SMART RETAIL ENSEMBLE TECHNIQUES FOR CUSTOMER ANALYSIS

Jayavelu Balaji

Illinois Institute of Technology Chicago and University of Chicago

Email: Jbalaji.ai@gmail.com

Abstract

Bagging Technique is one of the method introduced by Leo Bierman at UC Berkley. This method emphasis reducing the variance by retaining the Bias. This is possible if predictors average is taken in different spaces of the taken input feature Space. Its an Effective Technique because accuracy of the model is increased by using the multiple copies of it trained on different sets of data.

This project applies the bagging technique on Retail Mart Customer Data, and identifies the potential Customers by predicting the Customers who are Pregnant and who are expecting the babies in nearby Time. In this approach we can practically apply the models and accuracy of the model is plotted in ROC. This Project entirely relies on mathematical functions that is being executed on Excel.

Executive Summary

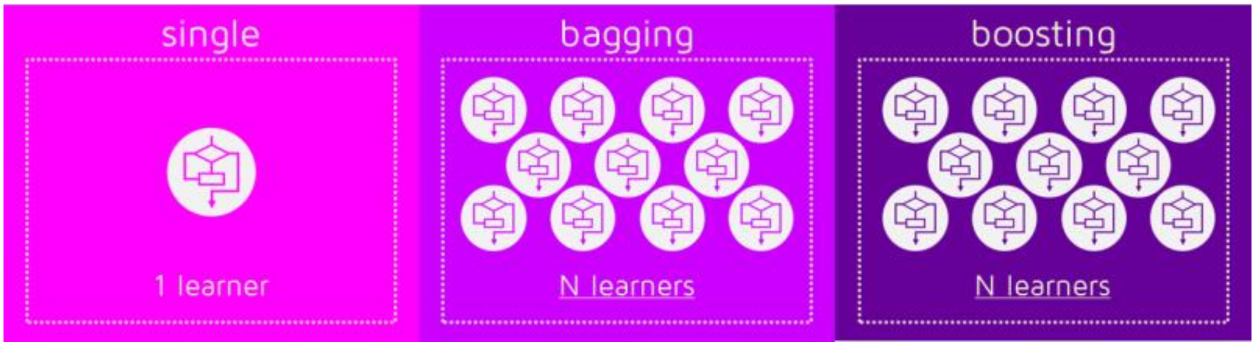
The Current Case study is taken from Ch 7 in Data Smart Book. This Project is improvised Version of the Ensemble Technique that is being applied on Retail Smart Customers.

Bagging Technique that Combines the Strong and Weak Learners are used to Predict the Predict Customers. This Classification Problem helps to identify the Potential Customers of Kids/Babies Product. In this Project implementation key mathematical functions are implemented using Excel.

Chapter-1

Introduction

Ensemble is a machine learning concept that uses same learning method to train multiple models. The ensemble makes multi classifiers to fuse together and makes the prediction. Trees are basic Building blocks for Bagging, Boosting and Random Forest. These are powerful Prediction Techniques.



Random Forest:

Classification trees can represent arbitrarily complex probability spaces P and hence hypothesis spaces H. As an instance, one can subdivide the unit interval [0,1] on the real line into segments of length with a tree that divides each parent segment in half. In higher multiple dimensions, to subdivide [0,1] into small hypercubes of side , build a tree that interleaves *D* interval-splitting trees. One can then assign separate label distributions to each hypercube, thereby approximating any desired distribution function to any degree.

Because of their expressiveness, classification trees must be curbed lest they overfit. The complexity of individual classification trees is typically controlled by pruning them after expansion. Empirical evidence shows that a better alternative is to use random forest classifiers, which combine the predictions of several trees.

A *random forest classifier* is a classifier that consists of a collection of classifier trees $f_m(\mathbf{x})$ for $m = 1,...,M$ that depend on independent identically distributed sets of parameters and each tree casts a unit vote for the most popular class at input **x**.

Several ways have been proposed and compared to each other [2, 7] to randomize the trees. These include the following: Bagging, that is, training each tree on a random subset of training samples drawn independently and uniformly at random from T with replacement.

Boosting, in which the random subsets of samples are drawn in sequence, each subset is drawn from a distribution that favors samples on which previous classifiers in the sequence failed, and classifiers are given a vote weight proportional to their performance .

Arcing, similar to boosting but without the final weighting of votes .

A particularly successful form of randomization combines bagging with random feature selection for each node of every tree in the forest [1, 5]. In Breiman's words,

(I) Its accuracy is as good as Adaboost and sometimes better.

(II) It's relatively robust to outliers and noise.

(III) It's faster than bagging or boosting.

(IV) It gives useful internal estimates of error, strength, correlation and variable importance. v It's simple and easily parallelized.

1.2 Objectives

- To classify RetailMart customers in a household
- To train an AI model based on Ensemble techniques
- To model the relationships between a dependent variable and one or more independent variable.
- Bagging Technique

Chapter-2

Mathematical Analysis

Given a set of n independent observations O1,O2,O3……On

Each having a variance of σ^2 and the variance of the mean $\bar{O}$ of the observations is given by σ^2/n

Averaging the set of observations would reduce the Variance and increase the prediction accuracy of the model by considering many training sets and build a separate prediction and averaging those predictions

In other Words we would calculate $f_1(x), f_2(x), f_3(x)..f_n(x)$ using N training sets and taking average of them would yield a single statistical low variance model

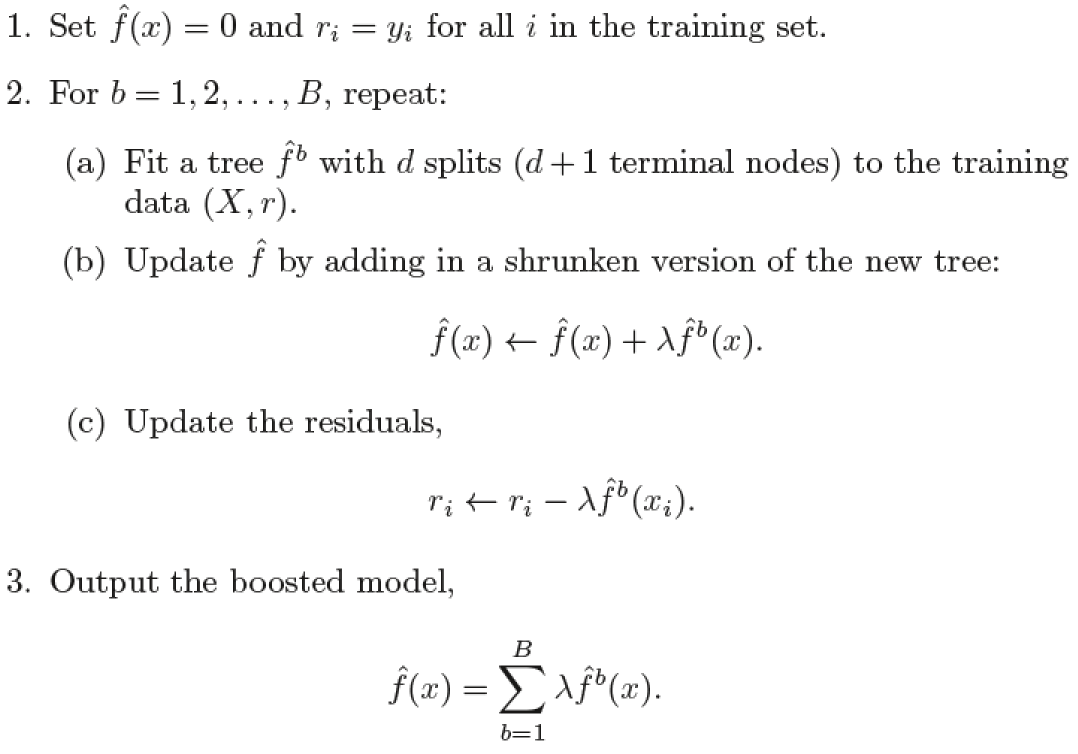
$$f_{avg}(x) \; = \; \frac{1}{N} \sum_{n=1}^{N} f(x)$$

