are the

woman that the model predicts correctly for not using contraceptive method. However, there are some false positives and false negatives observed as well.
False positives are detected when we affirmatively predict the usage of contraceptives when in fact, the person does not use contraceptives i.e. in our matrix 87 and False Negatives indicated when we incorrectly predict the person is not using a contraceptive, when they are in reality using contraceptives.

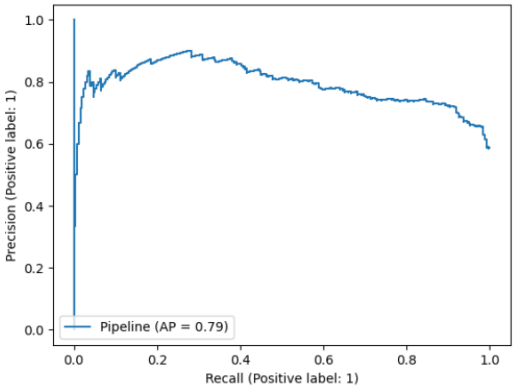

Figure 9 Precision – Recall curve

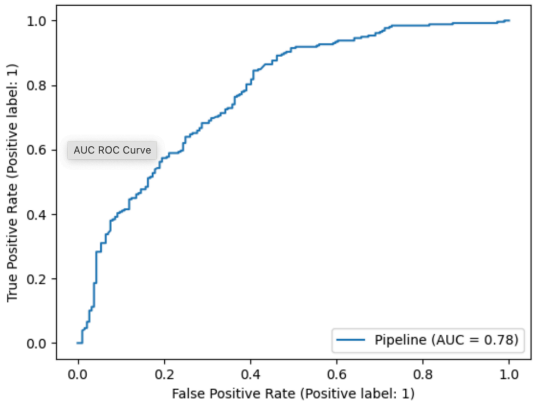

Figure 10 ROC curve

The Precision Recall tradeoff was analyzed using the PR curve Fig 9 where the model achieved a good score of 79%. The ROC curve of our model for different thresholds were plotted Fig 10. We achieved an AUC of 78% which tell us the probability that our model can classify a woman who uses contraceptive over a woman who doesn't use them.

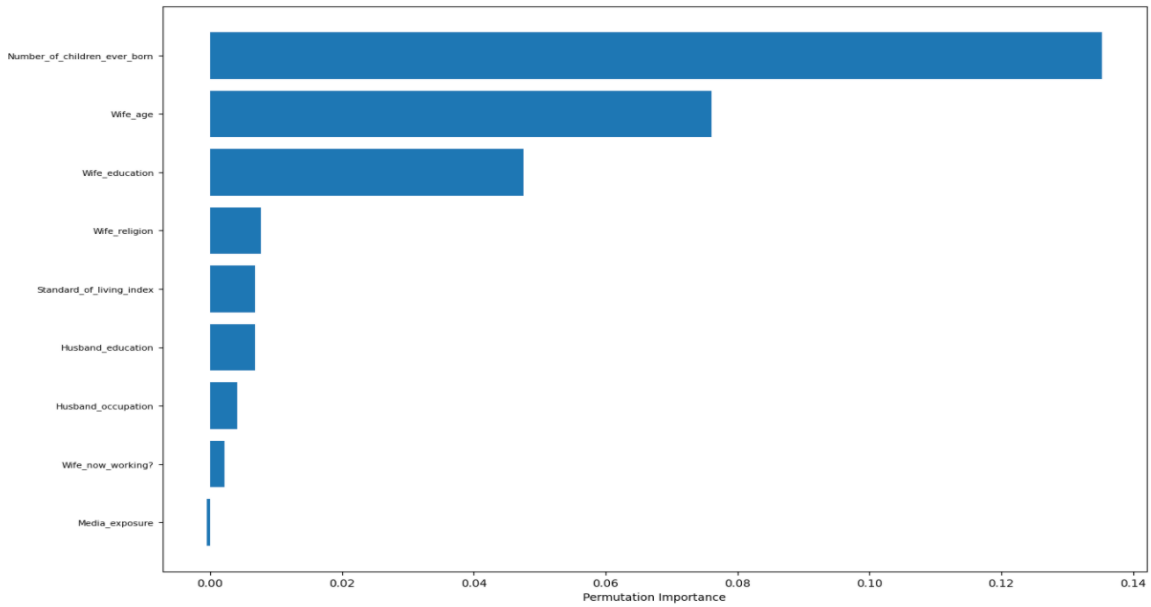

Figure 11 Feature importance

The model also helps users identify the reasons contributing to the usage of contraceptives. As we can see in figure 11. Factors such as number of children born to the married woman, the age and her education matter the most in deciding if she decides to use contraceptives or not. Though

the model does not clearly say what values of these features contribute to the use of contraceptive. It is a work in progress to show such a relation. As mentioned before Media exposure which looked like an important feature before doesn't really play a major role in the usage of contraceptives.

Conclusion

The feature importance helps the company decide on which sector to invest more on. In this instance, it feels that the company could invest more of its funds into micro segmenting the market based on the number of children in a particular province and the age of married woman to develop different strategies in those segmented areas. It also shows that media exposure isn't the prime reason for the sale of contraceptives, therefore we can cut corners in terms of funding for marketing strategies. These data driven decisions can reduce unnecessary cost to the company while putting their money where the rewards are more.

The predictive algorithm can help predict the number of customers which will use contraceptives with an accuracy of 74%. The model can be used in new regions which the company is trying to explore its sales by predicting on a sample if the audience are likely to use a contraceptive. This helps business to take a calculated risk on their investment. As mentioned, the model would also give out factors contributing to the usage of contraceptives which can help the business understand that region better.

References

[1] Key Facts of family planning and contraception methods https://www.who.int/news-room/fact-sheets/detail/family-planning-contraception
[2] Market segmentation of Contraceptive industry https://www.fortunebusinessinsights.com/industry-reports/contraceptives-market-100064
[3] Percentage of contraception usage among sexually experienced women has increased over time. https://www.cdc.gov/nchs/products/databriefs/db112.htm
[4] Unintended pregnancy rates in the USA https://www.guttmacher.org/fact-sheet/unintended-pregnancy-united-states
[5] USAID's Intelligent forecasting: A competition to model future contraceptive use https://www.usaid.gov/global-health/health-areas/family-planning/usaids-intelligent-forecasting-competition-model-future
[6] Contraceptive Method Choice Data Set https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice