201574114.pdf

by Surya Govindasamy Ganesan

Submission date: 10-May-2022 01:54AM (UTC+0100)
Submission ID: 179252003
File name: 201574114.pdf (483.56K)
Word count: 2310
Character count: 12155

Leeds University Business School

UNIVERSITY OF LEEDS

Assessed Coursework Coversheet

For use with *individual* assessed work

Student ID Number:	
Module Code:	LUBS1530
Module Title:	Business Analytics 1
Module Leader:	Dr Bill Gerrard
Declared Word Count:	1858

Please Note:

Your declared word count must be accurate, and should not mislead. Making a fraudulent statement concerning the work submitted for assessment could be considered academic malpractice and investigated as such. If the amount of work submitted is higher than that specified by the word limit or that declared on your word count, this may be reflected in the mark awarded and noted through individual feedback given to you.

It is not acceptable to present matters of substance, which should be included in the main body of the text, in the appendices ("appendix abuse"). It is not acceptable to attempt to hide words in graphs and diagrams; only text which is strictly necessary should be included in graphs and diagrams.

By submitting an assignment you confirm you have read and understood the University of Leeds **Declaration of Academic Integrity** (http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf).

LUBS 1530

Pregnancy and Customer Purchase Behavior

Table of Contents

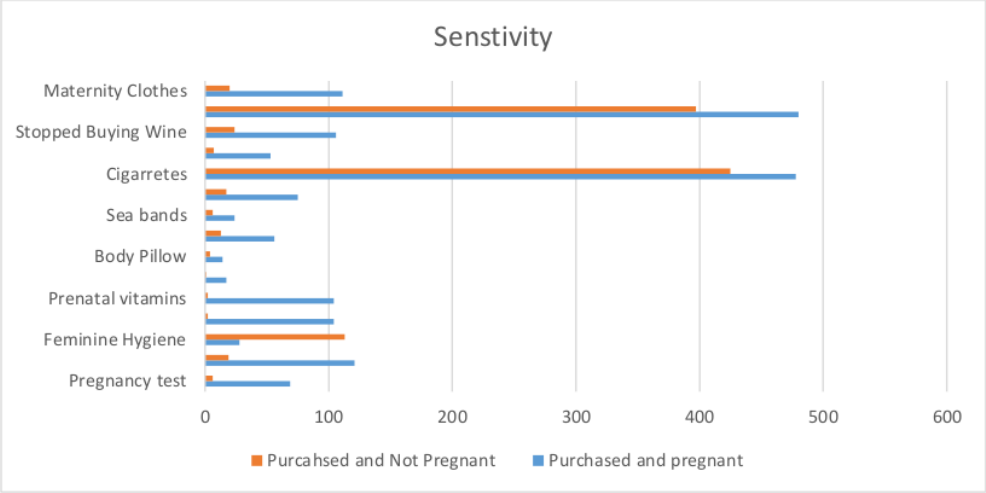
Introduction

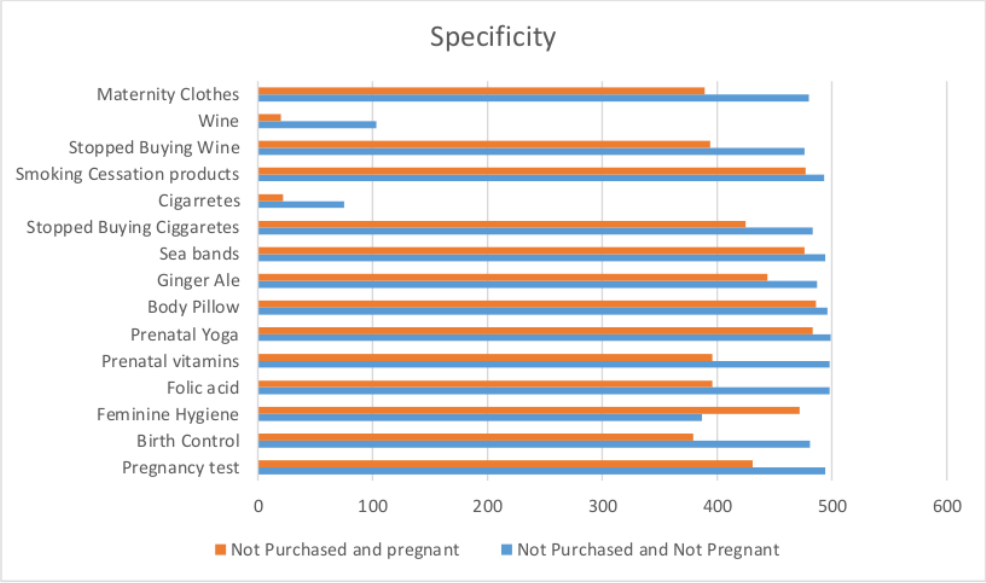
The purchasing history of 1000 customers have been recorded and compiled as a dataset. This dataset is an amalgam of 18 categorical variables, that provides information about customer behaviors on buying 12 products, gender, address types and pregnancies. 16 of the categorical variables are binary with the responses 0 and 1, while two other variables are trinary. The given dataset holds data about the purchase history of customers, half of whom are pregnant, and the other half aren't. The purpose of this report is to concoct a predictive tool to correctly identify the households with pregnancies that will be valid for any dataset.

Key Features of Data

The recorded dataset of the purchase history of 1000 customers has 19 variables, each of which will be analyzed to test their fit as a predictor variable for our predictive tool. The dataset contains 18 categorical variables and 1 continuous variable. The continuous variable, 'Account holder ID', is the unique identification variable for each customer. The 'gender' and 'Home/apt/PO' variables are trinary categorical variables, that record customer gender and address types. The 'pregnant' feature is a binary variable that tells if a household has a pregnancy. This is also the reference variable that will be used to measure the accuracy of the proposed predictive tool. The rest of the 15 binary variables inform on the purchase behavior of customers on 12 different products, with two responses – 0 and 1.

Different predictor variables can have different levels of power to contribute to the output variable. While some predictors are more probable to identify purchasers who are pregnant, others are strongly suited to identify non-pregnancies among non-purchasers. The base assumption in the graphic table below is that the purchase of an item contributes towards higher chances of pregnancy and non-purchases lowers the chances of pregnancy. The exception here is for these categories: Wine, Cigarettes and Birth control, where the opposite is true. The sensitivity bar helps us to draw the likelihood of pregnancy in a household when an item is purchased. The specificity bar displays the likelihood of non-pregnancy when a certain item hasn't been purchased recently.





While it is clear from the sensitivity graph that there are more people who've purchased cigarettes and wine than recorded in the 'Stopped buying' categories, hence giving us a bigger dataset to analyze pregnancy likelihood, customers recorded in the latter category has a higher ratio of pregnant vs not pregnant. This could imply that these two categories are better indicators to sensitivity than cigarettes and wine. On the other hand, 'Feminine Hygiene' displays a poor ability in both sensitivity and specificity.

Not all the several features available in the dataset contribute to the results of pregnancy. To analyze the importance of each variable, if any, we are required to test the significance of each feature in contributing to the final output results. As both the predictor variables and output variable, in this case the pregnancy results, are categorical variables.

A Chi-squared test has been used to understand the significance of each customer item that has been purchased. This is done by first building a matrix table with the pregnancy results against purchasing decisions in the following format.

Prenatal yoga

Observed frequencies

	Not Pregnant	Pregnant	Total
Not purchased	499	483	982
Purchased	1	17	18
Total	**500**	**500**	**1000**

Expected frequencies

	Not Pregnant	Pregnant	Total
Not purchased	491	491	982
Purchased	9	9	18
Total	**500**	**500**	**1000**

p-value	0.000141437
Test-statistic	14.48291469

For the next step a null hypothesis is assumed that the pregnancy results are independent of the items purchased by customers. Considering that the significant p-value to be 0.05, below which the null hypothesis is void, a table of the features against their p-values are listed in a descending order.

Feature	p - value
Body Pillow	0.017381985
Sea Bands	0.000847587
Prenatal Yoga	0.000141437
Ginger Ale	8.09529E-08
Cigarettes	1.48754E-08
Smoking Cessation Products	9.05858E-10
Stopped Buying Cigarettes	2.21172E-10
Pregnancy test	3.91513E-14
Stopped Buying Wine	1.25326E-14
Feminine Hygiene	1.132E-14
Wine	1.33358E-15
Maternity Clothes	1.47675E-17
Birth control	1.46099E-20
Folic acid	1.09023E-25
Prenatal vitamin	1.09023E-25

The body pillow feature with a p-value more than 0.05, identifies itself as a statistically insignificant variable in disproving the null hypothesis. While on the other hand, all other p-values are lower than 0.05, proving their significance to the output as far as the chi-squared test is concerned.

Predictive tool

All potential predictor variables in the dataset were found to be categorical, while the output variable of predicted pregnancy results will also be categorical and binary – 0 and 1. Therefore Logistic regression reaped the best results for the project. The presence of an extensive array of predictor variables made the use of classification trees an inefficient method considering the enormous number of branches that will be divided to produce the results. For this the set of binary variables that is most apt to predict the outcome variable is chosen, as displayed below. The birth control feature has been converted into its negative, 'No birth control' so that all variables involved have positive influence over pregnancy chances. The 'stopped buying cigarettes' and 'stopped buying wine' were better indicators of pregnant households than the 'cigarettes' and 'wine' features, making it reasonable to include only one variable informing a product's sales behavior. The 'Feminine hygiene' feature was eliminated based on the sensitivity and specificity charts as mentioned in the previous section.

Binary Variables

1. No Birth control
2. Pregnancy test
3. Folic acid
4. Prenatal vitamin
5. Prenatal Yoga
6. Ginger Ale
7. Sea Bands
8. Stopped Buying Cigarettes
9. Smoking Cessation Products
10. Stopped Buying Wine
11. Maternity Clothes

With the list of the above-mentioned variables a linear equation needs to be constructed to find the odds of pregnancy for each customer. If P is the probability of pregnancy, then odds of pregnancy is represented by:

$$\textbf{Odds} = P / (1 - P)$$

To find the odds, the log function of Odds, referred to by the phrase 'logit', is represented as a linear function of said variables.

$$\textbf{Logit} = a + bx + cy + dz \ \ldots$$

Here **'a'** is the intercept, b, c, d and the following first letters are co-effacements. The seconded letters represented by x, y and z are the filtered binary variables. The odds of pregnancy solved through the equation is then used to find the probability of pregnancy, leading to another function called Log Likelihood. Likelihood is the probability that those results of pregnancy, irrespective of the outcome, could be true. To find the best suited coefficients to predict pregnancies, the sum of the Log Likelihood for all customers should try to be maximized by changing the coefficients of the variables accordingly. This process is necessary to find the output variable for each customer - the probability of pregnancy and is carried out by the solver add-in in excel. On solving the equation, the following results were obtained:

Coefficients											
Intercept	No Birth Control	Pregnancy Test	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Ginger Ale	Sea Bands	Stopped Buying Cigarretes	Smoking Cessation products	Stopped Buying Wine	Maternity Clothes
3.2793	-2.1916	-2.4238	-4.0814	-2.3590	-2.5856	-1.7284	-1.1534	-1.4962	-1.7786	-1.5950	-1.8161