We then train our method on the nth bootstrapped training datasets and then we finally average all the predictions

$$f_{bag}(x) \; = \frac{1}{N} \sum_{n=1}^{N} f^{*}(x)$$

This is called as Bagging

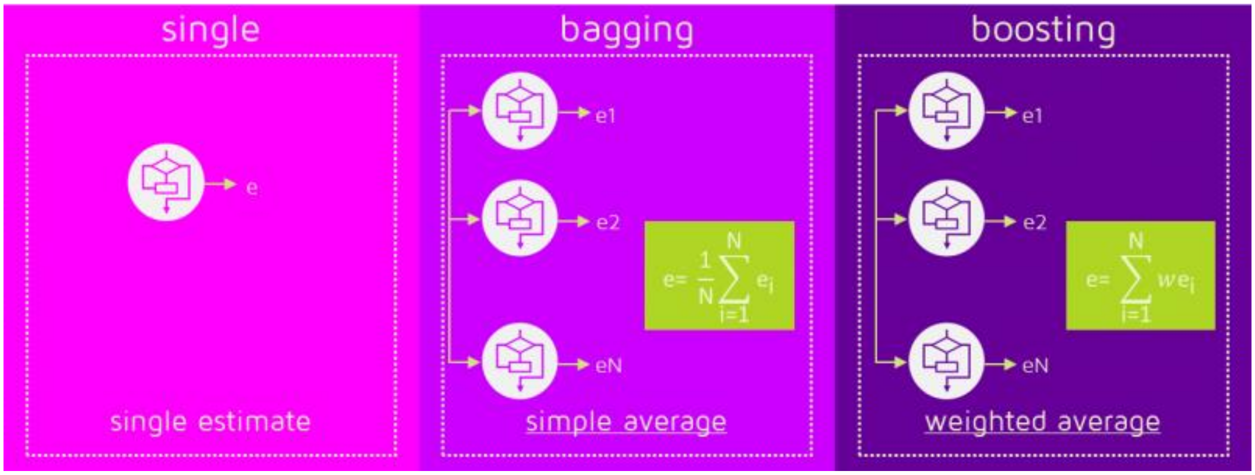
Algorithm Can be Depited as:



```
1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all $i$ in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
    (a) Fit a tree $\hat{f}^b$ with $d$ splits ($d + 1$ terminal nodes) to the training data $(X, r)$.
    (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
            $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$
    (c) Update the residuals,
            $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$
3. Output the boosted model,
            $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$
```

Bagging improves prediction for many Regression Methods, but it is highly useful in Decision Trees. In order to implement bagging on regression trees, we construct *N* regression trees using *N* bootstrapped training sets and average the resulting predictions.

These trees are grown deep and not allowed to prune. Each Individual tree has low bias and high variance .Applying average reduce the variance. Bagging has given high accuracies by combining hundreds or even thousands of trees. Bagging can be extended to classification to predict the Categorical Variable. Boosting is a family of methods. It produces a series of Classifiers. The training set used for each member of the series is chosen based on the performance of the earlier classifier in the series.



Software analysis

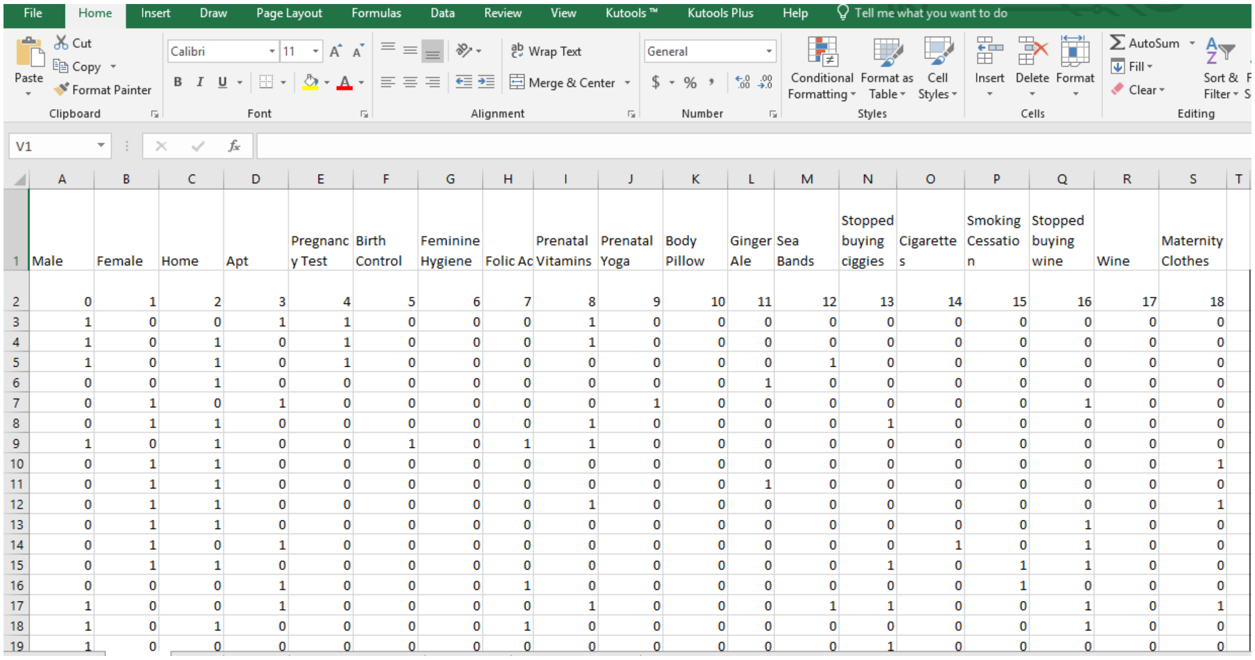
Data set given is Retail Customer data. We have to use our Statistical Approaches to model the data. We have many tools to execute our model and many free programming languages to make this model. In this Project we are using Microsft Excel to perform our Analysis.

Excel is one of the classic tools to perform data analysis. Here we have used the Excel by plugging in Data Analysis Tool Pack. Microsoft Excel is one of the less expensive tools for data analysis. We have utilized mathematical functions and graphing techniques for Model Accuracy Visualizations.

Data is downloaded from wiley.com and the data consist of .The following columns are in data set:

1. Account Holder is Male/Female/Unknown by matching surname to census data
2. Account holder address is a home, apartment or PO box
3. Recently purchased a pregnancy test
4. Recently purchased birth control
5. Recently purchased feminine hygiene products
6. Recently purchased folic acid supplements
7. Recently purchased prenatal vitamins
8. Recently purchased prenatal yoga DVD
9. Recently purchased body pillow
10. Recently purchased ginger ale
11. Recently purchased Sea-Bands
12. Bought cigarettes regularly until recently, then stopped
13. Recently purchased cigarettes
14. Recently purchased smoking cessation products (gum, patch, etc.)
15. Bought wine regularly until recently, then stopped
16. Recently purchased wine
17. Recently purchased maternity clothing

Sample Snapshot of the data :



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Male	Female	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Ac	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	
2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
3	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
4	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
5	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
6	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
7	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
8	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	
9	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	
10	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
11	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
12	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	
13	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
14	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	
15	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	
16	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	
17	1	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0	1	0	1	
18	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	
19	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	

Data Cleaning

The data provided has no holes in it. Hence we can apply the same data for our analysis.

Chapter 3

Ensemble Technique

Its an Machine Learning Technique that uses trees basically. It is Used both as Classifier and Regressor. But in our Project we have utilized Ensemble Technique as Classifier to Predict the Customers whether they are Pregnant. The Models intention is to classify the Customers in to two categories:

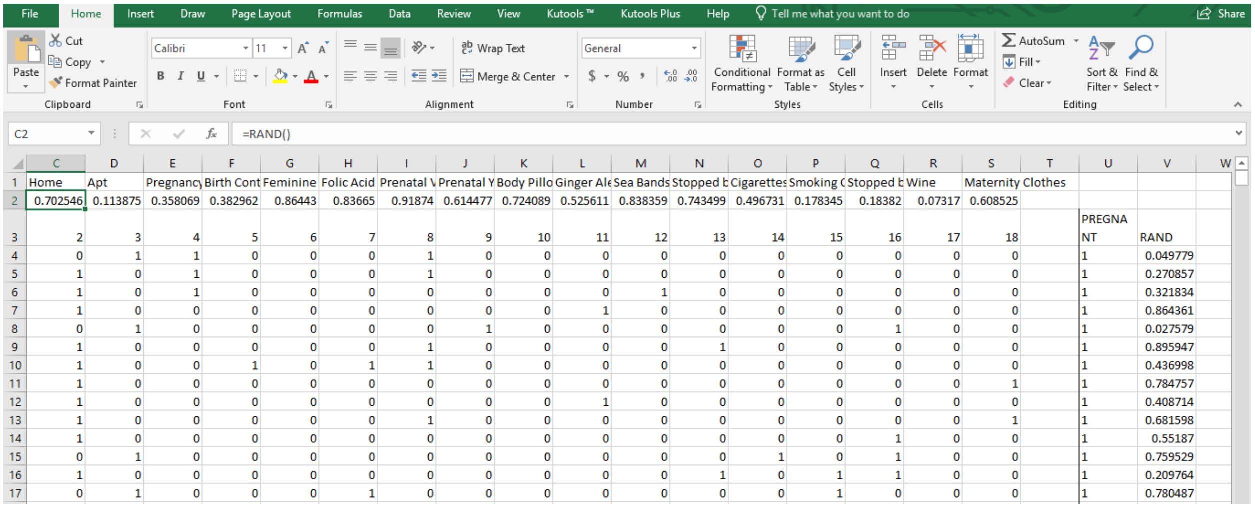
1. Pregnant
2. Non Pregnant Customers

Based on which the product promotion and products are endorsed to customers. This project really proves the data is highly valuable .In below sections we clearly explain the implementation of the Bagging Technique.

Implementation

Random Sampling:

Random Sampling is done using Rand( ) Function. Rand formula is added to row 2 and the same is depicted in the below figure:

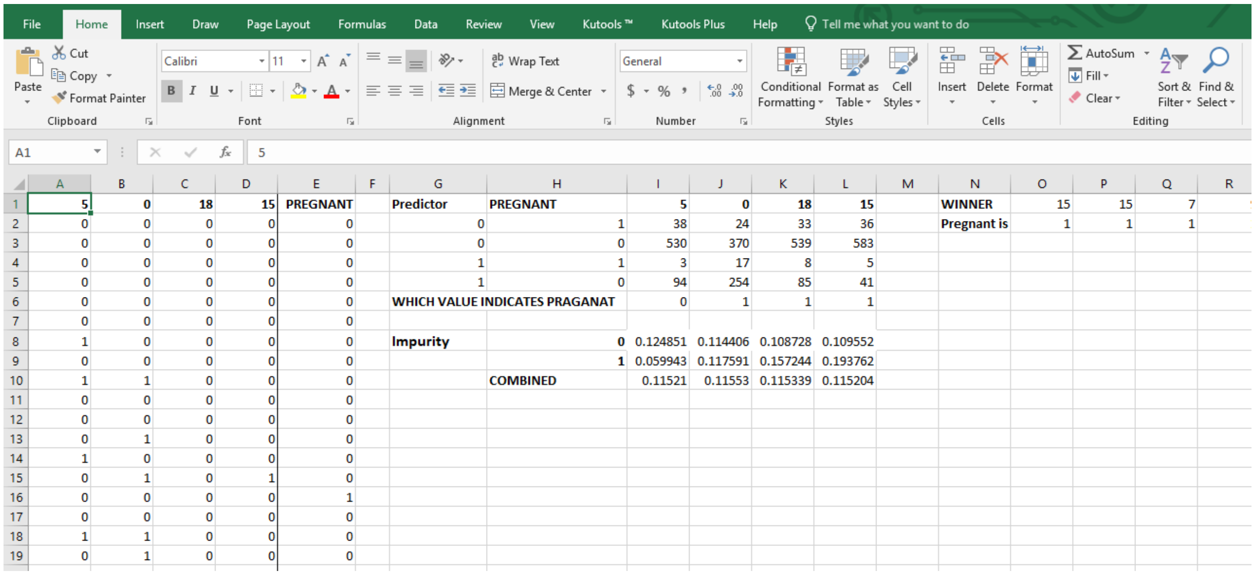


C2 =RAND()

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Home	Apt	Pregnancy	Birth Cont	Feminine	Folic Acid	Prenatal V	Prenatal Y	Body Pillo	Ginger Ale	Sea Bands	Stopped b	Cigarettes	Smoking C	Stopped b	Wine	Maternity Clothes			
2	0.702546	0.113875	0.358069	0.382962	0.86443	0.83665	0.91874	0.614477	0.724089	0.525611	0.838359	0.743499	0.496731	0.178345	0.18382	0.07317	0.608525			
3	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		PREGNANT	RAND
4	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.049779
5	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.270857
6	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0		1	0.321834
7	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.864361
8	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0		1	0.027579
9	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0		1	0.895947
10	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0		1	0.436998
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		1	0.784757
12	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.408714
13	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1		1	0.681598
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		1	0.55187
15	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0		1	0.759529
16	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0		1	0.209764
17	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0		1	0.780487

Column and rows are Sorted randomly. Custom Sort Window is opened and options button is used to sort left to right to sort the columns. Certain measures are taken to make sure that Row 2, the row with random numbers is selected as the row to sort by.

Same Procedure is followed for rows:



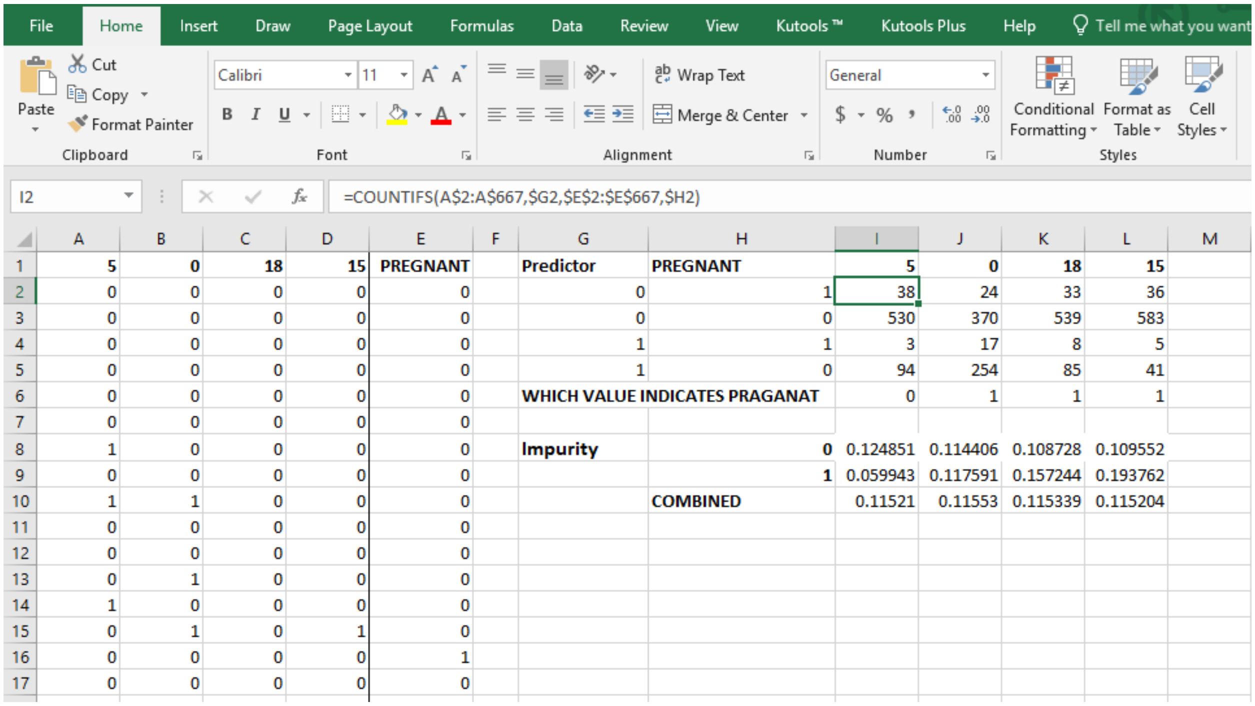
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	15	15	7	
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583						
4	0	0	0	0	0		1	1	3	17	8	5						
5	0	0	0	0	0		1	0	94	254	85	41						
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1						
7	0	0	0	0	0													
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552						
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762						
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204						
11	0	0	0	0	0													
12	0	0	0	0	0													
13	0	1	0	0	0													
14	1	0	0	0	0													
15	0	1	0	1	0													
16	0	0	0	0	1													
17	0	0	0	0	0													
18	1	1	0	0	0													
19	0	1	0	0	0													

Decision Stumps

Here we see for features that could determine the Model. Here we use countifs to count number of training rows in which feature 5 is a 0 and pregnant column is 1.

Formula :

=COUNTIFS(A$2:A$667,$G2,$E$2:$E$667,$H2)

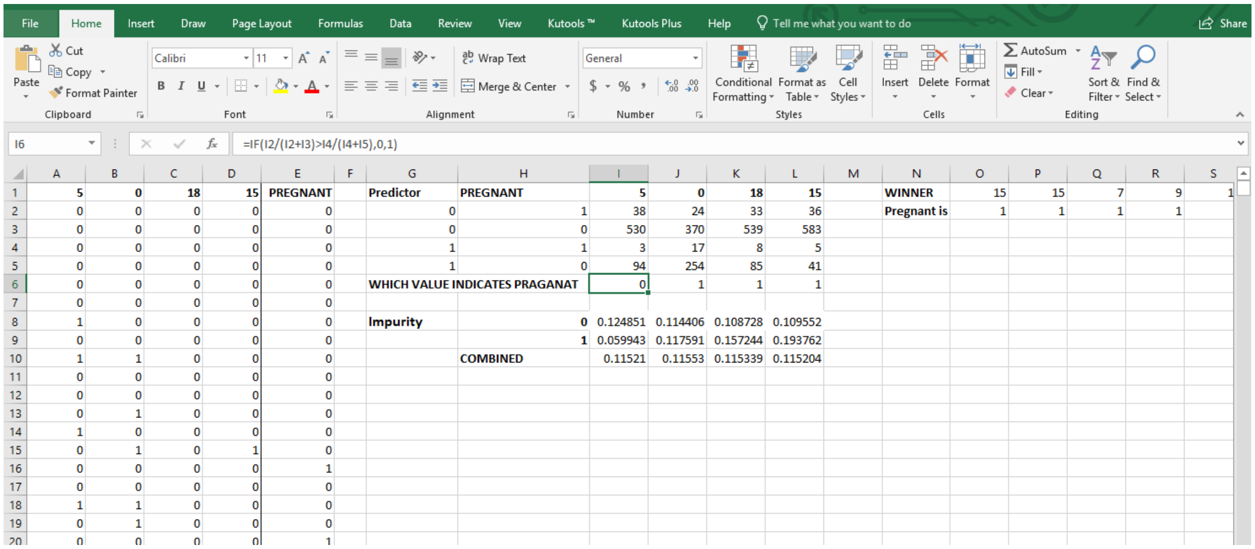


I2 =COUNTIFS(A$2:A$667,$G2,$E$2:$E$667,$H2)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15	
2	0	0	0	0	0		0	1	38	24	33	36	
3	0	0	0	0	0		0	0	530	370	539	583	
4	0	0	0	0	0		1	1	3	17	8	5	
5	0	0	0	0	0		1	0	94	254	85	41	
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1	
7	0	0	0	0	0								
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552	
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762	
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204	
11	0	0	0	0	0								
12	0	0	0	0	0								
13	0	1	0	0	0								
14	1	0	0	0	0								
15	0	1	0	1	0								
16	0	0	0	0	1								
17	0	0	0	0	0								