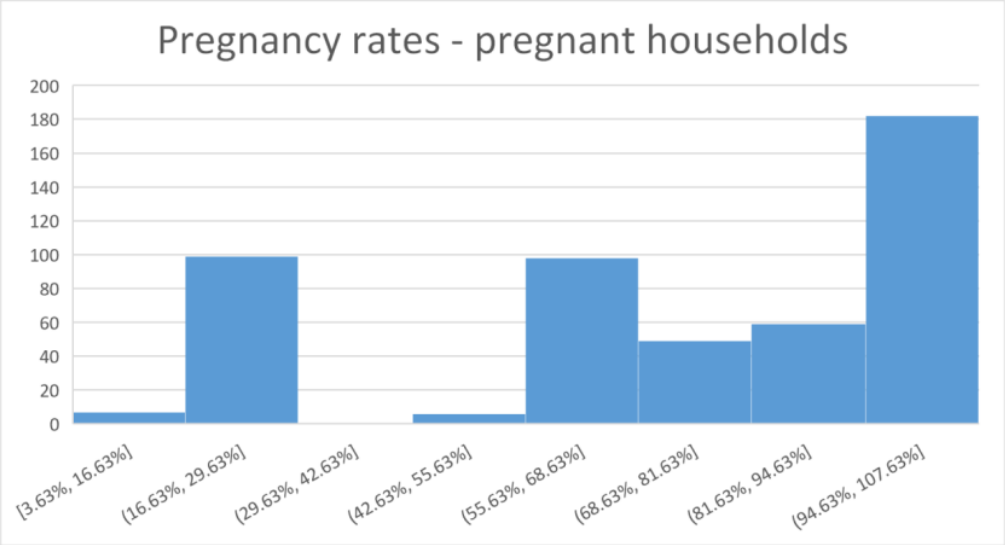
The sum log likelihood obtained through the solutions was -426.127. The final coefficients depict the purchase of folic acid to hold the highest predictive power among all features, followed by 'prenatal yoga' and 'pregnancy test'. The solutions also reveal sea bands to have the least significance on deciding the output variable.

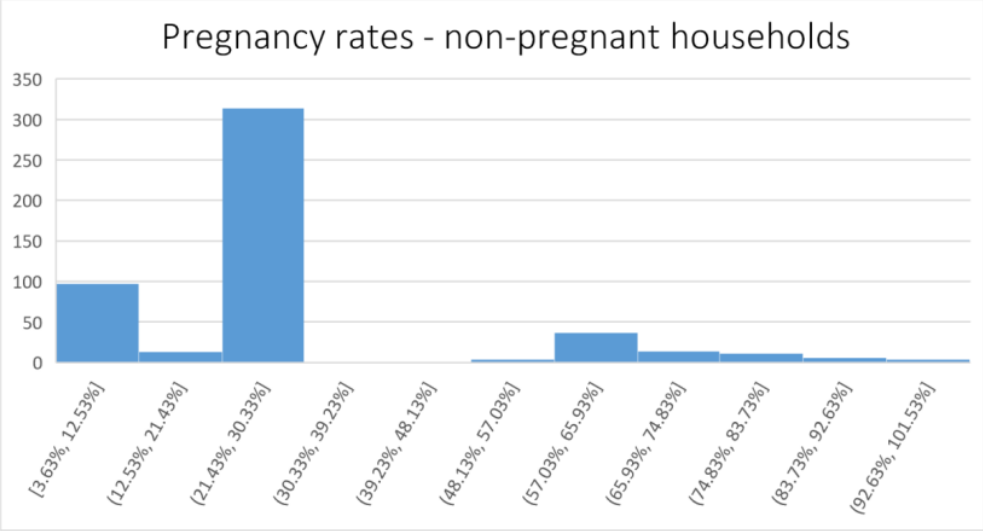
With the probability of pregnancy for all customers, it is now necessary to decide a cut-off value for the variable. Customers with pregnancy rates above this cut-off value are deemed pregnant, while customers below the value are marked not pregnant. This is the final binary outcome variable called prediction. As different cut-off values will have different values for the prediction feature, a measure of their accuracy is taken to determine the best cut-off value for use. The following table weighs the accuracy measures of different cut-off values against each other.

Measusures	20%	30%	40%	50%	60%	70%
Overall error rate	40.00%	18.20%	18.20%	18.20%	18.40%	23.20%
Sensitivity	98.00%	78.80%	78.80%	78.80%	77.60%	57.80%
Specificity	22.00%	84.80%	84.80%	84.80%	85.60%	95.80%
Precision	55.68%	83.83%	83.83%	83.83%	84.35%	93.23%

The absence of customers with pregnancy rates between 30% and 50% has resulted in the accuracy measures of 30%, 40% and 50% being the same. Nevertheless, 50% was found to have the lowest error rate and highest sensitivity. In deciding the cut-off values, the overall error rate and sensitivity are given preference as getting more right pregnancies has been prioritized. Therefore, the ideal cut-off value to determine the output variable, 'prediction', is 50%.