Formula given below is copied and applied to remaining cells to get counts. This would give us the value with highest concentration of customers with pregnancy in the given sample data. The formula that used to calculate is

=IF(I2/(I2+I3)>I4/(I4+I5),0,1)



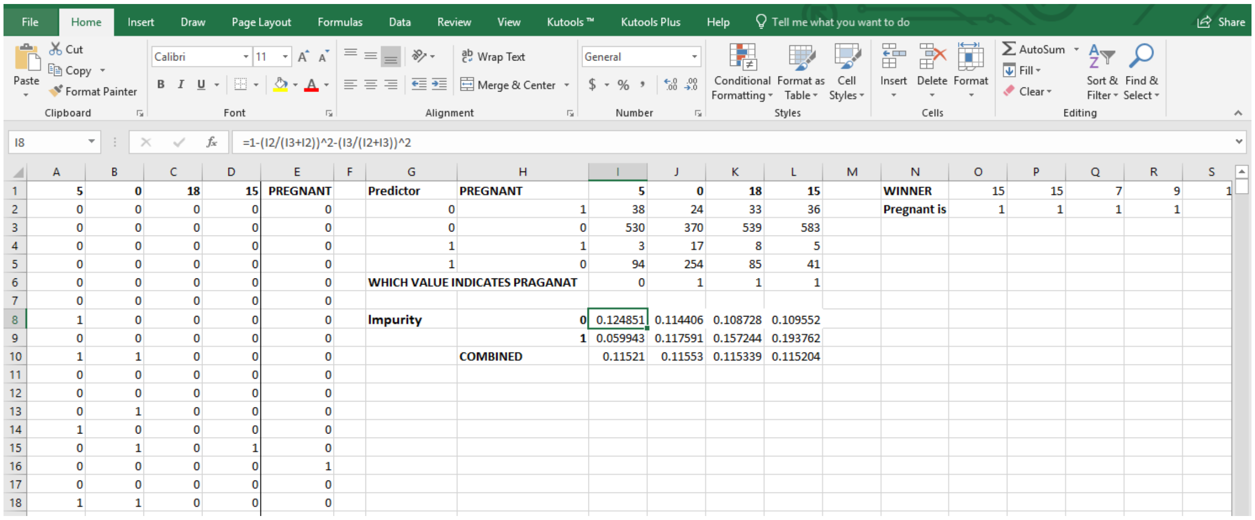
I6 =IF(I2/(I2+I3)>I4/(I4+I5),0,1)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														
19	0	1	0	0	0														
20	0	0	0	0	1														

If pregnant customers with 0 value for the feature (I2/ (I2+I3)) is greater than the associated value with (I4/ (I4+I5)),then 0 is the stump that predicts the pregnancy. Otherwise 1.

Impurity for 0 is calculated using

=1-(I2/(I3+I2))^2-(I3/(I2+I3))^2

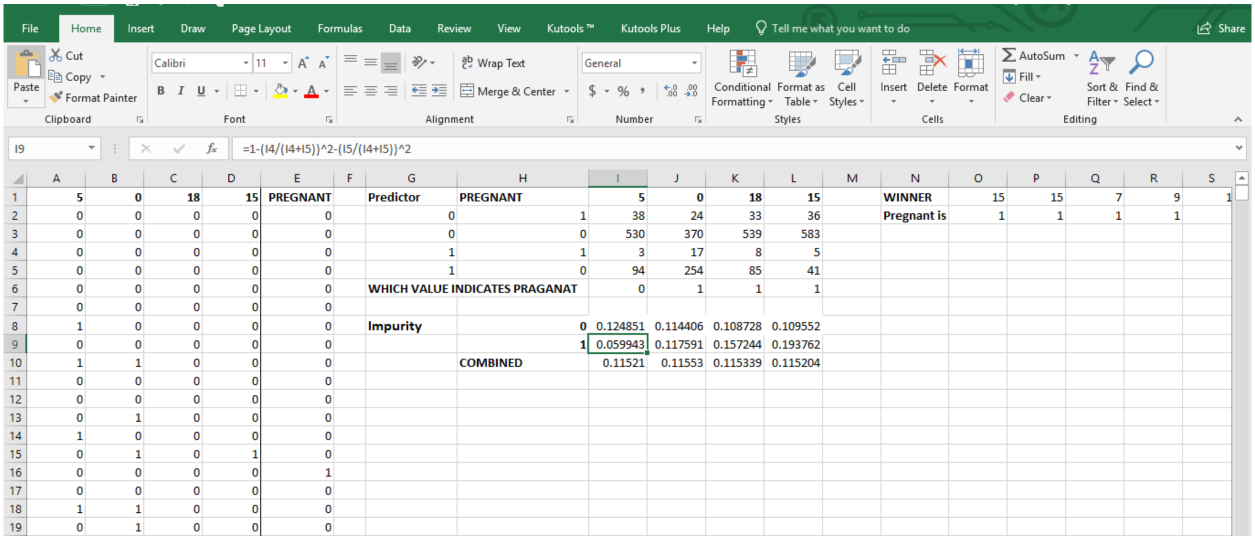


I8 =1-(I2/(I3+I2))^2-(I3/(I2+I3))^2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														

Impurity for 1 is calculated using

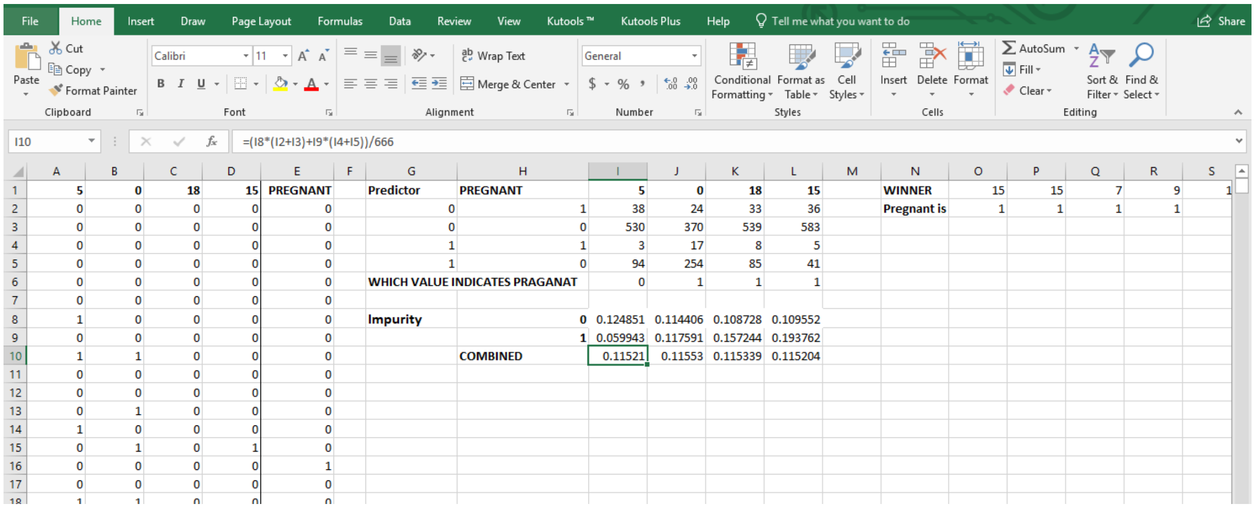
=1-(I4/(I4+I5))^2-(I5/(I4+I5))^2



I9 =1-(I4/(I4+I5))^2-(I5/(I4+I5))^2

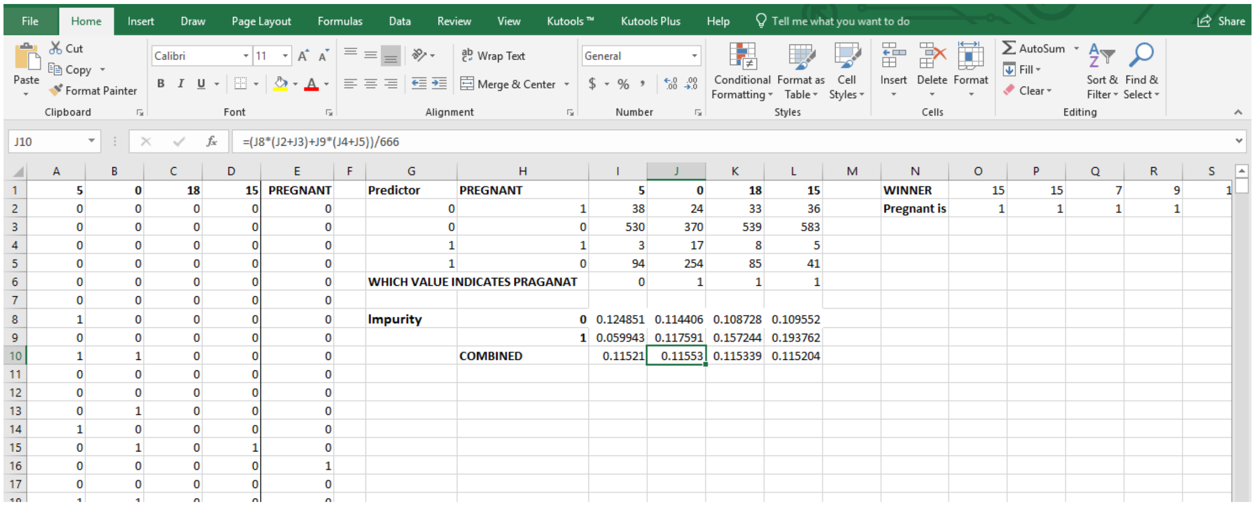
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														
19	0	1	0	0	0														

Sum of all is added and divided by 666:



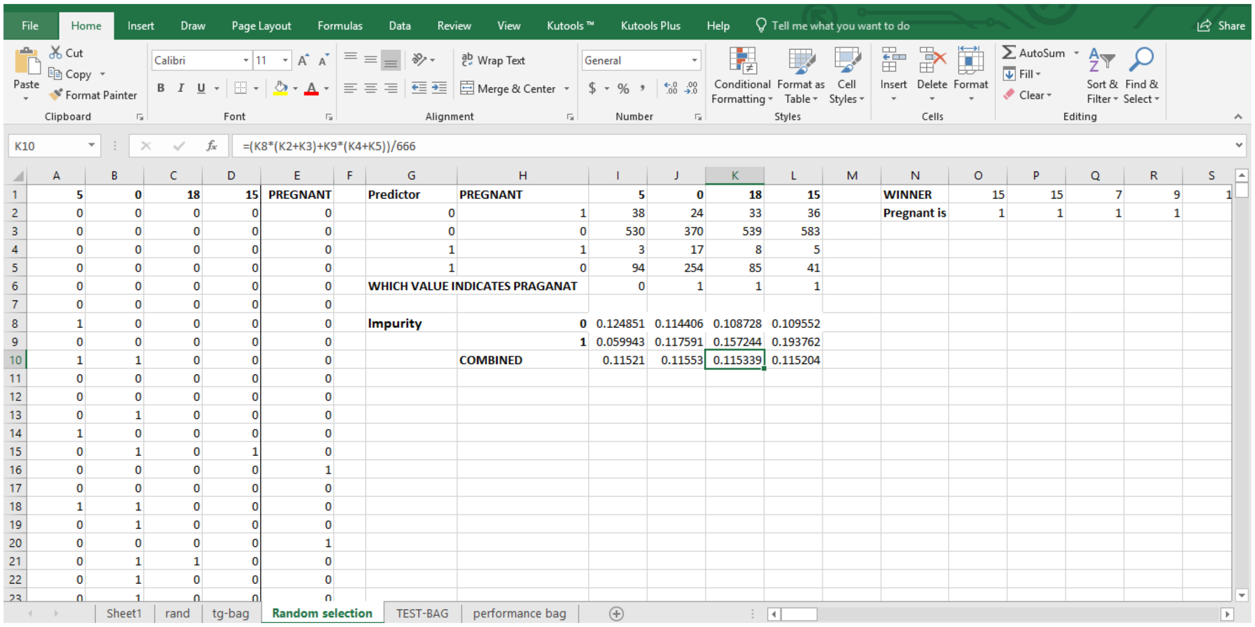
I10 =(I8*(I2+I3)+I9*(I4+I5))/666

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**5**	**0**	**18**	**15**	**PREGNANT**		**Predictor**	**PREGNANT**	**5**	**0**	**18**	**15**		**WINNER**	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		**Pregnant is**	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		**WHICH VALUE INDICATES PRAGANAT**		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		**Impurity**	**0**	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			**1**	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			**COMBINED**	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														



J10 =(J8*(J2+J3)+J9*(J4+J5))/666

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**5**	**0**	**18**	**15**	**PREGNANT**		**Predictor**	**PREGNANT**	**5**	**0**	**18**	**15**		**WINNER**	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		**Pregnant is**	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		**WHICH VALUE INDICATES PRAGANAT**		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		**Impurity**	**0**	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			**1**	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			**COMBINED**	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														



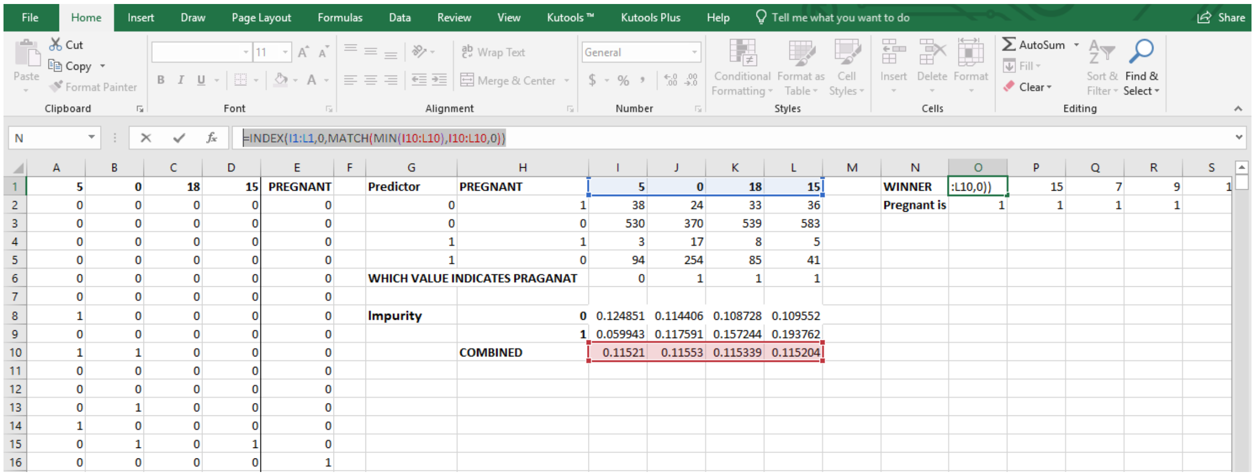
K10 =(K8*(K2+K3)+K9*(K4+K5))/666

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														
19	0	1	0	0	0														
20	0	0	0	0	1														
21	0	1	1	0	0														
22	0	1	0	0	0														
23	0	1	0	0	0														

Winner:

The indexes are matched using the formula

==INDEX(I1:L1,0,MATCH(MIN(I10:L10),I10:L10,0))

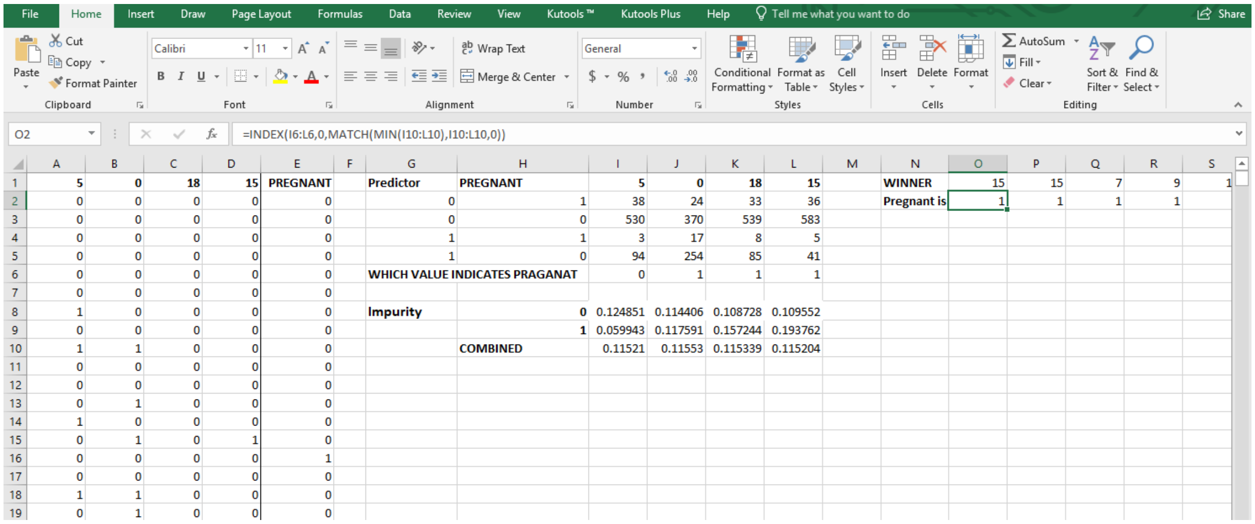


N =INDEX(I1:L1,0,MATCH(MIN(I10:L10),I10:L10,0))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	:L10,0))	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														