Through the method of logistic regression, the developed predictive tool has correctly identified 78.8% of pregnant households, correctly identified 84.8% of non-pregnant households, while misclassifying 18.2% of the recorded households. Furthermore, analyzing the histogram summarized below depicts the distributional properties of pregnancy rates among both pregnant and non-pregnant households.


Pregnancy rates - pregnant households
200
180
160
140
120
100
80
60
40
20
0
[3.63%, 16.63%]
(16.63%, 29.63%]
(29.63%, 42.63%]
(42.63%, 55.63%]
(55.63%, 68.63%]
(68.63%, 81.63%]
(81.63%, 94.63%]
(94.63%, 107.63%]


Pregnancy rates - non-pregnant households
350
300
250
200
150
100
50
0
[3.63%, 12.53%]
(12.53%, 21.43%]
(21.43%, 30.33%]
(30.33%, 39.23%]
(39.23%, 48.13%]
(48.13%, 57.03%]
(57.03%, 65.93%]
(65.93%, 74.83%]
(74.83%, 83.73%]
(83.73%, 92.63%]
(92.63%, 101.53%]

Among pregnant households, the highest concentration of pregnancy rates is between [94.63%, 100%]. While this result is foreseeable, it is counterintuitive that the second highest concentration of pregnancy rates is in two categories, [16.63%, 29.63%] and [55.63%, 68.63%]. Another trend noted in the pregnancy rates among non-pregnant households is the higher concentration of non-pregnant people in the range [21.63%, 30.63%] than in the range [0%, 21.63%]. Overall, the high accuracy measures speak to the model's generalizability and application to other datasets or even as a tool for forecasting. While the histogram does not reflect a bell curve, its unconventional distribution could be an effect of the absence of pregnancy rates between 30% and 50%. This could be further improved with better or more predictors in the predictive tool.

8

9

10

Reflection

There are several key factors that determine the success of a data analytics project. For any such data analytics project the purpose of the project must be well defined to provide a direction to it. Finding the purpose of the project requires the analyst to communicate effectively with the clients to understand their needs and the goals they're trying to achieve through the project. In the exploration stage of the project, it is important to understand the type of variables available for analysis and the type of outcome variable needed. This allows for any data analytics project to decide on the right tests for variables and models to build the predictive tool. After noting the type of variables available and needed appropriate tests should be conducted on variables to discover their influence on the outcome variable. This is crucial as not all features directly influence the outcome variable. The test on the variable's influence on the outcome can help filter out only the significant variables. This step of data transformation helps build a better predictive tool as more features doesn't directly translate to higher accuracy in results.

While building the predictive tool is the heart of any data analytics project, effective data is of little use when it is not properly represented. Graphical and tabular display of obtained results can effectively convey the results of a project to clients or help clients be more informed on the data given by them. Hence graphical description is an important factor among others for any data analytics project to be successful. Measuring accuracy of the obtained results helps check the hypothesis of an effective predictive tool indicating the how successful a data analytics project is. The higher is the accuracy of the results the better is the projects generalizability and applicability to other datasets. A project must be applicable to a different data set for it to be most useful beyond the current scenario and marks a successfully built data analytics project.

201574114.pdf

GRADEMARK REPORT

FINAL GRADE

70/100

GENERAL COMMENTS

Instructor

PAGE 1

Comment 1

ID?

PAGE 2

PAGE 3

PAGE 4

Comment 2

technical?

Comment 3

4. Also add the reflecion.

PAGE 5

PAGE 6

PAGE 7

Comment 4

10. Some good context and distributional statistics but what are the consequences of outliers?

PAGE 8

Comment 5

6. Some useful informational structure.

PAGE 9

Comment 6

6. Good proposed model.

PAGE 10

PAGE 11

Comment 7

12. Good technical details presented.

PAGE 12

PAGE 13

Comment 8

7. Good discussion. Training data?

Comment 9

4. Well written thtoughout.

Comment 10

7. Good and relevant figures and tables but add labels.

PAGE 14

Comment 11

14. Some useful insights.