Using this formula the below index is matched:

=INDEX(I6:L6,0,MATCH(MIN(I10:L10),I10:L10,0))

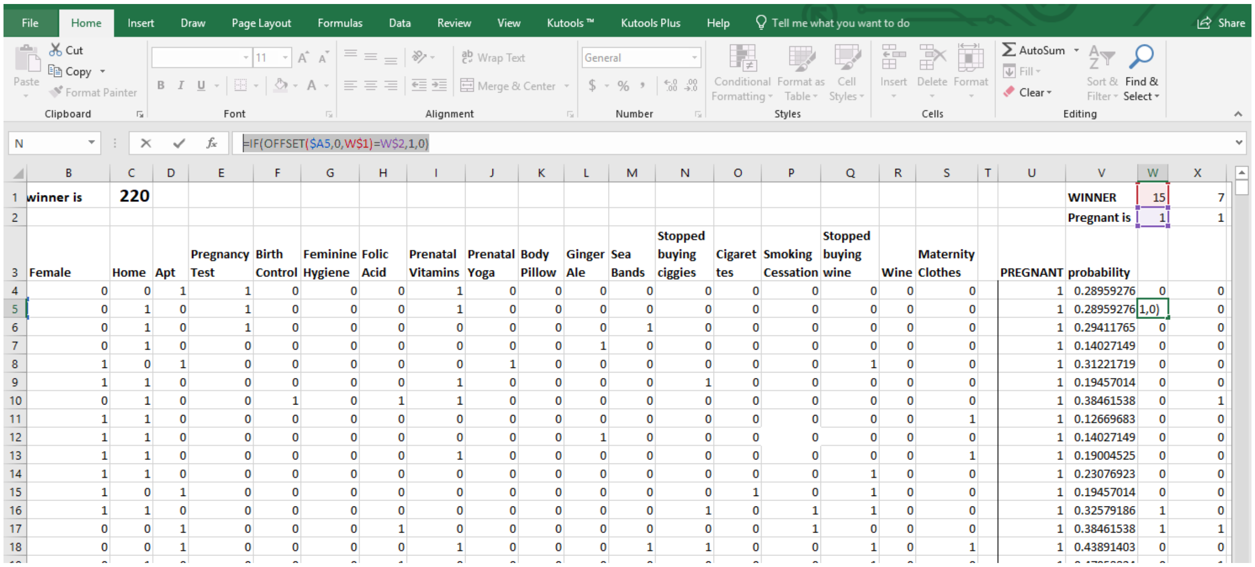


O2 =INDEX(I6:L6,0,MATCH(MIN(I10:L10),I10:L10,0))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	5	0	18	15	PREGNANT		Predictor	PREGNANT	5	0	18	15		WINNER	15	15	7	9	1
2	0	0	0	0	0		0	1	38	24	33	36		Pregnant is	1	1	1	1	
3	0	0	0	0	0		0	0	530	370	539	583							
4	0	0	0	0	0		1	1	3	17	8	5							
5	0	0	0	0	0		1	0	94	254	85	41							
6	0	0	0	0	0		WHICH VALUE INDICATES PRAGANAT		0	1	1	1							
7	0	0	0	0	0														
8	1	0	0	0	0		Impurity	0	0.124851	0.114406	0.108728	0.109552							
9	0	0	0	0	0			1	0.059943	0.117591	0.157244	0.193762							
10	1	1	0	0	0			COMBINED	0.11521	0.11553	0.115339	0.115204							
11	0	0	0	0	0														
12	0	0	0	0	0														
13	0	1	0	0	0														
14	1	0	0	0	0														
15	0	1	0	1	0														
16	0	0	0	0	1														
17	0	0	0	0	0														
18	1	1	0	0	0														
19	0	1	0	0	0														

Now finally we can predict the pregnancy of the customer using the offset formula:

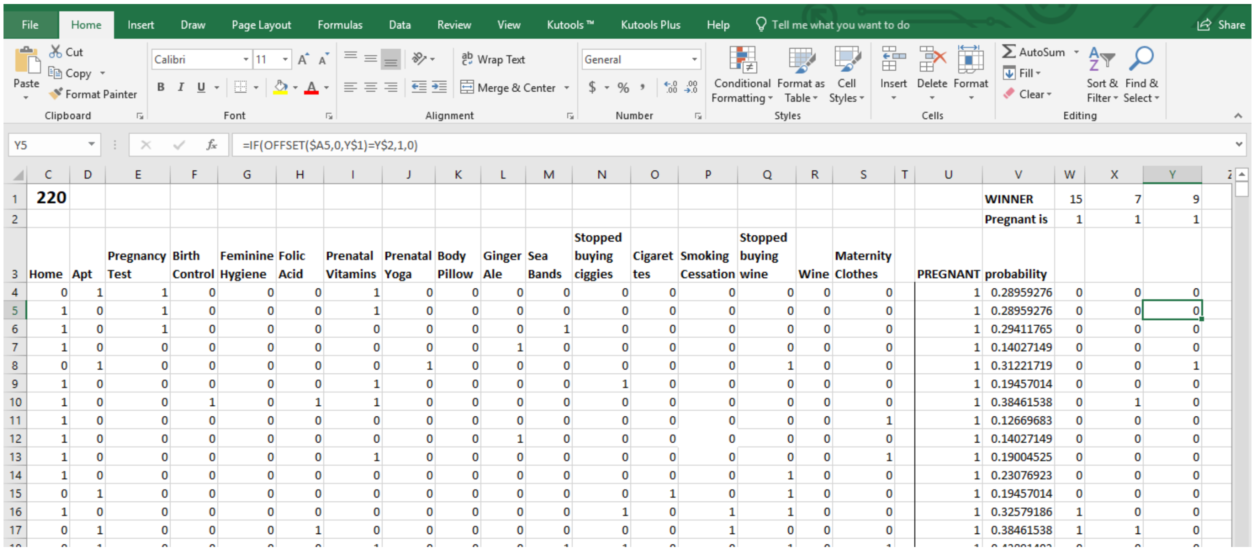
=IF(OFFSET($A5,0,W$1)=W$2,1,0)



N =IF(OFFSET($A5,0,W$1)=W$2,1,0)

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	winner is	220																			WINNER	15	7
2																					Pregnant is	1	1
3	Female	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigaret tes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes		PREGNANT	probability		
4	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.28959276	0	0
5	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.28959276	1,0)	0
6	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0		1	0.29411765	0	0
7	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.14027149	0	0
8	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0		1	0.31221719	0	0
9	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0		1	0.19457014	0	0
10	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0		1	0.38461538	0	1
11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		1	0.12669683	0	0
12	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.14027149	0	0
13	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1		1	0.19004525	0	0
14	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		1	0.23076923	0	0
15	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0		1	0.19457014	0	0
16	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0		1	0.32579186	1	0
17	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0		1	0.38461538	1	1
18	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0	1	0	1		1	0.43891403	0	0

The same formula is dragged and applied to all the cells:



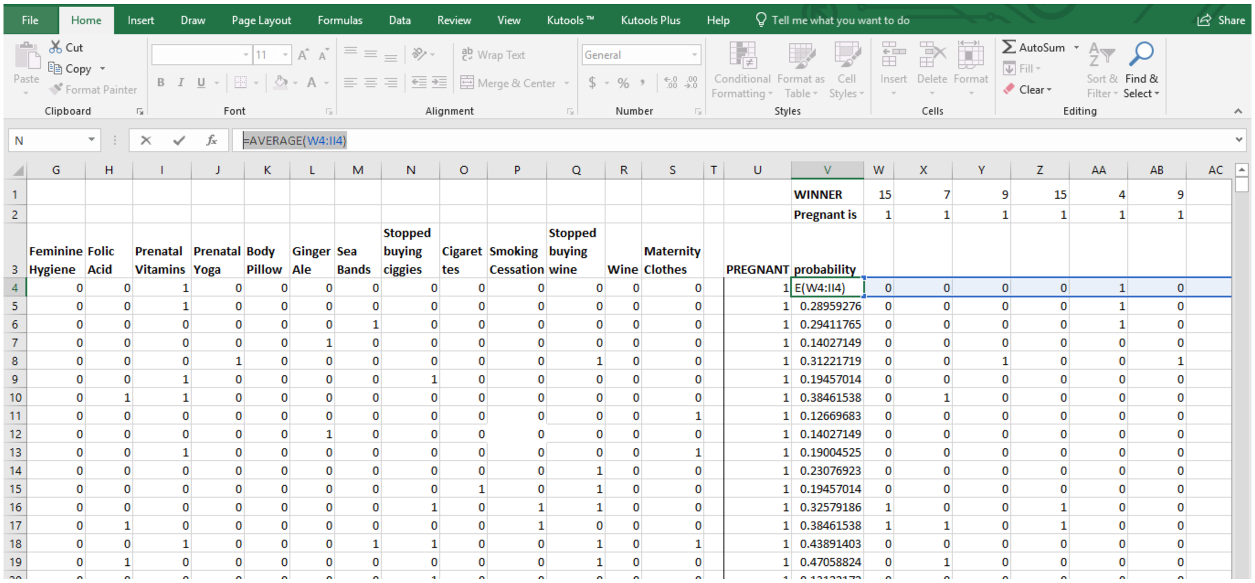
Y5 =IF(OFFSET($A5,0,Y$1)=Y$2,1,0)

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	220																			WINNER	15	7	9
2																				Pregnant is	1	1	1
3	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes		PREGNANT	probability			
4	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.28959276	0	0	0
5	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.28959276	0	0	0
6	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0		1	0.29411765	0	0	0
7	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.14027149	0	0	0
8	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0		1	0.31221719	0	0	1
9	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0		1	0.19457014	0	0	0
10	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0		1	0.38461538	0	1	0
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		1	0.12669683	0	0	0
12	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.14027149	0	0	0
13	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1		1	0.19004525	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		1	0.23076923	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0		1	0.19457014	0	0	0
16	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0		1	0.32579186	1	0	0
17	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0		1	0.38461538	1	1	0

After all these Steps, key step in our analysis is calculating the Probabalities and determining the winners.

The probabilities are calculated using the formula:

=AVERAGE(W4:II4)

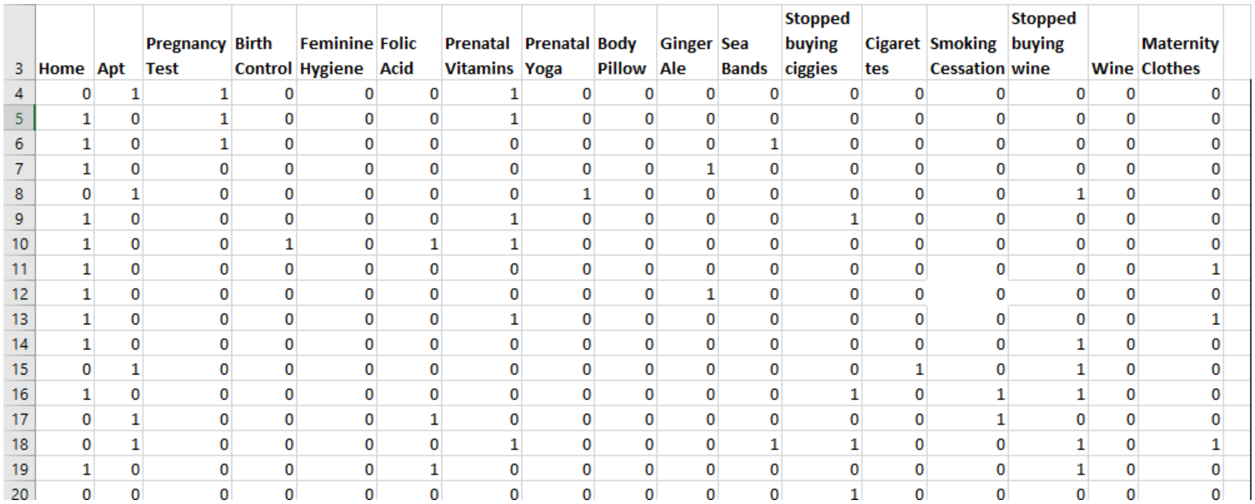


N =AVERAGE(W4:II4)

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1																WINNER	15	7	9	15	4	9	
2																Pregnant is	1	1	1	1	1	1	
3	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes		PREGNANT	probability							
4	0	0	1	0	0	0	0	0	0	0	0	0	0		1	E(W4:II4)	0	0	0	0	1	0	
5	0	0	1	0	0	0	0	0	0	0	0	0	0		1	0.28959276	0	0	0	0	1	0	
6	0	0	0	0	0	0	1	0	0	0	0	0	0		1	0.29411765	0	0	0	0	1	0	
7	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.14027149	0	0	0	0	0	0	
8	0	0	0	1	0	0	0	0	0	0	1	0	0		1	0.31221719	0	0	1	0	0	1	
9	0	0	1	0	0	0	0	1	0	0	0	0	0		1	0.19457014	0	0	0	0	0	0	
10	0	1	1	0	0	0	0	0	0	0	0	0	0		1	0.38461538	0	1	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	1		1	0.12669683	0	0	0	0	0	0	
12	0	0	0	0	0	1	0	0	0	0	0	0	0		1	0.14027149	0	0	0	0	0	0	
13	0	0	1	0	0	0	0	0	0	0	0	0	1		1	0.19004525	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	1	0	0		1	0.23076923	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	0	1	0	1	0	0		1	0.19457014	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	1	0	1	1	0	0		1	0.32579186	1	0	0	1	0	0	
17	0	1	0	0	0	0	0	0	0	1	0	0	0		1	0.38461538	1	1	0	1	0	0	
18	0	0	1	0	0	0	1	1	0	0	1	0	1		1	0.43891403	0	0	0	0	0	0	
19	0	1	0	0	0	0	0	0	0	0	1	0	0		1	0.47058824	0	1	0	0	0	0	

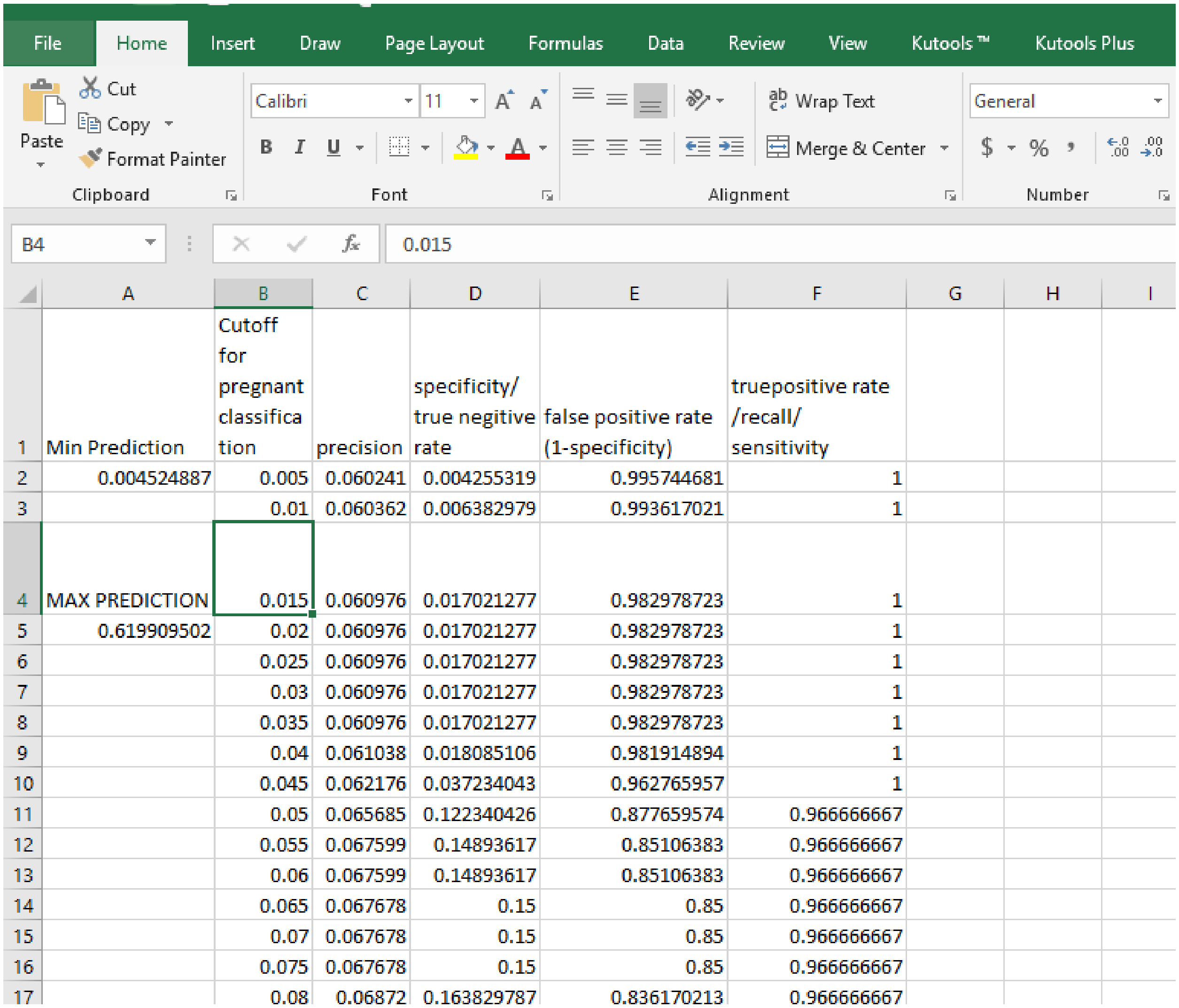
Outcomes and Evaluation:

Lets evaluate the model using the test data. Test-Bag is created in order to evaluate the performance of the model



3	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes
4	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
9	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
10	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
13	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
15	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
16	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
17	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
18	0	1	0	0	0	0	1	0	0	0	1	1	0	0	1	0	1
19	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
20	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

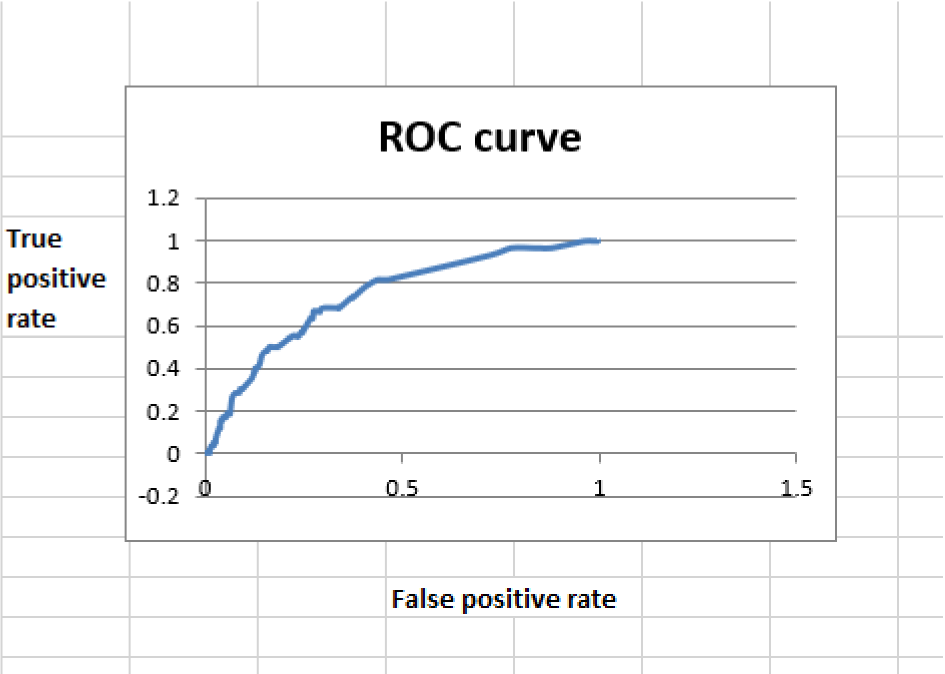
Performance is evaluated using ROC Curves,That is plotted across True Positive Rate and False Positive Rate .



	A	B	C	D	E	F
1	Min Prediction	Cutoff for pregnant classification	precision	specificity/ true negitive rate	false positive rate (1-specificity)	truepositive rate /recall/ sensitivity
2	0.004524887	0.005	0.060241	0.004255319	0.995744681	1
3		0.01	0.060362	0.006382979	0.993617021	1
4	MAX PREDICTION	0.015	0.060976	0.017021277	0.982978723	1
5	0.619909502	0.02	0.060976	0.017021277	0.982978723	1
6		0.025	0.060976	0.017021277	0.982978723	1
7		0.03	0.060976	0.017021277	0.982978723	1
8		0.035	0.060976	0.017021277	0.982978723	1
9		0.04	0.061038	0.018085106	0.981914894	1
10		0.045	0.062176	0.037234043	0.962765957	1
11		0.05	0.065685	0.122340426	0.877659574	0.966666667
12		0.055	0.067599	0.14893617	0.85106383	0.966666667
13		0.06	0.067599	0.14893617	0.85106383	0.966666667
14		0.065	0.067678	0.15	0.85	0.966666667
15		0.07	0.067678	0.15	0.85	0.966666667
16		0.075	0.067678	0.15	0.85	0.966666667
17		0.08	0.06872	0.163829787	0.836170213	0.966666667

Precision and True Negative rates and False Positive rate are calculated as part of Metrics Evaluation.

Roc Curve of the Model:



The above ROC Curve is plotted between the True Positive Rate and False Positive Rate. It clearly explains the tradeoff between Sensitivity and specifity. The Curve is not that close to Left hand. Hence the Accuracy of the model can be improved via Boosting.

The area of the ROC curve can be calculated and the Accuracy can be roughly classified using the below values (Area)

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

The area of the above ROC Curve is 75.Hence the model is fair. But it can be boosted Via Ada Boosting.

SUMMARY

The project implemented the Ensemble Technique using Excel. First the data that has been collected is very clean hence we did not apply an Data Cleaning Techniques to clean the data. The data had some of the variables that are not numerical hence we have done some of the dummy variable techniques to convert them in to categorical. Now our data is ready to perform our Analysis. We have applied the Random Function on our Rows and Columns using Rand() Function. Then follows the calculation of Impurity of node 1 and Node 0 ,and all of its indexes are matched. The average has been calculated and the Probabilities has been calculated using the average Formula.

The winner was predicted and the Performance of the model is evaluated using test Bag. The Roc curve and Roc Area is Calculated to asses the model and the finally the Theoretical Explanation of the model seems to be sophisticated in theory but implementation proves that wrong. The Roc Area of the curve shows the Model is fair to predict the Output but models Prediction can be improved Further.

CONCLUSION

Bagging Technique proves to be one of the powerful methods for the Classification Problem. The strong learners and Weak Learners are bagged together and Predicted. This technique practically reduces the variance that was clearly observed. Logistic Regression, One of the oldest techniques used in the Industry for the solving the classification Problem has less Accuracy compared to the Bagging Techniques Accuracy.

This Technique relies on the Output of multiple models Trained on the same Dataset. The prediction outcome is taken as average and made the aggregated prediction. In this Project we had an opportunity to implement this technique from scratch without using any pack. This approach has utilized Rand Function and the trade off between the Sensitivity and specificity is clearly observed in the ROC Graph. The models accuracy is fair according to the area of the ROC Curve, the models performance can be boosted using the Ada Boosting Approaches. The area of the Roc Curve is 78 and according to the rough estimates its less than the standard accuracy of > 90.

References:

[1]. *Working Analytics*, workinganalytics.com/case-studies-2/.

[2]. Crossman, Ashley. "Data Cleaning for Data Analysis in Sociology." *ThoughtCo*, www.thoughtco.com/data-cleaning-3026541.

[3]. *Data Analysis*, ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html.

[4]. "Data Transformation." *MuleSoft*, 26 Oct. 2017, www.mulesoft.com/resources/esb/data-transformation.

[5]. "Data Transformation." *Wikipedia*, Wikimedia Foundation, 13 Oct. 2017, en.wikipedia.org/wiki/Data_transformation.

[6]. Ray, Sunil, et al. "7 Types of Regression Techniques You Should Know." *Analytics Vidhya*, 4 May 2017, www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/.

[7]. statisticaldesignmethods.com/files/regression-techniques.pdf++.

[8]. Http://Statlab.stat.yale.edu/Workshops/IntroRegression/StatLab-IntroRegressionFa08.Pdf.

[9]. Frost, Jim. "The Danger of Overfitting Regression Models." *Minitab*, 3 Sept. 1970, blog.minitab.com/blog/adventures-in-statistics-2/the-danger-of-overfitting-regression-models.

[10]. Http://Sites.stat.psu.edu/~ajw13/SpecialTopics/Multicollinearity.pdf.

[11]. "Assumptions of Linear Regression." *Statistics Solutions*, www.statisticssolutions.com/assumptions-of-linear-regression/.

[12].Http://www.dianneballanceportfolio.com/Uploads/1/2/8/2/12825938/assumptions_in_multiple_regression.Pdf.

[13]. *Regression Using Excel's Solver*, archives.math.utk.edu/ICTCM/VOL13/C013/paper.html.

[14]. "Linear Regression in Excel: 3 Alternative Methods." *EngineerExcel*, 20 June 2016, www.engineerexcel.com/linear-regression-in-excel-3-alternative-methods/.

[15]. Oppenheimer, Diego. "New and Improved Solver." *Office Blogs*, 22 Sept. 2009, blogs.office.com/en-us/2009/09/21/new-and-improved-solver/.

[16]. "Using Solver to Assign Items to Buckets." *Chandoo.org - Learn Excel & Charting Online*, 20 June 2011, chandoo.org/wp/2011/05/11/using-solver-to-assign-item/.

[17].Http://ww.eng.auburn.edu/~Clemept/CEANALYSIS_FALL2011/Week1/non_Linearregression_paper.Pdf.pp.199-200

[18].Http://www.eng.auburn.edu/~Clemept/CEANALYSIS_FALL2011/Week1/non_Linearregression_paper.Pdf. pp.198

[19]. "19. Http://www.math.utah.edu/~Gamez/Files/ROC-Curves.pdf."

[20]. Https://Synapse.koreamed.org/Synapse/Data/PDFData/0006JKAN/Jkan-43-154.Pdf.pp.155

[21]. "Logistic Regression with Stata Chapter 1: Introduction to Logistic Regression with Stata."*IDRE Stats*, stats.idre.ucla.edu/stata/webbooks/logistic/chapter1/logistic-regression-with-statachapter-1-introduction-to-logistic-regression-with-stata/

[22]. Https://Synapse.koreamed.org/Synapse/Data/PDFData/0006JKAN/Jkan-43-154.Pdf. pp. 160

[23]. Http://www.kentohio.org/Dep/KentTradeAreaHouseholdSegmentation.pdf.